Feature Induction in Machine Learning

Ganesh Ramakrishnan

Collaboration: J. Saketha Nath, Pratik J.,
Naveen Nair and Amrita Saha.

October. 21, 2012

Primary areas of my research

- Efficient Feature Induction
- Information Extraction and Disambiguation (in the enterprise domain)
- Search over Entities and Relationships (in the enterprise domain)

Primary areas of my research

- Efficient Feature Induction
- Information Extraction and Disambiguation (in the enterprise domain)
- Search over Entities and Relationships (in the enterprise domain)

Primary areas of my research

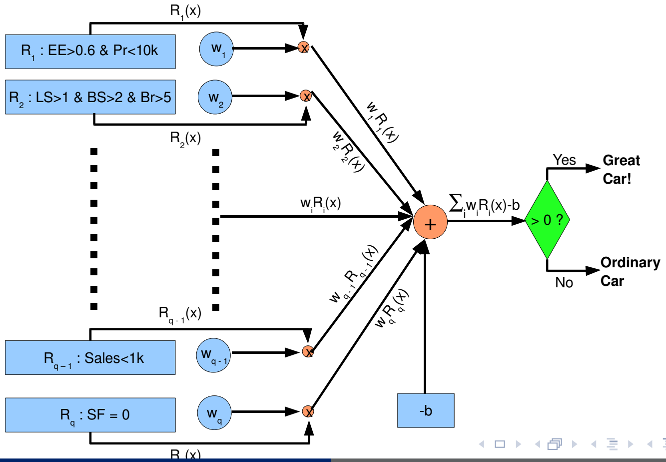
- Efficient Feature Induction
- Information Extraction and Disambiguation (in the enterprise domain)
- Search over Entities and Relationships (in the enterprise domain)

Outline

1 Introduction

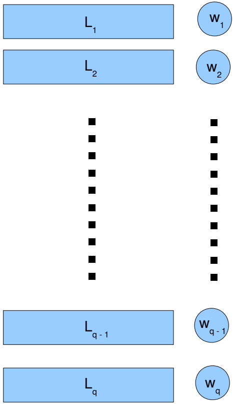
Feature and Parameter Induction for Statistical Modeling

- Example: Learning model for classifying cars (Cohen&Singer, 99)
- Simple *boolean* features (look like rules)



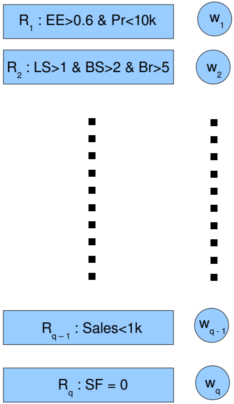
Feature and Parameter Induction for Statistical Modeling

- Example: Learning model for classifying cars (Cohen&Singer, 99)
- Simple *boolean* features (look like rules)



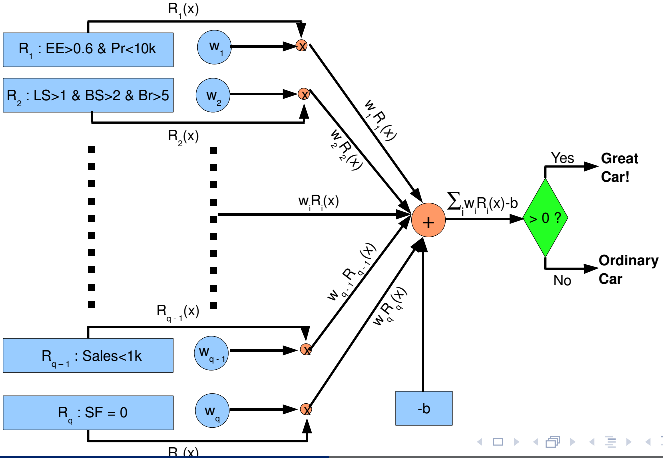
Feature and Parameter Induction for Statistical Modeling

- Example: Learning model for classifying cars (Cohen&Singer, 99)
- Simple *boolean* features (look like rules)



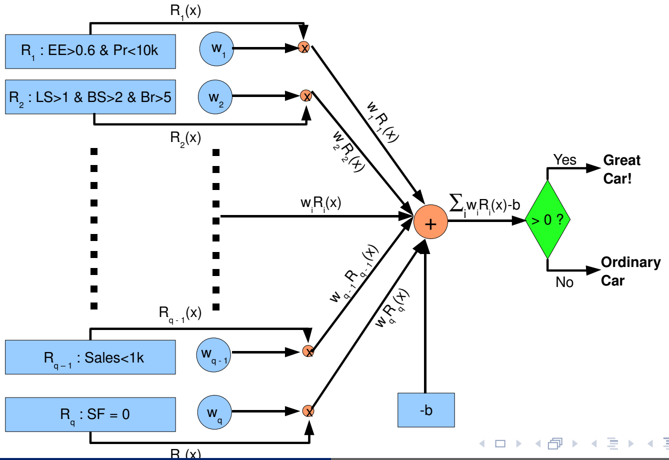
Feature and Parameter Induction for Statistical Modeling

- Example: Learning model for classifying cars (Cohen&Singer, 99)
- Simple *boolean* features (look like rules)



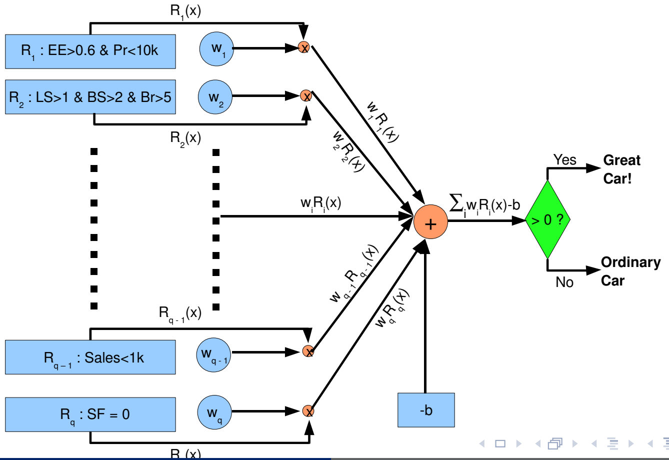
Feature and Parameter Induction for Statistical Modeling

- Example: Learning model for classifying cars (Cohen&Singer, 99)
- Simple *boolean* features (look like rules)



Feature and Parameter Induction for Statistical Modeling

- Example: Learning model for classifying cars (Cohen&Singer, 99)
- Simple *boolean* features (look like rules)



Feature Induction

- Learning features for any predictive learning task
  - identifying key attributes $\Rightarrow$ characterise the domain
  - good generalization performance
  - human-interpretable.
- to induce a compact set of relevant features
- understand how they interact with the strengths and limitations of the predictive learner

Feature Induction

- Learning features for any predictive learning task
  - identifying key attributes $\Rightarrow$ characterise the domain
  - good generalization performance
  - human-interpretable.
- to induce a compact set of relevant features
- understand how they interact with the strengths and limitations of the predictive learner

Feature Induction

- Learning features for any predictive learning task
  - identifying key attributes $\Rightarrow$ characterise the domain
  - good generalization performance
  - human-interpretable.
- to induce a compact set of relevant features
- understand how they interact with the strengths and limitations of the predictive learner

Feature Induction

- Learning features for any predictive learning task
  - identifying key attributes $\Rightarrow$ characterise the domain
  - good generalization performance
  - human-interpretable.
- to induce a compact set of relevant features
- understand how they interact with the strengths and limitations of the predictive learner

Feature Induction

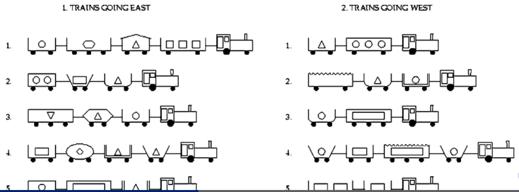
- Learning features for any predictive learning task
  - identifying key attributes ⇒ characterise the domain
  - good generalization performance
  - human-interpretable.
- to induce a compact set of relevant features
- understand how they interact with the strengths and limitations of the predictive learner

Feature Induction

- Learning features for any predictive learning task
  - identifying key attributes $\Rightarrow$ characterise the domain
  - good generalization performance
  - human-interpretable.
- to induce a compact set of relevant features
- understand how they interact with the strengths and limitations of the predictive learner

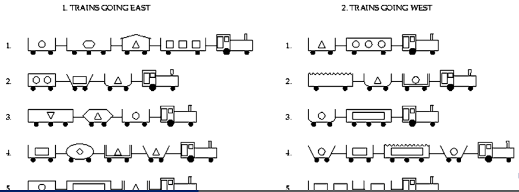
Kinds of Propositional Features/Rules

- Simple Conjunctive features [ICML2011, AAAI2012, CIKM2012]
  - Eg. Features Like $sleeping \leftarrow bedroomDoor \wedge toiletDoor \wedge bathroomDoor$ (Activity Recognition)
- Simple Disjunctive features (under review)
  - Eg. Rules Like $PositiveSentiment \leftarrow exquisite \vee elegant$ (Sentiment Analysis)
- Features in the more general language of First order logic [MLJ2009, ILP2012a, EMNLP2012, ILP2012b]
  - $eastbound(A) \leftarrow hasCar(A,B),\ hasCar(A,C),\ short(B),\ closed(C)$



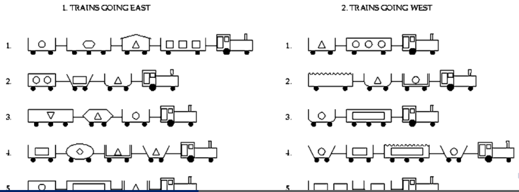
Kinds of Propositional Features/Rules

- Simple Conjunctive features [ICML2011, AAAI2012, CIKM2012]
  - Eg. Features Like $sleeping \leftarrow bedroomDoor \land toiletDoor \land bathroomDoor$ (Activity Recognition)
- Simple Disjunctive features (under review)
  - Eg. Rules Like $PositiveSentiment \leftarrow exquisite \lor elegant$ (Sentiment Analysis)
- Features in the more general language of First order logic [MLJ2009, ILP2012a, EMNLP2012, ILP2012b]
  - $eastbound(A) \leftarrow hasCar(A,B),\ hasCar(A,C),\ short(B),\ closed(C)$



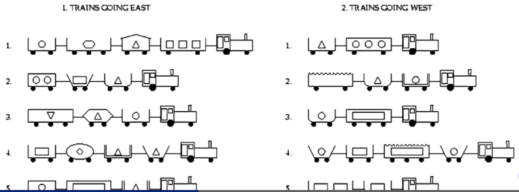
Kinds of Propositional Features/Rules

- Simple Conjunctive features [ICML2011, AAAI2012, CIKM2012]
  - Eg. Features Like $sleeping \leftarrow bedroomDoor \wedge toiletDoor \wedge bathroomDoor$ (Activity Recognition)
- Simple Disjunctive features (under review)
  - Eg. Rules Like $PositiveSentiment \leftarrow exquisite \vee elegant$ (Sentiment Analysis)
- Features in the more general language of First order logic [MLJ2009, ILP2012a, EMNLP2012, ILP2012b]
  - $eastbound(A) \leftarrow hasCar(A,B), hasCar(A,C), short(B), closed(C)$



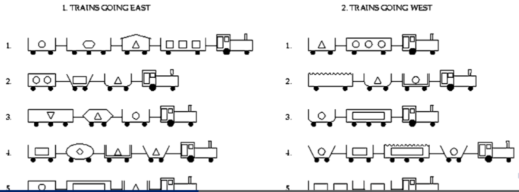
Kinds of Propositional Features/Rules

- Simple Conjunctive features [ICML2011, AAAI2012, CIKM2012]
  - Eg. Features Like $sleeping \leftarrow bedroomDoor \wedge toiletDoor \wedge bathroomDoor$ (Activity Recognition)
- Simple Disjunctive features (under review)
  - Eg. Rules Like $PositiveSentiment \leftarrow exquisite \vee elegant$ (Sentiment Analysis)
- Features in the more general language of First order logic [MLJ2009, ILP2012a, EMNLP2012, ILP2012b]
  - $eastbound(A) \leftarrow hasCar(A,B),\ hasCar(A,C),\ short(B),\ closed(C)$



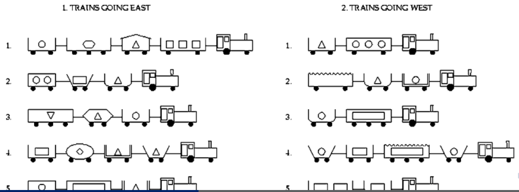
Kinds of Propositional Features/Rules

- Simple Conjunctive features [ICML2011, AAAI2012, CIKM2012]
  - Eg. Features Like $sleeping \leftarrow bedroomDoor \wedge toiletDoor \wedge bathroomDoor$ (Activity Recognition)
- Simple Disjunctive features (under review)
  - Eg. Rules Like $PositiveSentiment \leftarrow exquisite \vee elegant$ (Sentiment Analysis)
- Features in the more general language of First order logic [MLJ2009, ILP2012a, EMNLP2012, ILP2012b]
  - $eastbound(A) \leftarrow hasCar(A,B),\ hasCar(A,C),\ short(B),\ closed(C)$



Kinds of Propositional Features/Rules

- Simple Conjunctive features [ICML2011, AAAI2012, CIKM2012]
  - Eg. Features Like $sleeping \leftarrow bedroomDoor \wedge toiletDoor \wedge bathroomDoor$ (Activity Recognition)
- Simple Disjunctive features (under review)
  - Eg. Rules Like $PositiveSentiment \leftarrow exquisite \vee elegant$ (Sentiment Analysis)
- Features in the more general language of First order logic [MLJ2009, ILP2012a, EMNLP2012, ILP2012b]
  - $eastbound(A) \leftarrow hasCar(A,B),\ hasCar(A,C),\ short(B),\ closed(C)$



Desired Features

- Highly **interpretable** hypothesis
  - Small set of features/rules i.e., **low** q
  - *Simple* features e.g., **short conjunctive propositions**
- Better **generalization** than conventional rule learners

Desired Features

- Highly **interpretable** hypothesis
  - Small set of features/rules i.e., **low** q
  - *Simple* features e.g., **short conjunctive propositions**
- Better **generalization** than conventional rule learners

Desired Features

- Highly **interpretable** hypothesis
  - Small set of features/rules i.e., **low** q
  - *Simple* features e.g., **short conjunctive propositions**
- Better **generalization** than conventional rule learners

Desired Features

- Highly **interpretable** hypothesis
  - Small set of features/rules i.e., **low** q
  - *Simple* features e.g., **short conjunctive propositions**
- Better generalization than conventional rule learners

Desired Features

- Highly **interpretable** hypothesis
  - Small set of features/rules i.e., **low** q
  - *Simple* features e.g., **short conjunctive propositions**
- Better **generalization** than conventional rule learners

Formal Problem Definition

Input:

- Training Set: $\mathscr{D}=\{(\mathbf{x}^1,\mathbf{y}^1),\ldots,(\mathbf{x}^m,\mathbf{y}^m)\}$, $\mathbf{x}^i\in\mathbb{R}^n$ and $\mathbf{y}^i\in\mathscr{C}^r$
  - $\mathscr{C}$ is set of class labels. For binary: $\mathscr{C}\in\{-1,1\}$
  - If $r>1$, we are dealing with a sequence prediction problem
  - Initially $y_i\in\{-1,1\}$. That is, $\mathscr{C}\in\{-1,1\}$ and $r=1$.
- Basic propositions regarding input features (say, p in number)

  Nominal e.g., $x_i=a$ and $x_i\neq a$

  Numeric e.g., $x_j\geq b$ and $x_j\leq b$

Goal:

Construct interpretable features/rules from basic propositions

Formal Problem Definition

Input:

- Training Set: $\mathscr{D}=\{(\mathbf{x}^1,\mathbf{y}^1),\ldots,(\mathbf{x}^m,\mathbf{y}^m)\}$, $\mathbf{x}^i\in\mathbb{R}^n$ and $\mathbf{y}^i\in\mathscr{C}^r$
  - $\mathscr{C}$ is set of class labels. For binary: $\mathscr{C}\in\{-1,1\}$
  - If $r>1$, we are dealing with a sequence prediction problem
  - Initially $y_i\in\{-1,1\}$. That is, $\mathscr{C}\in\{-1,1\}$ and $r=1$.
- Basic propositions regarding input features (say, p in number)
  - Nominal e.g., $x_i = a$ and $x_i \neq a$
  - Numeric e.g., $x_j \geq b$ and $x_j \leq b$

Goal:

- Construct interpretable features/rules from basic propositions

Formal Problem Definition

Input:

- Training Set: $\mathcal{D}=\{(\mathbf{x}^1,\mathbf{y}^1),\ldots,(\mathbf{x}^m,\mathbf{y}^m)\}$, $\mathbf{x}^i\in\mathbb{R}^n$ and $\mathbf{y}^i\in\mathcal{C}^r$
  - $\mathcal{C}$ is set of class labels. For binary: $\mathcal{C}\in\{-1,1\}$
  - If $r>1$, we are dealing with a sequence prediction problem
  - Initially $y_i\in\{-1,1\}$. That is, $\mathcal{C}\in\{-1,1\}$ and $r=1$.
- Basic propositions regarding input features (say, p in number)
  - Nominal e.g., $x_i=a$ and $x_i\neq a$
  - Numeric e.g., $x_j\geq b$ and $x_j\leq b$

Goal:

Construct interpretable features/rules from basic propositions

Formal Problem Definition

Input:

- Training Set: $\mathscr{D}=\{(\mathbf{x}^1,\mathbf{y}^1),...,(\mathbf{x}^m,\mathbf{y}^m)\}$, $\mathbf{x}^i \in \mathbb{R}^n$ and $\mathbf{y}^i \in \mathscr{C}^r$
  - $\mathscr{C}$ is set of class labels. For binary: $\mathscr{C} \in \{-1,1\}$
  - If $r > 1$, we are dealing with a sequence prediction problem
  - Initially $y_i \in \{-1,1\}$. That is, $\mathscr{C} \in \{-1,1\}$ and $r = 1$.
- Basic propositions regarding input features (say, p in number)
  - Nominal e.g., $x_i = a$ and $x_i \neq a$
  - Numeric e.g., $x_j \geq b$ and $x_j \leq b$

Goal:

- Construct interpretable features/rules from basic propositions

Formal Problem Definition

Input:

- Training Set: $\mathscr{D}=\{(\mathbf{x}^1,\mathbf{y}^1),\ldots,(\mathbf{x}^m,\mathbf{y}^m)\}$, $\mathbf{x}^i\in\mathbb{R}^n$ and $\mathbf{y}^i\in\mathscr{C}^r$
  - $\mathscr{C}$ is set of class labels. For binary: $\mathscr{C}\in\{-1,1\}$
  - If $r>1$, we are dealing with a sequence prediction problem
  - Initially $y_i\in\{-1,1\}$. That is, $\mathscr{C}\in\{-1,1\}$ and $r=1$.
- Basic propositions regarding input features (say, p in number)

  Nominal e.g., $x_i=a$ and $x_i\neq a$

  Numeric e.g., $x_j\geq b$ and $x_j\leq b$

Goal:

Construct interpretable features/rules from basic propositions

Formal Problem Definition

Input:

- Training Set: $\mathcal{D}=\left\{(\mathbf{x}^1,\mathbf{y}^1),\dots,(\mathbf{x}^m,\mathbf{y}^m)\right\}$, $\mathbf{x}^i \in \mathbb{R}^n$ and $\mathbf{y}^i \in \mathcal{C}^r$
  - $\mathcal{C}$ is set of class labels. For binary: $\mathcal{C} \in \{-1,1\}$
  - If $r > 1$, we are dealing with a sequence prediction problem
  - Initially $y_i \in \{-1,1\}$. That is, $\mathcal{C} \in \{-1,1\}$ and $r = 1$.
- Basic propositions regarding input features (say, p in number)

  Nominal e.g., $x_i = a$ and $x_i \neq a$

  Numeric e.g., $x_j \geq b$ and $x_j \leq b$

Goal:

Construct interpretable features/rules from basic propositions

Formal Problem Definition

Input:

- Training Set: $\mathscr{D}=\{(\mathbf{x}^1,\mathbf{y}^1),\ldots,(\mathbf{x}^m,\mathbf{y}^m)\}$, $\mathbf{x}^i \in \mathbb{R}^n$ and $\mathbf{y}^i \in \mathscr{C}^r$
  - $\mathscr{C}$ is set of class labels. For binary: $\mathscr{C} \in \{-1,1\}$
  - If $r > 1$, we are dealing with a sequence prediction problem
  - Initially $y_i \in \{-1,1\}$. That is, $\mathscr{C} \in \{-1,1\}$ and $r = 1$.
- Basic propositions regarding input features (say, p in number)

  Nominal e.g., $x_i = a$ and $x_i \neq a$

  Numeric e.g., $x_j \geq b$ and $x_j \leq b$

Goal:

Construct interpretable features/rules from basic propositions

Formal Problem Definition

Input:

- Training Set: $\mathscr{D}=\{(\mathbf{x}^1,\mathbf{y}^1),\ldots,(\mathbf{x}^m,\mathbf{y}^m)\}$, $\mathbf{x}^i \in \mathbb{R}^n$ and $\mathbf{y}^i \in \mathscr{C}^r$
  - $\mathscr{C}$ is set of class labels. For binary: $\mathscr{C} \in \{-1,1\}$
  - If $r > 1$, we are dealing with a sequence prediction problem
  - Initially $y_i \in \{-1,1\}$. That is, $\mathscr{C} \in \{-1,1\}$ and $r = 1$.
- Basic propositions regarding input features (say, p in number)

  Nominal e.g., $x_i = a$ and $x_i \neq a$

  Numeric e.g., $x_j \geq b$ and $x_j \leq b$

Goal:

Construct interpretable features/rules from basic propositions

Formal Problem Definition

Input:

- Training Set: $\mathscr{D}=\{(\mathbf{x}^1,\mathbf{y}^1),\ldots,(\mathbf{x}^m,\mathbf{y}^m)\}$, $\mathbf{x}^i \in \mathbb{R}^n$ and $\mathbf{y}^i \in \mathscr{C}^r$
  - $\mathscr{C}$ is set of class labels. For binary: $\mathscr{C} \in \{-1, 1\}$
  - If $r > 1$, we are dealing with a sequence prediction problem
  - Initially $y_i \in \{-1, 1\}$. That is, $\mathscr{C} \in \{-1, 1\}$ and $r = 1$.
- Basic propositions regarding input features (say, p in number)

  Nominal e.g., $x_i = a$ and $x_i \neq a$

  Numeric e.g., $x_j \geq b$ and $x_j \leq b$

Goal:

- Construct interpretable features/rules from basic propositions

Formal Problem Definition

Input:

- Training Set: $\mathscr{D}=\left\{(\mathbf{x}^1,\mathbf{y}^1),\ldots,(\mathbf{x}^m,\mathbf{y}^m)\right\}$, $\mathbf{x}^i \in \mathbb{R}^n$ and $\mathbf{y}^i \in \mathscr{C}^r$
  - $\mathscr{C}$ is set of class labels. For binary: $\mathscr{C} \in \{-1,1\}$
  - If $r > 1$, we are dealing with a sequence prediction problem
  - Initially $y_i \in \{-1,1\}$. That is, $\mathscr{C} \in \{-1,1\}$ and $r = 1$.
- Basic propositions regarding input features (say, p in number)

  Nominal e.g., $x_i = a$ and $x_i \neq a$
  Numeric e.g., $x_j \geq b$ and $x_j \leq b$

Goal:

- Construct interpretable features/rules from basic propositions
  - Few in number
  - Short conjunctions
- Compute corresponding weights (w, b)

Formal Problem Definition

Input:

- Training Set: $\mathscr{D}=\{(\mathbf{x}^1,\mathbf{y}^1),\ldots,(\mathbf{x}^m,\mathbf{y}^m)\}$, $\mathbf{x}^i \in \mathbb{R}^n$ and $\mathbf{y}^i \in \mathscr{C}^r$
  - $\mathscr{C}$ is set of class labels. For binary: $\mathscr{C} \in \{-1,1\}$
  - If $r > 1$, we are dealing with a sequence prediction problem
  - Initially $y_i \in \{-1,1\}$. That is, $\mathscr{C} \in \{-1,1\}$ and $r = 1$.
- Basic propositions regarding input features (say, p in number)

  Nominal e.g., $x_i = a$ and $x_i \neq a$
  Numeric e.g., $x_j \geq b$ and $x_j \leq b$

Goal:

- Construct interpretable features/rules from basic propositions
  - **Few** in number
  - **Short** conjunctions
- Compute corresponding **weights** (w, b)

Formal Problem Definition

Input:

- Training Set: $\mathscr{D}=\{(\mathbf{x}^1,\mathbf{y}^1),...,(\mathbf{x}^m,\mathbf{y}^m)\}$, $\mathbf{x}^i \in \mathbb{R}^n$ and $\mathbf{y}^i \in \mathscr{C}^r$
  - $\mathscr{C}$ is set of class labels. For binary: $\mathscr{C} \in \{-1, 1\}$
  - If $r > 1$, we are dealing with a sequence prediction problem
  - Initially $y_i \in \{-1, 1\}$. That is, $\mathscr{C} \in \{-1, 1\}$ and $r = 1$.
- Basic propositions regarding input features (say, p in number)

  Nominal e.g., $x_i = a$ and $x_i \neq a$

  Numeric e.g., $x_j \geq b$ and $x_j \leq b$

Goal:

- Construct interpretable features/rules from basic propositions
  - **Few** in number
  - **Short** conjunctions
- Compute corresponding **weights** (w, b)

Formal Problem Definition

Input:

- Training Set: $\mathscr{D}=\left\{(\mathbf{x}^1,\mathbf{y}^1),\ldots,(\mathbf{x}^m,\mathbf{y}^m)\right\}$, $\mathbf{x}^i\in\mathbb{R}^n$ and $\mathbf{y}^i\in\mathscr{C}^r$
  - $\mathscr{C}$ is set of class labels. For binary: $\mathscr{C}\in\{-1,1\}$
  - If $r>1$, we are dealing with a sequence prediction problem
  - Initially $y_i\in\{-1,1\}$. That is, $\mathscr{C}\in\{-1,1\}$ and $r=1$.
- Basic propositions regarding input features (say, p in number)

  Nominal e.g., $x_i = a$ and $x_i \neq a$
  Numeric e.g., $x_j \geq b$ and $x_j \leq b$

Goal:

- Construct interpretable features/rules from basic propositions
  - **Few** in number
  - **Short** conjunctions
- Compute corresponding **weights** (w, b)

Formal Problem Definition

Input:

- Training Set: $\mathscr{D}=\{(\mathbf{x}^1,\mathbf{y}^1),...,(\mathbf{x}^m,\mathbf{y}^m)\}$, $\mathbf{x}^i\in\mathbb{R}^n$ and $\mathbf{y}^i\in\mathscr{C}^r$
  - $\mathscr{C}$ is set of class labels. For binary: $\mathscr{C} \in \{-1, 1\}$
  - If $r > 1$, we are dealing with a sequence prediction problem
  - Initially $y_i \in \{-1, 1\}$. That is, $\mathscr{C} \in \{-1, 1\}$ and $r = 1$.
- Basic propositions regarding input features (say, p in number)

  Nominal e.g., $x_i = a$ and $x_i \neq a$

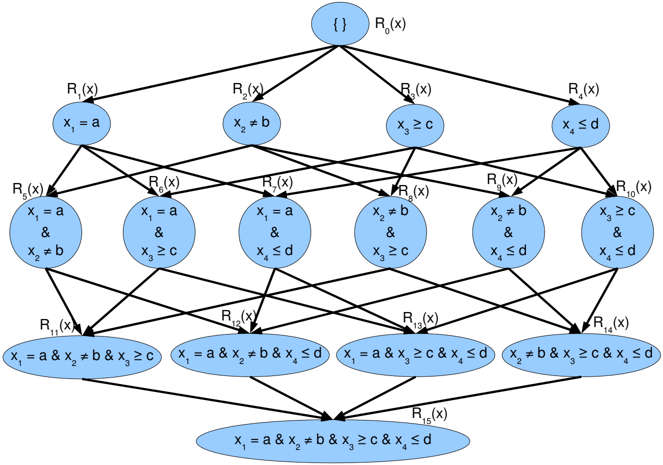
  Numeric e.g., $x_j \geq b$ and $x_j \leq b$

Goal:

- Construct interpretable features/rules from basic propositions
  - **Few** in number
  - **Short** conjunctions
- Compute corresponding **weights** $(\mathbf{w}, b)$

Challenge:

Extremely **large** search space over features! Atleast $O(2^n)$ (conjunctive/disjunctive features)



Existing Methods: Greedy and/or suboptimal

SLIPPER(Cohen&Singer, 99): AdaBoost + RIPPER — greedy

RuleFit(Friedman&Popescu, 08): ISLE + decision tree — greedy

ELCS(Gao et.al., 07): Genetic Alg. + post-pruning — sub-optimal

ENDER(Dembczynski et.al., 10): Minimization of empirical risk — greedy

Existing Methods: Greedy and/or suboptimal

SLIPPER(Cohen&Singer, 99): AdaBoost + RIPPER — **greedy**

RuleFit(Friedman&Popescu, 08): ISLE + decision tree — **greedy**

ELCS(Gao et.al., 07): Genetic Alg. + post-pruning — **sub-optimal**

ENDER(Dembczynski et.al., 10): Minimization of empirical risk — **greedy**

Existing Methods: Greedy and/or suboptimal

SLIPPER(Cohen&Singer, 99): AdaBoost + RIPPER — **greedy**

RuleFit(Friedman&Popescu, 08): ISLE + decision tree — **greedy**

ELCS(Gao et.al., 07): Genetic Alg. + post-pruning — **sub-optimal**

ENDER(Dembczynski et.al., 10): Minimization of empirical risk — **greedy**

Existing Methods: Greedy and/or suboptimal

SLIPPER(Cohen&Singer, 99): AdaBoost + RIPPER — **greedy**

RuleFit(Friedman&Popescu, 08): ISLE + decision tree — **greedy**

ELCS(Gao et.al., 07): Genetic Alg. + post-pruning — **sub-optimal**

ENDER(Dembczynski et.al., 10): Minimization of empirical risk — **greedy**

Existing Methods: Greedy and/or suboptimal

SLIPPER(Cohen&Singer, 99): AdaBoost + RIPPER — **greedy**

RuleFit(Friedman&Popescu, 08): ISLE + decision tree — **greedy**

ELCS(Gao et.al., 07): Genetic Alg. + post-pruning — **sub-optimal**

ENDER(Dembczynski et.al., 10): Minimization of empirical risk — **greedy**

Existing Methods: Greedy and/or suboptimal

SLIPPER(Cohen&Singer, 99): AdaBoost + RIPPER — **greedy**

RuleFit(Friedman&Popescu, 08): ISLE + decision tree — **greedy**

ELCS(Gao et.al., 07): Genetic Alg. + post-pruning — **sub-optimal**

ENDER(Dembczynski et.al., 10): Minimization of empirical risk — **greedy**

Proposed Methodology — Overview

Optimal search for rules over **all** conjunctions

- **Regularized** loss minimization
- **Convex** formulation
- Discovers **compact** ruleset (small set with short rules)

Technical Contribution:

Efficient mirror-descent based active set method

Key Structure Exploited:

Sub-lattices with long features are discouraged.

Proposed Methodology — Overview

Optimal search for rules over **all** conjunctions

- **Regularized** loss minimization
- **Convex** formulation
- Discovers **compact** ruleset (small set with short rules)

Technical Contribution:

Efficient mirror-descent based active set method

Key Structure Exploited:

Sub-lattices with long features are discouraged.

Proposed Methodology — Overview

Optimal search for rules over **all** conjunctions

- **Regularized** loss minimization
- **Convex** formulation
- Discovers **compact** ruleset (small set with short rules)

Technical Contribution:

Efficient mirror-descent based active set method

Key Structure Exploited:

Sub-lattices with long features are discouraged.

Proposed Methodology — Overview

Optimal search for rules over **all** conjunctions

- **Regularized** loss minimization
- **Convex** formulation
- Discovers **compact** ruleset (small set with short rules)

Technical Contribution:

Efficient mirror-descent based active set method

Key Structure Exploited:

Sub-lattices with long features are discouraged.

Proposed Methodology — Overview

Optimal search for rules over **all** conjunctions

- **Regularized** loss minimization
- **Convex** formulation
- Discovers **compact** ruleset (small set with short rules)

Technical Contribution:

Efficient mirror-descent based active set method

- Complexity: **polynomial** in active set size ($\ll 2^p$)

Key Structure Exploited:

Sub-lattices with long features are discouraged.

Proposed Methodology — Overview

Optimal search for rules over **all** conjunctions

- **Regularized** loss minimization
- **Convex** formulation
- Discovers **compact** ruleset (small set with short rules)

Technical Contribution:

Efficient mirror-descent based active set method

- Complexity: **polynomial** in active set size ($\ll 2^p$)

Key Structure Exploited:

Sub-lattices with long features are discouraged.

Proposed Methodology — Overview

Optimal search for rules over **all** conjunctions

- **Regularized** loss minimization
- **Convex** formulation
- Discovers **compact** ruleset (small set with short rules)

Technical Contribution:

Efficient mirror-descent based active set method

- Complexity: **polynomial** in active set size ($\ll 2^p$)

Key Structure Exploited:

Sub-lattices with long features are discouraged.

Proposed Methodology — Overview

Optimal search for rules over **all** conjunctions

- **Regularized** loss minimization
- **Convex** formulation
- Discovers **compact** ruleset (small set with short rules)

Technical Contribution:

Efficient mirror-descent based active set method

- Complexity: **polynomial** in active set size ($\ll 2^p$)

Key Structure Exploited:

Sub-lattices with **long features are discouraged**.

A Primitive Formulation

- Decision function[1]: sign $\left(\sum_{v\in\mathcal{V}} w_v R_v(\mathbf{x}) - b\right)$
- l_1 regularizer to force many w_v to zero

l_1 regularized formulation:

$$\min_{\mathbf{w},b} \frac{1}{2}\left(\sum_{v\in\mathcal{V}} |w_v|\right)^2 + C\sum_{i=1}^{m} L\left(y^i, \sum_{v\in\mathcal{V}} w_v R_v(\mathbf{x}^i) - b\right)$$

A Primitive Formulation

- Decision function[1]: sign $\left(\sum_{v \in \mathcal{V}} w_v R_v(\mathbf{x}) - b\right)$
- l_1 regularizer to force many w_v to zero

l_1 regularized formulation:

$$\min_{\mathbf{w},b} \frac{1}{2}\left(\sum_{v \in \mathcal{V}} |w_v|\right)^2 + C\sum_{i=1}^{m} L\left(y^i, \sum_{v \in \mathcal{V}} w_v R_v(\mathbf{x}^i) - b\right)$$

A Primitive Formulation

- Decision function[1]: sign $\left(\sum_{v\in\mathcal{V}} w_v R_v(\mathbf{x}) - b\right)$
- l_1 regularizer to force many w_v to zero

l_1 regularized formulation:

$$\min_{\mathbf{w},b} \frac{1}{2}\left(\sum_{v\in\mathcal{V}} |w_v|\right)^2 + C\sum_{i=1}^{m} L\left(y^i, \sum_{v\in\mathcal{V}} w_v R_v(\mathbf{x}^i) - b\right)$$

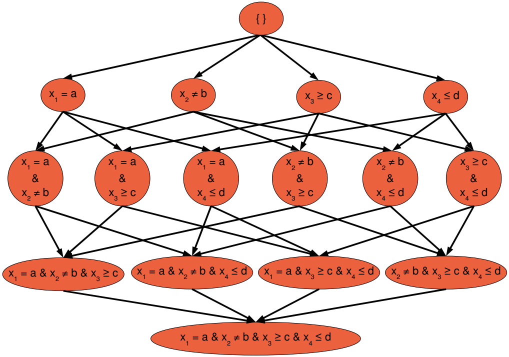
A Primitive Formulation

- Decision function[1]: sign $\left(\sum_{v \in \mathcal{V}} w_v R_v(\mathbf{x}) - b\right)$
- l_1 regularizer to force many w_v to zero

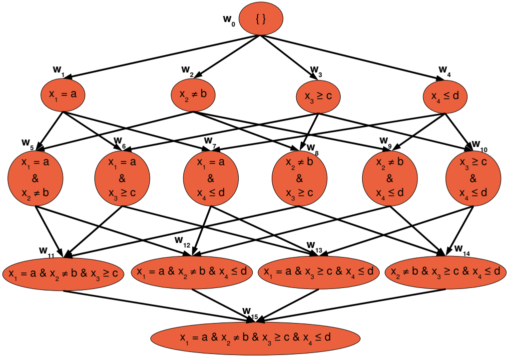
l_1 regularized formulation:

$$\min_{\mathbf{w},b} \frac{1}{2}\left(\sum_{v \in \mathcal{V}} |w_v|\right)^2 + C\sum_{i=1}^{m} L\left(y^i, \sum_{v \in \mathcal{V}} w_v R_v(\mathbf{x}^i) - b\right)$$

A Primitive Formulation



A Primitive Formulation

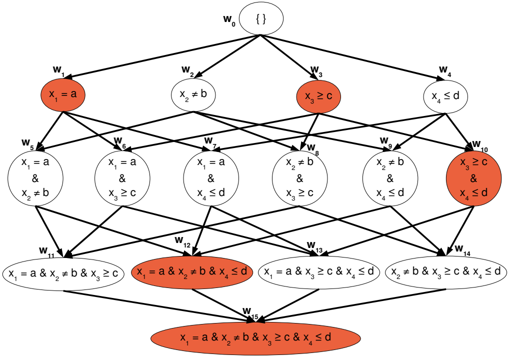


Short-comings:

may be selected

Computationally problem

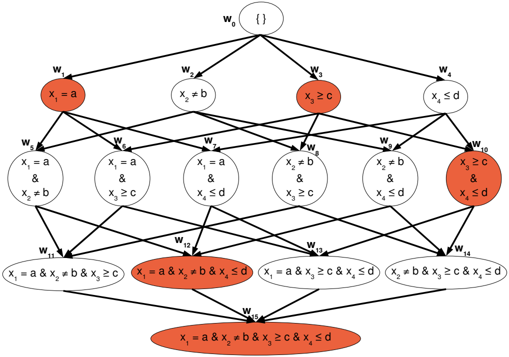
A Primitive Formulation



Short-comings:

- ... may be selected
- Computationally ... problem

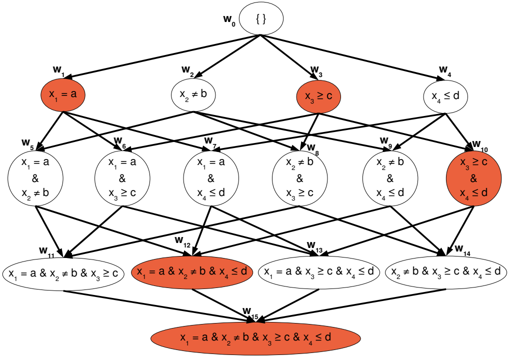
A Primitive Formulation



Short-comings:

- **long rules** may be selected
- Computationally **difficult** problem

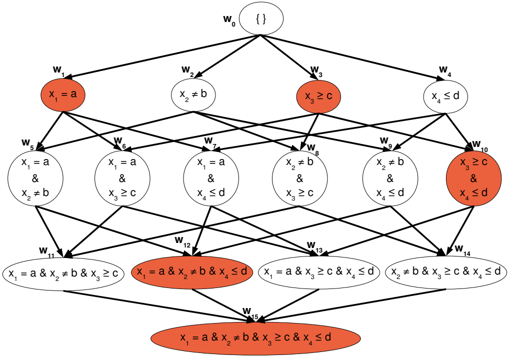
A Primitive Formulation



Short-comings:

- **long rules** may be selected
- Computationally **difficult** problem

A Primitive Formulation



Short-comings:

- **long rules** may be selected
- Computationally **difficult** problem

An Improved Formulation

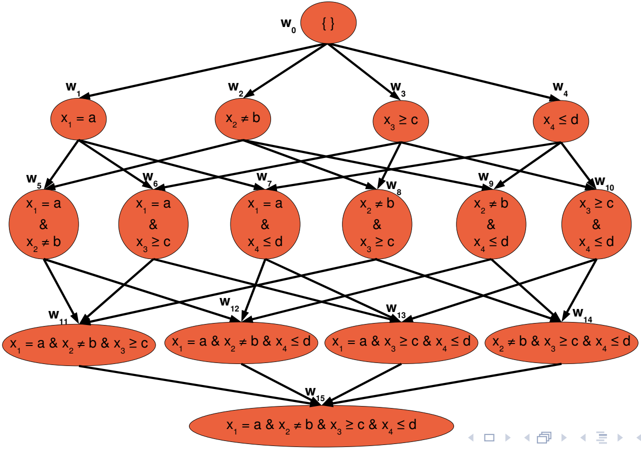
Key Idea:

Block l_1 regularizer discourages long rules: $\left(\sum_{v \in \mathcal{V}} \|\mathbf{w}_{D(v)}\|_2\right)^2$

An Improved Formulation

Key Idea:

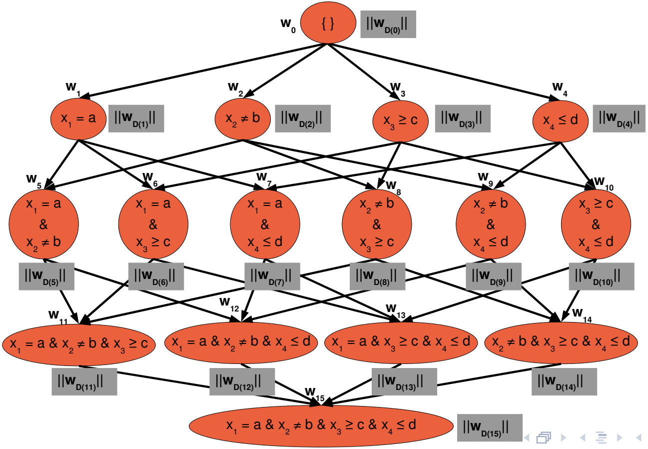
Block l_1 regularizer discourages long rules: $\left(\sum_{v \in \mathscr{V}} \|\mathbf{w}_{D(v)}\|_2\right)^2$



An Improved Formulation

Key Idea:

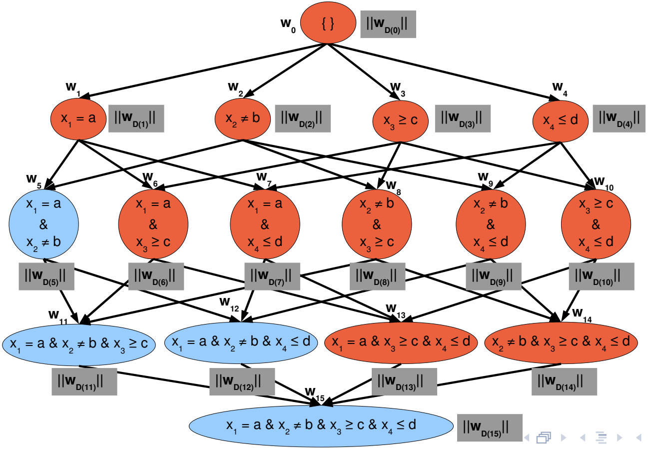
Block l_1 regularizer discourages long rules: $\left(\sum_{v\in\mathcal{V}} \|\mathbf{w}_{D(v)}\|_2\right)^2$



An Improved Formulation

Key Idea:

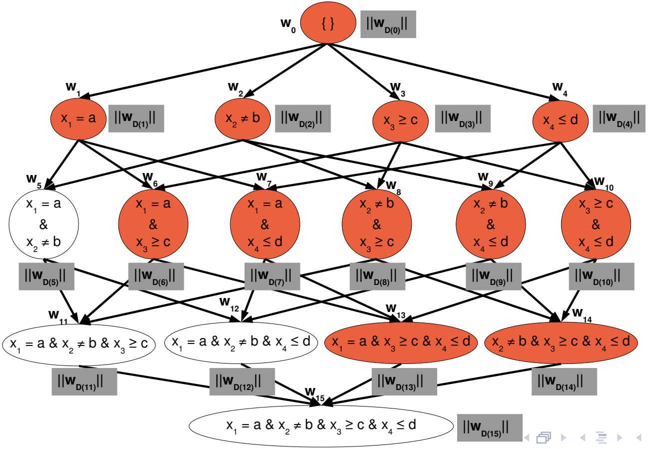
Block l_1 regularizer discourages long rules: $\left(\sum_{v \in \mathscr{V}} \|\mathbf{w}_{D(v)}\|_2\right)^2$



An Improved Formulation

Key Idea:

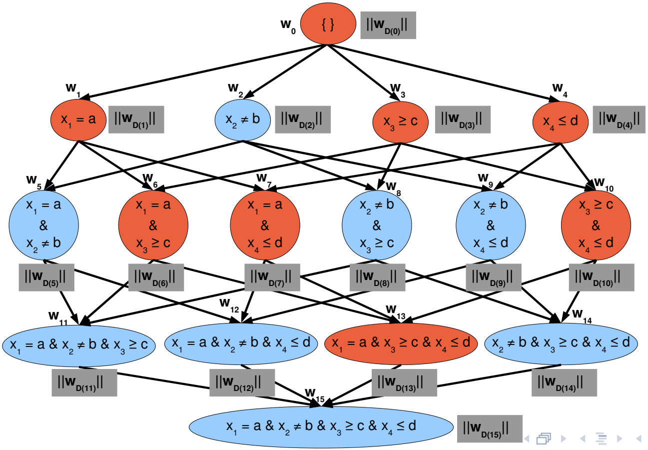
Block l_1 regularizer discourages long rules: $\left(\sum_{v \in \mathcal{V}} \|\mathbf{w}_{D(v)}\|_2\right)^2$



An Improved Formulation

Key Idea:

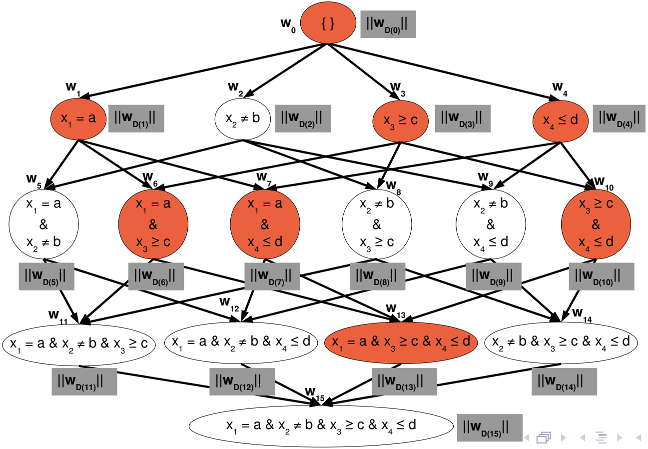
Block l_1 regularizer discourages long rules: $\left(\sum_{v \in \mathcal{V}} \|\mathbf{w}_{D(v)}\|_2\right)^2$



An Improved Formulation

Key Idea:

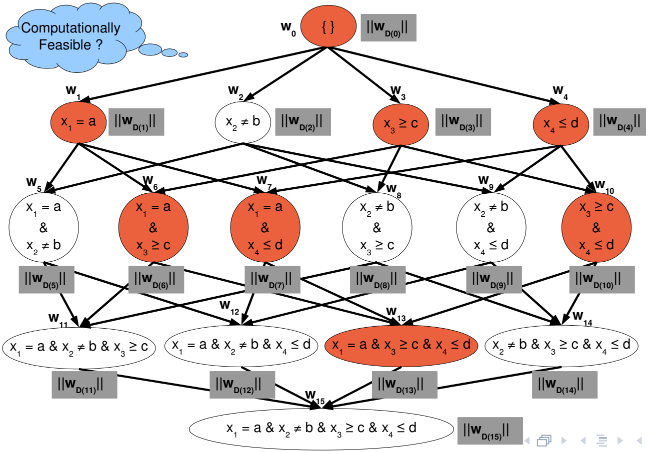
Block l_1 regularizer discourages long rules: $\left(\sum_{v \in \mathscr{V}} \|\mathbf{w}_{D(v)}\|_2\right)^2$



An Improved Formulation

Key Idea:

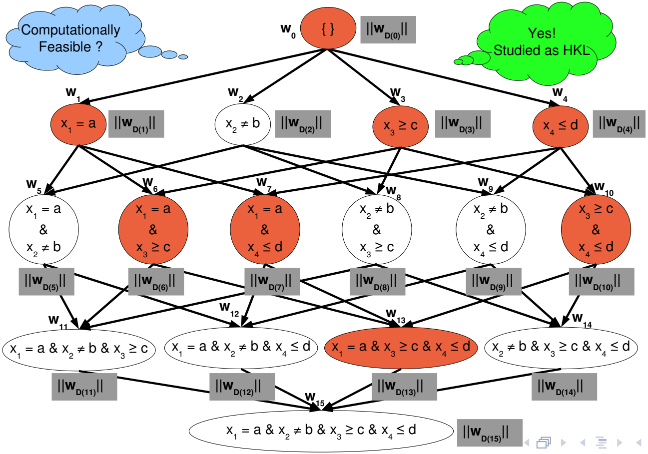
Block l_1 regularizer discourages long rules: $\left(\sum_{v \in \mathcal{V}} \|\mathbf{w}_{D(v)}\|_2\right)^2$



An Improved Formulation

Key Idea:

Block l_1 regularizer discourages long rules: $\left(\sum_{v \in \mathcal{V}} \|\mathbf{w}_{D(v)}\|_2\right)^2$



An Improved Formulation

Key Idea:

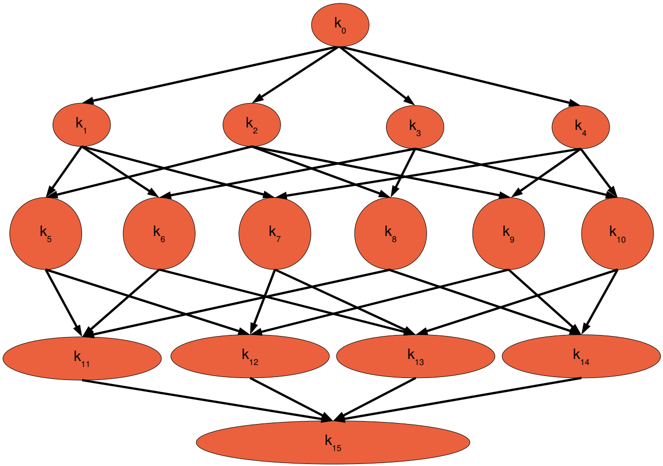
Block l_1 regularizer discourages long rules: $\left(\sum_{v \in \mathcal{V}} \|\mathbf{w}_{D(v)}\|_2\right)^2$

Hierarchical Kernel Learning (HKL)(Bach, 08)

- Kernels arranged on DAG (lattice) are given
- Optimal combination of kernels (Multiple Kernel Learning)

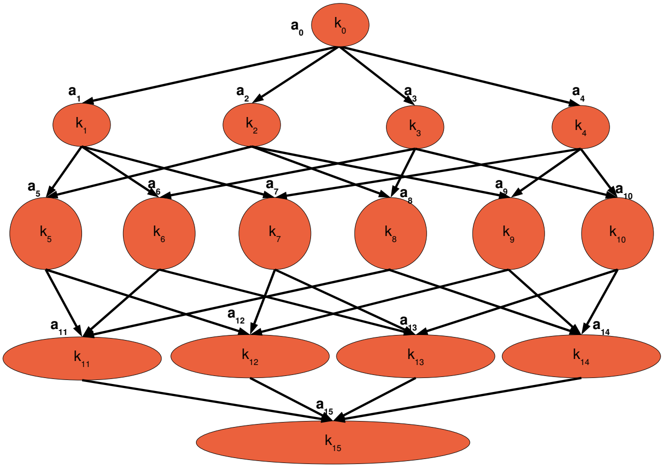
Hierarchical Kernel Learning (HKL)(Bach, 08)

- Kernels arranged on DAG (lattice) are given
- Optimal combination of kernels (Multiple Kernel Learning)



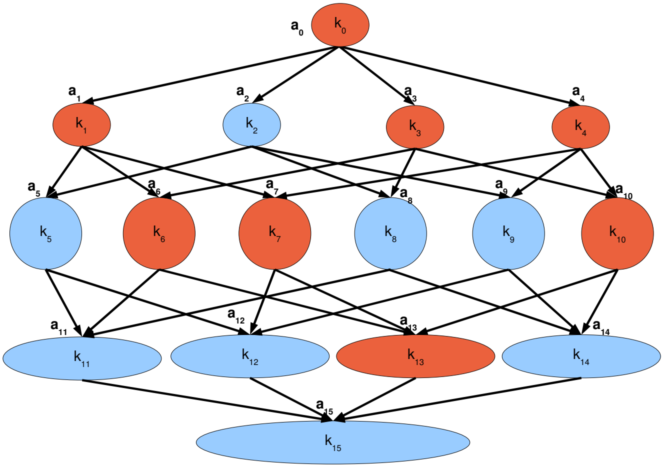
Hierarchical Kernel Learning (HKL)(Bach, 08)

- Kernels arranged on DAG (lattice) are given
- Optimal combination of kernels (Multiple Kernel Learning)



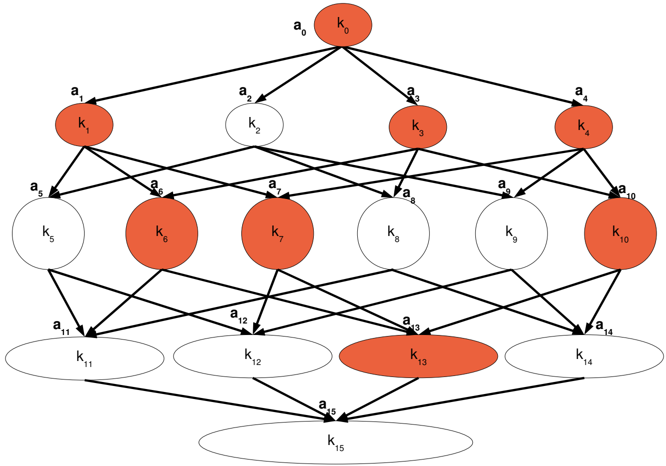
Hierarchical Kernel Learning (HKL)(Bach, 08)

- Kernels arranged on DAG (lattice) are given
- Optimal combination of kernels (Multiple Kernel Learning)



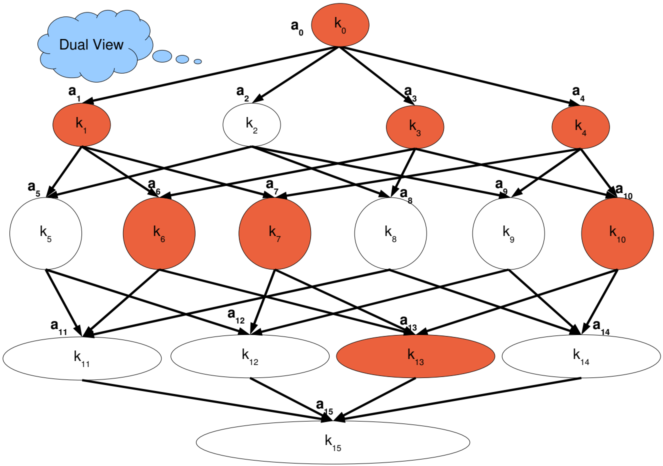
Hierarchical Kernel Learning (HKL)(Bach, 08)

- Kernels arranged on DAG (lattice) are given
- Optimal combination of kernels (Multiple Kernel Learning)



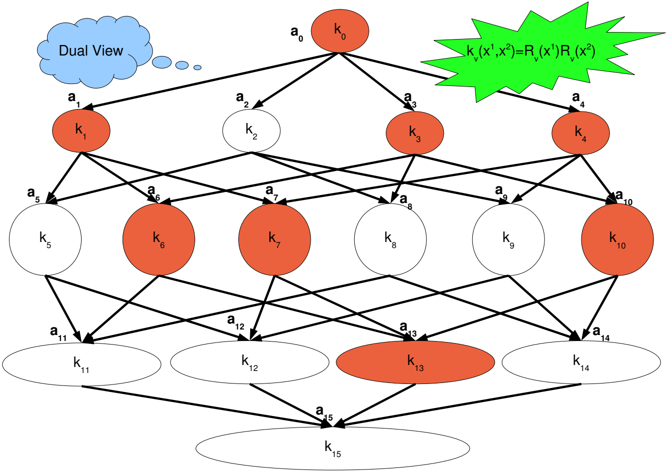
Hierarchical Kernel Learning (HKL)(Bach, 08)

- Kernels arranged on DAG (lattice) are given
- Optimal combination of kernels (Multiple Kernel Learning)



Hierarchical Kernel Learning (HKL)(Bach, 08)

- Kernels arranged on DAG (lattice) are given
- Optimal combination of kernels (Multiple Kernel Learning)



Hierarchical Kernel Learning (HKL)(Bach, 08)

- Kernels arranged on DAG (lattice) are given
- Optimal combination of kernels (Multiple Kernel Learning)

Hierarchical Kernel Learning (HKL)(Bach, 08)

- Kernels arranged on DAG (lattice) are given
- Optimal combination of kernels (Multiple Kernel Learning)

HKL — Key Result

Active Set Algorithm:

- Complexity: **Polynomial** in number of selected kernels
- Condition: kernels are summable in *linear* time over a sub-lattice

Our case:

- Kernels indeed easily summable

HKL — Key Result

Active Set Algorithm:

- Complexity: **Polynomial** in number of selected kernels
- Condition: kernels are summable in *linear* time over a sub-lattice

Our case:

- Kernels indeed easily summable

HKL — Key Result

Active Set Algorithm:

- Complexity: **Polynomial** in number of selected kernels
- Condition: kernels are summable in *linear* time over a sub-lattice

Our case:

- Kernels indeed easily summable

HKL — Key Result

Active Set Algorithm:

- Complexity: **Polynomial** in number of selected kernels
- Condition: kernels are summable in *linear* time over a sub-lattice

Our case:

- Kernels indeed easily summable

HKL — Key Result

Active Set Algorithm:

- Complexity: **Polynomial** in number of selected kernels
- Condition: kernels are summable in *linear* time over a sub-lattice

Our case:

- Kernels indeed easily summable
  - R_v is nothing but product of few base proposition evaluations
  - Sum of exponential no. terms = Product of linear no. terms
  - E.g., $1 + R_1 + R_2 + R_1 R_2 = (1 + R_1)(1 + R_2)$
  - Our problem can be solved in reasonable time

HKL — Key Result

Active Set Algorithm:

- Complexity: **Polynomial** in number of selected kernels
- Condition: kernels are summable in *linear* time over a sub-lattice

Our case:

- Kernels indeed easily summable
  - R_v is nothing but product of few base proposition evaluations
  - Sum of exponential no. terms = Product of linear no. terms
  - E.g., $1 + R_1 + R_2 + R_1 R_2 = (1 + R_1)(1 + R_2)$
  - Our problem can be solved in reasonable time

HKL — Key Result

Active Set Algorithm:

- Complexity: **Polynomial** in number of selected kernels
- Condition: kernels are summable in *linear* time over a sub-lattice

Our case:

- Kernels indeed easily summable
  - R_v is nothing but product of few base proposition evaluations
  - Sum of exponential no. terms = Product of linear no. terms
  - E.g., $1 + R_1 + R_2 + R_1 R_2 = (1 + R_1)(1 + R_2)$
  - Our problem can be solved in reasonable time

HKL — Key Result

Active Set Algorithm:

- Complexity: **Polynomial** in number of selected kernels
- Condition: kernels are summable in *linear* time over a sub-lattice

Our case:

- Kernels indeed easily summable
  - R_v is nothing but product of few base proposition evaluations
  - Sum of exponential no. terms = Product of linear no. terms
  - E.g., $1 + R_1 + R_2 + R_1R_2 = (1 + R_1)(1 + R_2)$
  - Our problem can be solved in reasonable time

HKL — Key Result

Active Set Algorithm:

- Complexity: **Polynomial** in number of selected kernels
- Condition: kernels are summable in *linear* time over a sub-lattice

Our case:

- Kernels indeed easily summable
  - R_v is nothing but product of few base proposition evaluations
  - Sum of exponential no. terms = Product of linear no. terms
  - E.g., $1 + R_1 + R_2 + R_1 R_2 = (1 + R_1)(1 + R_2)$
  - Our problem can be solved in reasonable time

HKL — Key Result

Active Set Algorithm:

- Complexity: **Polynomial** in number of selected kernels
- Condition: kernels are summable in *linear* time over a sub-lattice

Our case:

- Kernels indeed easily summable
  - R_v is nothing but product of few base proposition evaluations
  - Sum of exponential no. terms = Product of linear no. terms
  - E.g., $1+R_1+R_2+R_1R_2=(1+R_1)(1+R_2)$
  - Our problem can be solved in reasonable time

Performance Comparison

Dataset	RuleFit	SLI	ENDER	HKL
TIC-TAC-TOE	0.652 ± 0.068 (40, 2.51)	0.747 ± 0.026 (59, 2.35)	0.633 ± 0.011 (111, 2.46)	**0.889** ± 0.029 (129, **1.85**)
BALANCE	0.835 ± 0.034 (17, 2.18)	0.856 ± 0.027 (25, 1.88)	0.827 ± 0.013 (64, 1.99)	**0.893** ± 0.027 (65, **1.65**)
HABERMAN	0.512 ± 0.072 (6, 1.68)	0.565 ± 0.066 (8, **1.14**)	0.424 ± 0.000 (18, 1.87)	**0.594** ± 0.056 (32, 1.27)
CAR	0.913 ± 0.033 (34, 3.12)	0.895 ± 0.024 (141, 2.27)	0.755 ± 0.028 (80, 1.85)	**0.943** ± 0.024 (87, **1.78**)
BLOOD TRANS.	0.549 ± 0.092 (18, 1.99)	0.559 ± 0.100 (6, **1.07**)	0.489 ± 0.054 (58, 1.5)	**0.594** ± 0.009 (242, 1.64)
CMC	0.632 ± 0.013 (39, 2.41)	0.601 ± 0.041 (13, 2.13)	0.644 ± 0.026 (74, 2.65)	**0.656** ± 0.014 (127, **1.96**)

Performance Comparison

Dataset	RuleFit	SLI	ENDER	HKL
TIC-TAC-TOE	0.652 ± 0.068 (40, 2.51)	0.747 ± 0.026 (59, 2.35)	0.633 ± 0.011 (111, 2.46)	$\mathbf{0.889} \pm 0.029$ (129, **1.85**)
BALANCE	0.835 ± 0.034 (17, 2.18)	0.856 ± 0.027 (25, 1.88)	0.827 ± 0.013 (64, 1.99)	$\mathbf{0.893} \pm 0.027$ (65, **1.65**)
HABERMAN	0.512 ± 0.072 (6, 1.68)	0.565 ± 0.066 (8, **1.14**)	0.424 ± 0.000 (18, 1.87)	$\mathbf{0.594} \pm 0.056$ (32, 1.27)
CAR	0.913 ± 0.033 (34, 3.12)	0.895 ± 0.024 (141, 2.27)	0.755 ± 0.028 (80, 1.85)	$\mathbf{0.943} \pm 0.024$ (87, **1.78**)
BLOOD TRANS.	0.549 ± 0.092 (18, 1.99)	0.559 ± 0.100 (6, **1.07**)	0.489 ± 0.054 (58, 1.5)	$\mathbf{0.594} \pm 0.009$ (242, 1.64)
CMC	0.632 ± 0.013 (39, 2.41)	0.601 ± 0.041 (13, 2.13)	0.644 ± 0.026 (74, 2.65)	$\mathbf{0.656} \pm 0.014$ (127, **1.96**)

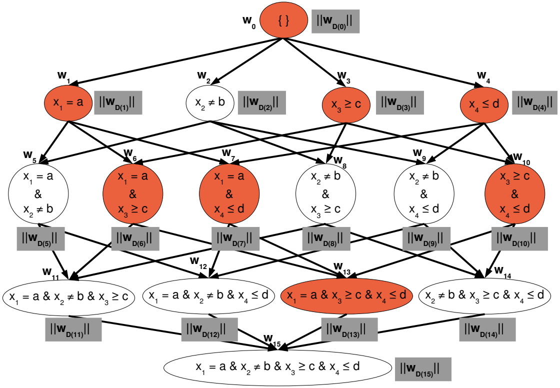
Performance Comparison

Dataset	RuleFit	SLI	ENDER	HKL
TIC-TAC-TOE	0.652 ± 0.068 (**40**, 2.51)	0.747 ± 0.026 (59, 2.35)	0.633 ± 0.011 (111, 2.46)	**0.889** ± 0.029 (129, **1.85**)
BALANCE	0.835 ± 0.034 (**17**, 2.18)	0.856 ± 0.027 (25, 1.88)	0.827 ± 0.013 (64, 1.99)	**0.893** ± 0.027 (65, **1.65**)
HABERMAN	0.512 ± 0.072 (**6**, 1.68)	0.565 ± 0.066 (8, **1.14**)	0.424 ± 0.000 (18, 1.87)	**0.594** ± 0.056 (32, 1.27)
CAR	0.913 ± 0.033 (**34**, 3.12)	0.895 ± 0.024 (141, 2.27)	0.755 ± 0.028 (80, 1.85)	**0.943** ± 0.024 (87, **1.78**)
BLOOD TRANS.	0.549 ± 0.092 (18, 1.99)	0.559 ± 0.100 (**6**, **1.07**)	0.489 ± 0.054 (58, 1.5)	**0.594** ± 0.009 (242, 1.64)
CMC	0.632 ± 0.013 (39, 2.41)	0.601 ± 0.041 (**13**, 2.13)	0.644 ± 0.026 (74, 2.65)	**0.656** ± 0.014 (127, **1.96**)

Performance Comparison

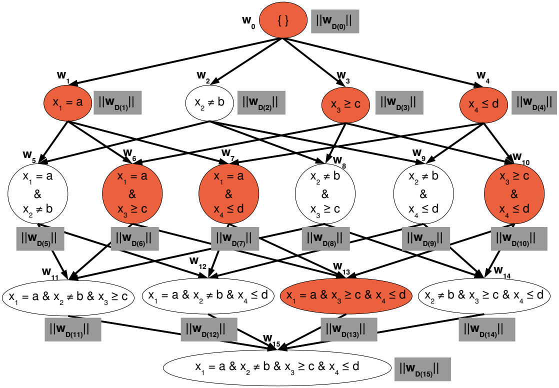
Dataset	RuleFit	SLI	ENDER	HKL
TIC-TAC-TOE	0.652 ± 0.068 (**40**, 2.51)	0.747 ± 0.026 (59, 2.35)	0.633 ± 0.011 (111, 2.46)	$\mathbf{0.889} \pm 0.029$ (**129**, **1.85**)
BALANCE	0.835 ± 0.034 (**17**, 2.18)	0.856 ± 0.027 (25, 1.88)	0.827 ± 0.013 (64, 1.99)	$\mathbf{0.893} \pm 0.027$ (65, **1.65**)
HABERMAN	0.512 ± 0.072 (**6**, 1.68)	0.565 ± 0.066 (8, **1.14**)	0.424 ± 0.000 (18, 1.87)	$\mathbf{0.594} \pm 0.056$ (**32**, 1.27)
CAR	0.913 ± 0.033 (**34**, 3.12)	0.895 ± 0.024 (141, 2.27)	0.755 ± 0.028 (80, 1.85)	$\mathbf{0.943} \pm 0.024$ (87, **1.78**)
BLOOD TRANS.	0.549 ± 0.092 (18, 1.99)	0.559 ± 0.100 (**6**, **1.07**)	0.489 ± 0.054 (58, 1.5)	$\mathbf{0.594} \pm 0.009$ (**242**, 1.64)
CMC	0.632 ± 0.013 (39, 2.41)	0.601 ± 0.041 (**13**, 2.13)	0.644 ± 0.026 (74, 2.65)	$\mathbf{0.656} \pm 0.014$ (**217**, **1.96**)

HKL — Introspection



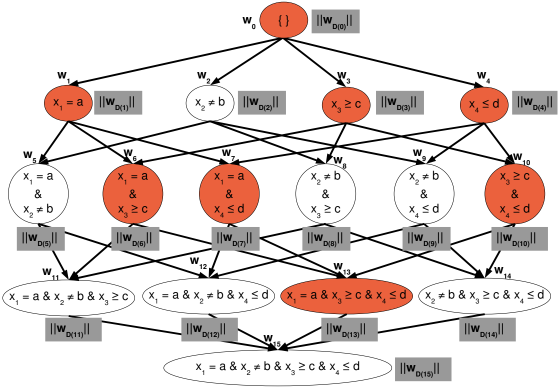
- Node selected only if all its ancesters are!
- l_1 promotes sparsity.
- l_2 promotes non-sparsity. Employ sparsity inducing norm!

HKL — Introspection



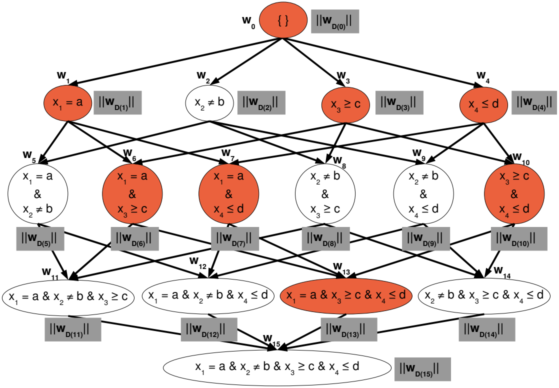
- Node selected **only** if all its ancesters are!
- l_1 promotes sparsity.
- l_2 promotes non-sparsity. Employ sparsity inducing norm!

HKL — Introspection



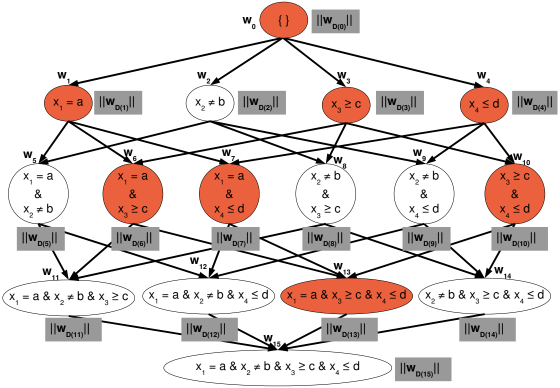
- Node selected **only** if all its ancesters are!
- l_1 promotes sparsity.
- l_2 promotes non-sparsity. **Employ sparsity inducing norm!**

HKL — Introspection



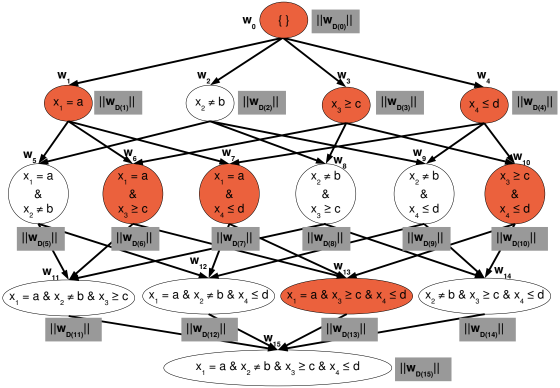
- Node selected **only** if all its ancesters are!
- l_1 promotes sparsity.
- l_2 promotes non-sparsity. **Employ sparsity inducing norm!**

HKL — Introspection



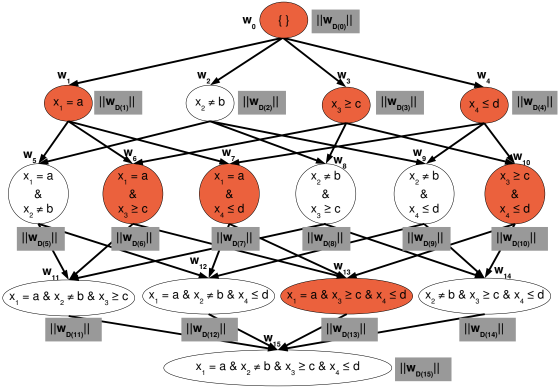
- Node selected **only** if all its ancesters are!
- l_1 promotes sparsity.
- l_2 promotes non-sparsity. **Employ sparsity inducing norm!**

HKL — Introspection



- Node selected **only** if all its ancesters are!
- l_1 promotes sparsity.
- l_2 promotes non-sparsity. **Employ sparsity inducing norm!**

HKL — Introspection



- Node selected **only** if all its ancesters are!
- l_1 promotes sparsity.
- l_2 promotes non-sparsity. **Employ sparsity inducing norm!**

Proposed Formulation

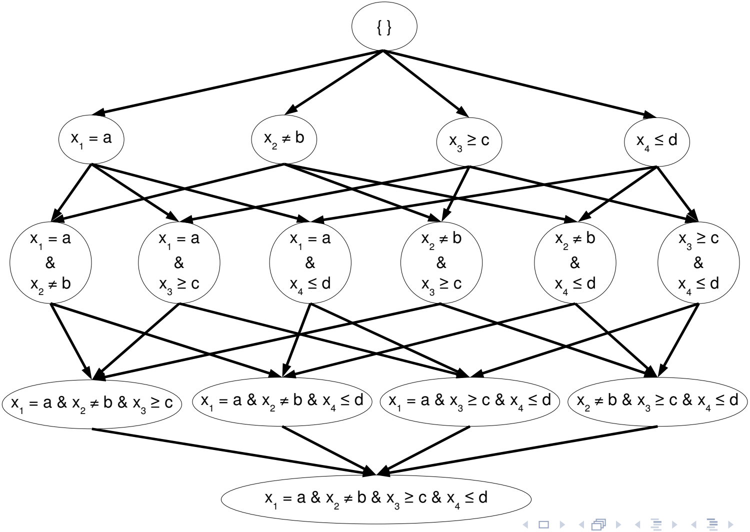
Generalized HKL

$$\min_{\mathbf{w},b} \frac{1}{2}\left(\sum_{v\in\mathcal{V}} d_v \|\mathbf{w}_{D(v)}\|_\rho\right)^2 + C\sum_{i=1}^{m} L\left(y^i, \sum_{v\in\mathcal{V}} w_v R_v(\mathbf{x}^i) - b\right)$$

where $1 < \rho \leq 2$.

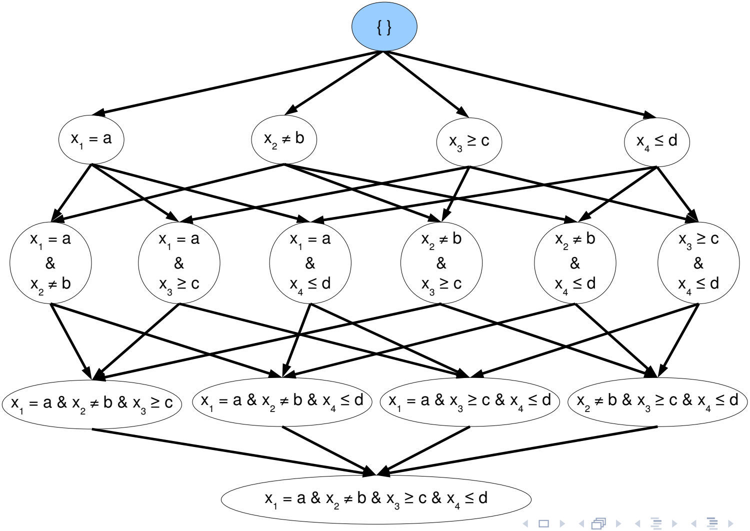
Active Set Method

p



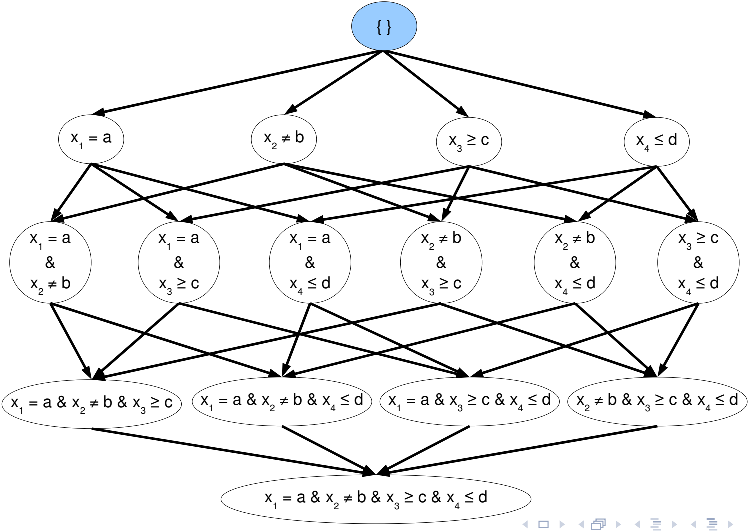
Active Set Method

Initialize active set with root node ($\mathscr{W} = \{0\}$).



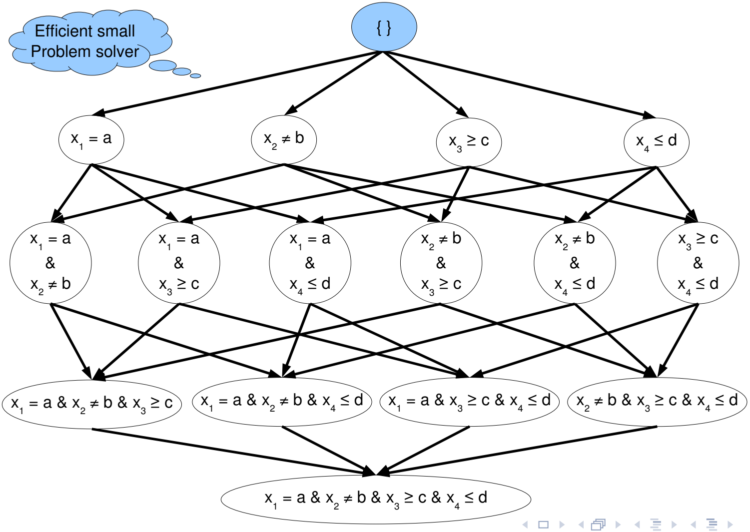
Active Set Method

Solve small problem



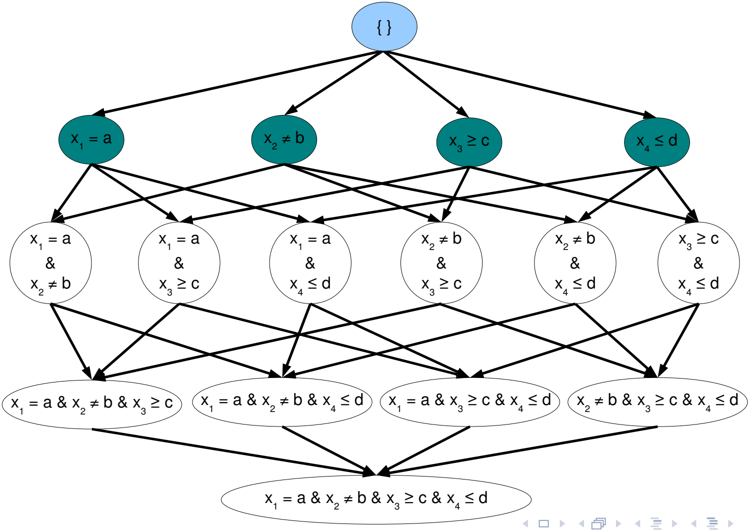
Active Set Method

Solve small problem



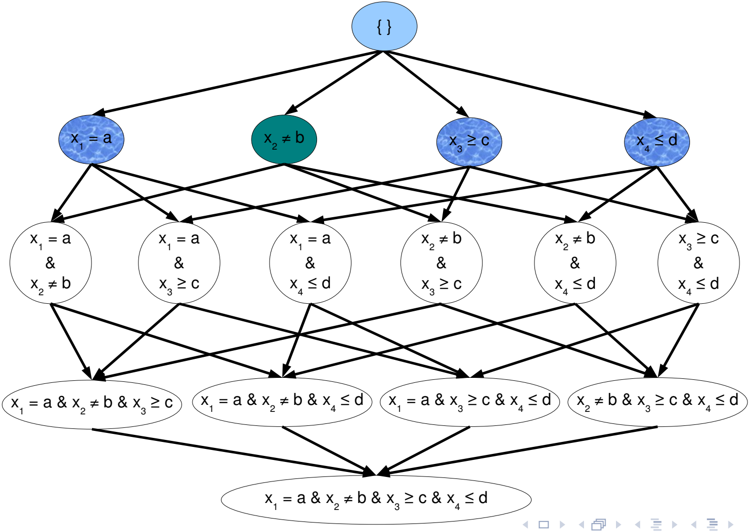
Active Set Method

Identify potential active set entries (i.e., $sources(\mathscr{W}^c)$)



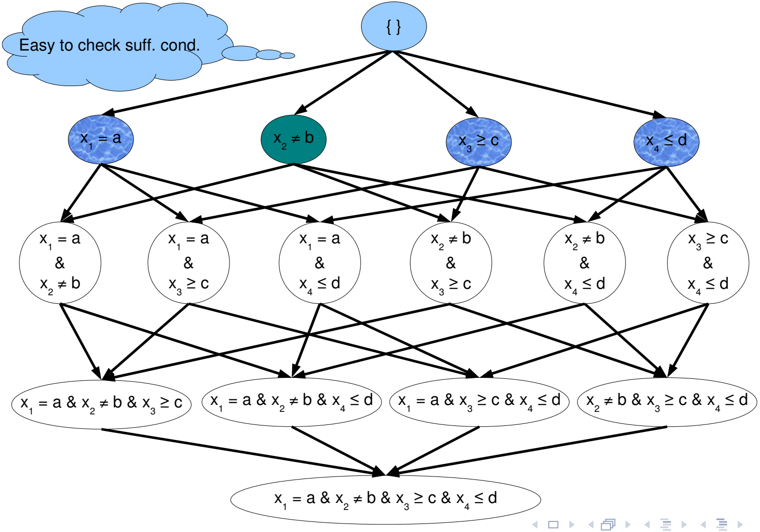
Active Set Method

Among them, optimality condition violators



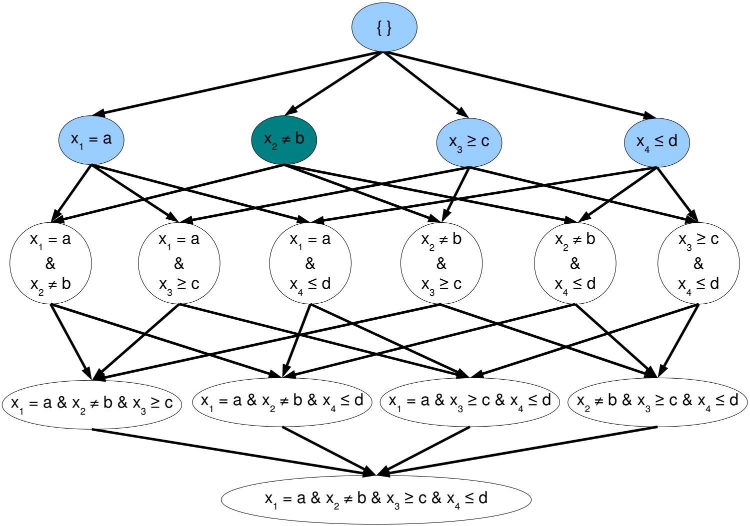
Active Set Method

Among them, optimality condition violators



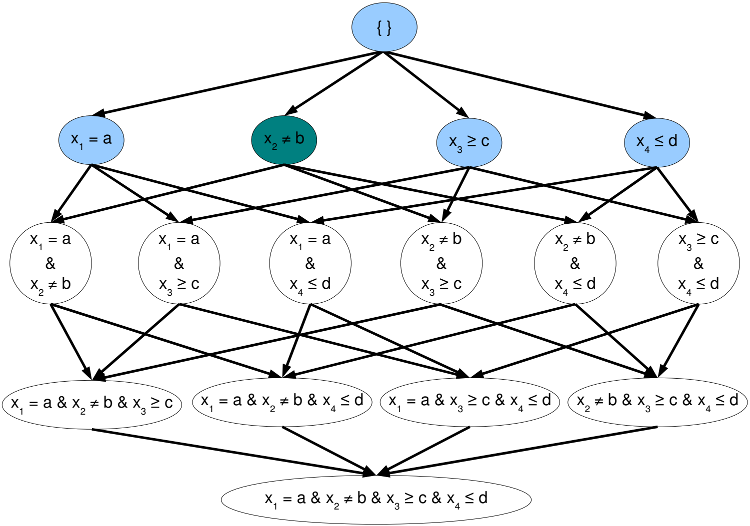
Active Set Method

Append them to active set ($\mathscr{W} = \{0, 1, 3, 4\}$). (repeat until suff. cond. satisfied)



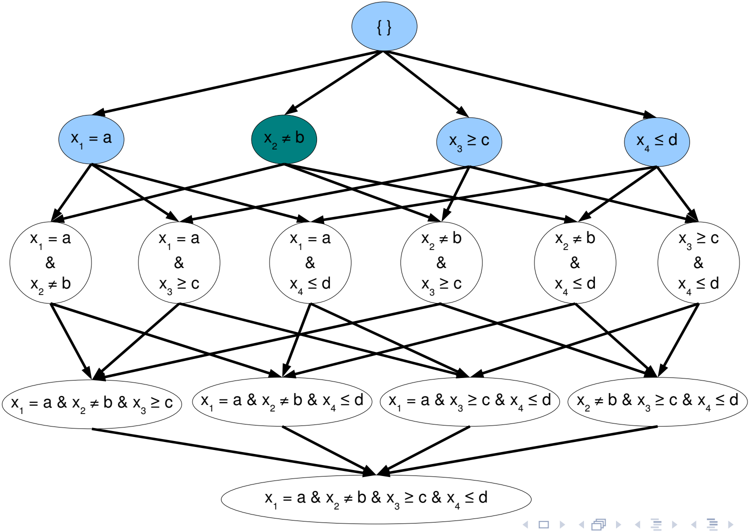
Active Set Method

Append them to active set ($\mathscr{W} = \{0, 1, 3, 4\}$). (repeat until suff. cond. satisfied)



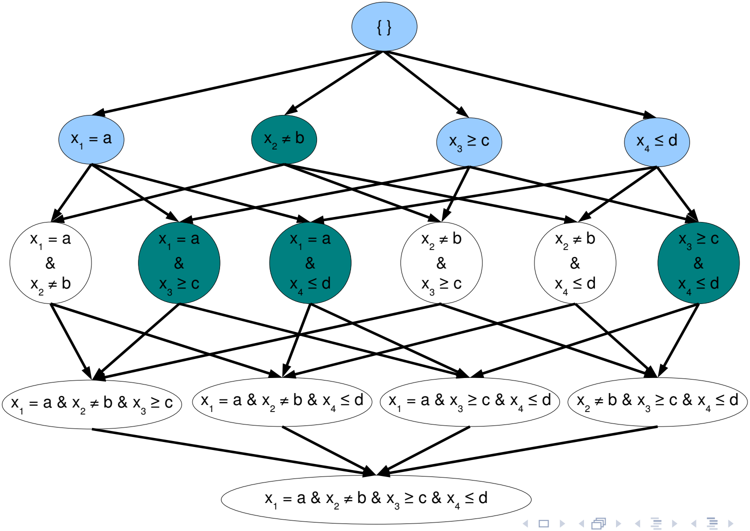
Active Set Method

Solve small problem



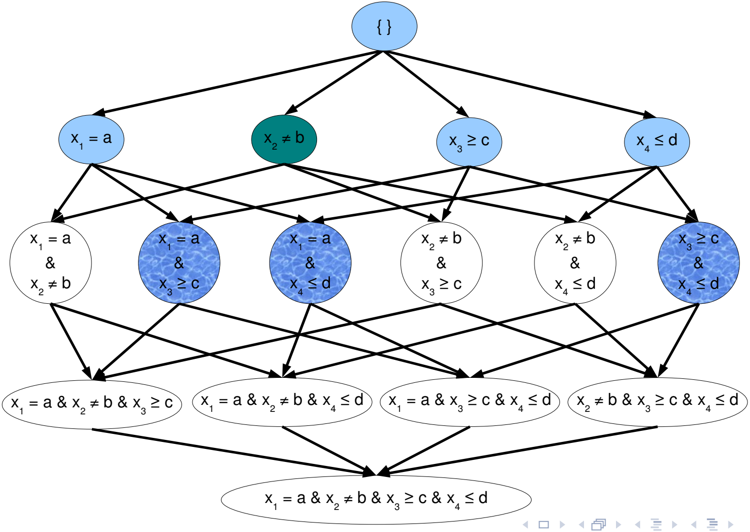
Active Set Method

Identify potential active set entries (i.e., $sources(\mathscr{W}^c)$)



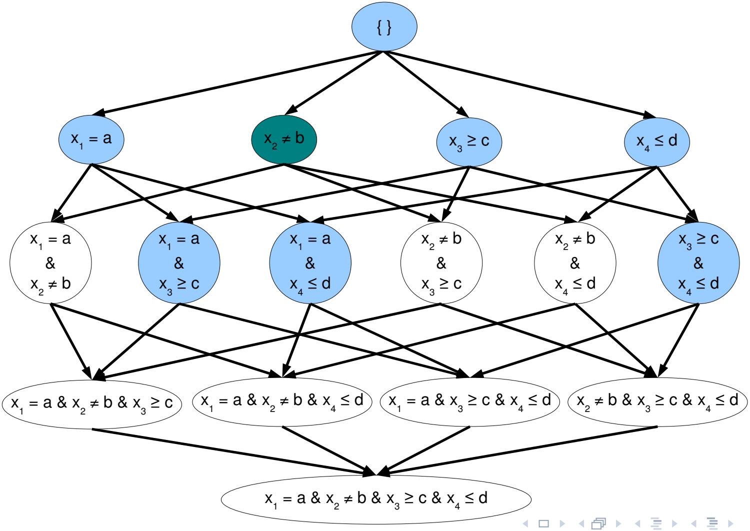
Active Set Method

Among them, optimality condition violators



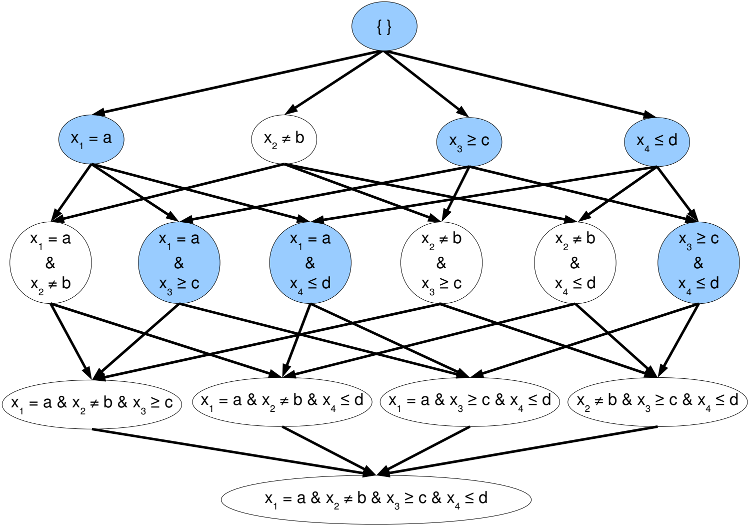
Active Set Method

Append them to active set ($\mathscr{W} = \{0, 1, 3, 4, 6, 7, 10\}$)



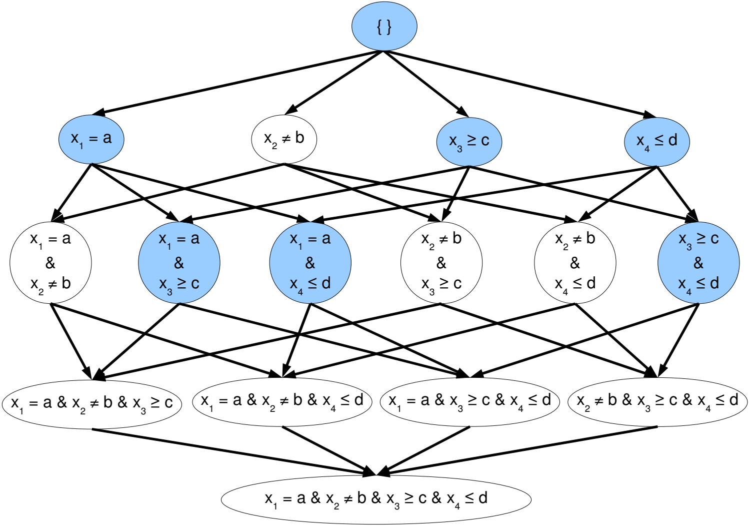
Active Set Method

Final active set: $\mathscr{W} = \{0, 1, 3, 4, 6, 7, 10\}$ (Complexity: Polynomial in active set size)



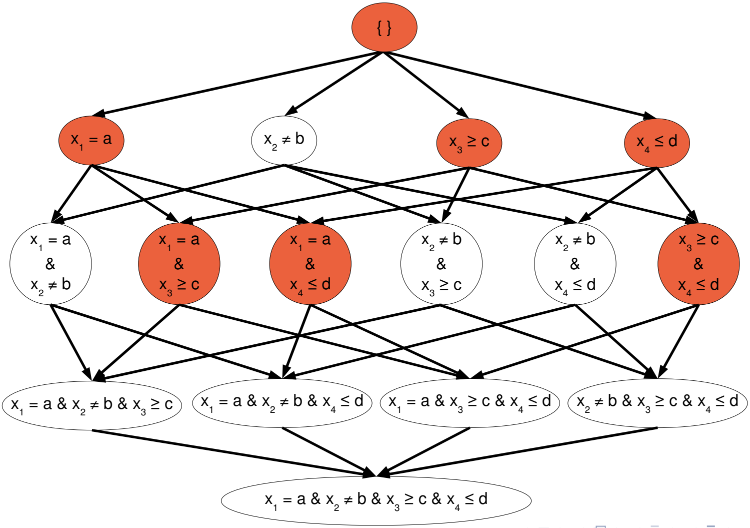
Active Set Method

Final active set: $\mathscr{W} = \{0, 1, 3, 4, 6, 7, 10\}$ (Complexity: Polynomial in active set size)



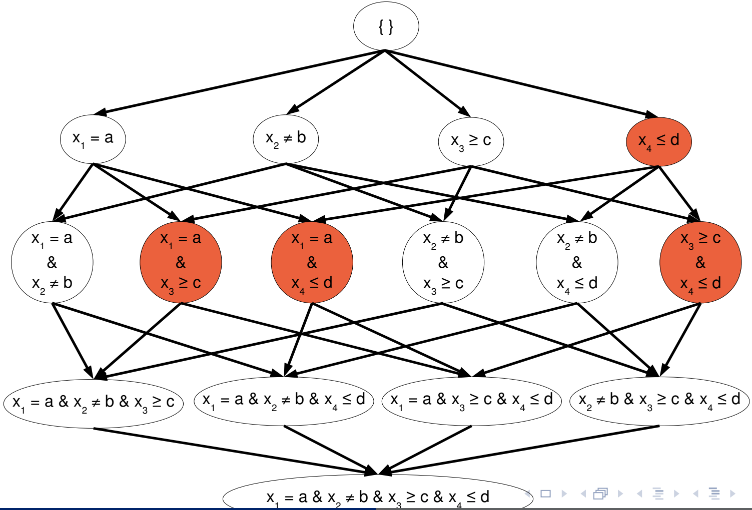
Active Set Method

Solution with HKL



Active Set Method

Key difference from HKL: Node selected without its ancestor!



Key Technical Result

Theorem

A highly specialized partial dual of generalized HKL is:

$$\min_{\eta \in^{|\mathscr{V}|}} \quad g(\eta) \qquad \text{s.t.} \quad \eta \geq 0, \ \sum_{v \in \mathscr{V}} \eta_v = 1$$

where $g(\eta)$ *is the optimal objective value of the following convex problem:*

$$\max_{\alpha \in^m} \quad \sum_{i=1}^m \alpha_i - \frac{1}{2}\left(\sum_{v \in \mathscr{V}} \zeta_v(\eta)\left(\alpha^\top \mathbf{K}_v \alpha\right)\right)^{1} \ \text{s.t.} \ 0 \leq \alpha_i \leq C, \ \sum_{i=1}^m \alpha_i y^i = 0.$$

where $\zeta_v(\eta) = \left(\sum_{u \in A(v)} d_u^\rho \eta_u^{1-\rho}\right)^{\frac{1}{1-\rho}}, = \frac{\rho}{2(\rho-1)}$ *and* $\mathbf{K}_v$ *is matrix with entries:* $y^i y^j k_v(\mathbf{x}^i, \mathbf{x}^j)$.

Key Technical Result

Theorem

A highly specialized partial dual of generalized HKL is:

$$\min_{\eta \in^{|\mathscr{V}|}} \quad g(\eta)$$
$$\text{s.t.} \quad \eta \geq 0,\ \sum_{v\in\mathscr{V}} \eta_v = 1$$

where $g(\eta)$ is the optimal objective value of the following convex problem:

$$\max_{\alpha\in^m} \quad \sum_{i=1}^m \alpha_i - \tfrac{1}{2}\left(\sum_{v\in\mathscr{V}} \zeta_v(\eta)\left(\alpha^\top \mathbf{K}_v \alpha\right)\right)^{\frac{1}{}} \text{ s.t. } 0\leq\alpha_i\leq C,\ \sum_{i=1}^m \alpha_i y^i = 0.$$

where $\zeta_v(\eta) = \left(\sum_{u\in A(v)} d_u^\rho \eta_u^{1-\rho}\right)^{\frac{1}{1-\rho}}$, $= \frac{\rho}{2(\rho-1)}$ and $\mathbf{K}_v$ is matrix with entries: $y^i y^j k_v(\mathbf{x}^i, \mathbf{x}^j)$.

Key Technical Result

Theorem

A highly specialized partial dual of generalized HKL is:

$$\begin{array}{ll} \min_{\eta \in |\mathscr{V}|} & g(\eta) \\ \text{s.t.} & \eta \geq 0,\ \sum_{v \in \mathscr{V}} \eta_v = 1 \end{array}$$

where $g(\eta)$ *is the optimal objective value of the following convex problem:*

$$\max_{\alpha \in m} \quad \sum_{i=1}^{m} \alpha_i - \tfrac{1}{2}\left(\sum_{v \in \mathscr{V}} \zeta_v(\eta)\left(\alpha^\top \mathbf{K}_v \alpha\right)\right)^{\frac{1}{}} \ \text{s.t. } 0 \leq \alpha_i \leq C,\ \sum_{i=1}^{m} \alpha_i y^i = 0.$$

where $\zeta_v(\eta) = \left(\sum_{u \in A(v)} d_u^{\rho} \eta_u^{1-\rho}\right)^{\frac{1}{1-\rho}}, = \frac{\rho}{2(\rho-1)}$ *and* $\mathbf{K}_v$ *is matrix with entries:* $y^i y^j k_v(\mathbf{x}^i, \mathbf{x}^j)$.

Solving small problem

- Dual is min. of convex, Lipschitz conts., sub-differential objective over a simplex.
- Mirror-descent — **highly scalable** alg. for such problems.
- Sub-gradient — solve l_p-MKL (Vishwanathan et.al., 10).

Solving small problem

- Dual is min. of convex, Lipschitz conts., sub-differential objective over a simplex.
- Mirror-descent — **highly scalable** alg. for such problems.
- Sub-gradient — solve l_p-MKL (Vishwanathan et.al., 10).

Solving small problem

- Dual is min. of convex, Lipschitz conts., sub-differential objective over a simplex.
- Mirror-descent — **highly scalable** alg. for such problems.
- Sub-gradient — solve l_p-MKL (Vishwanathan et.al., 10).

Key Technical Result

Theorem

Suppose the active set $\mathscr{W}$ is such that $\mathscr{W} = A(\mathscr{W})$. Let the reduced solution with this $\mathscr{W}$ be $(\mathbf{w}_{\mathscr{W}}, b_{\mathscr{W}})$ and the corresponding dual variables be $(\eta_{\mathscr{W}}, \alpha_{\mathscr{W}})$. Then the reduced solution is a solution to the full problem with a duality gap less than ε if:

$$\max_{t \in sources(\mathscr{W}^c)} \left(\sum_{v \in D(t)} \left(\frac{\alpha_{\mathscr{W}}^{\top} \mathbf{K}_v \alpha_{\mathscr{W}}}{\left(\sum_{u \in A(v) \cap D(t)} d_u \right)^2} \right) \right)^{\frac{1}{}} \leq (\Omega(\mathbf{w}_{\mathscr{W}}))^2 + 2(\varepsilon - \varepsilon_{\mathscr{W}})$$

where $\varepsilon_{\mathscr{W}}$ is a duality gap term associated with the computation of the reduced solution.

Complexity: Polynomial in size of $\mathscr{W}$?

Final Sufficiency Condition:

$$\max_{t \in sources(\mathscr{W}^c)} \left(\sum_{v \in D(t)} \left(\frac{\alpha_{\mathscr{W}}^\top \mathbf{K}_v \alpha_{\mathscr{W}}}{\left(\sum_{u \in A(v) \cap D(t)} d_u \right)^2} \right) \right)^{\frac{1}{\,}} \leq (\Omega(\mathbf{w}_{\mathscr{W}}))^2 + 2(\varepsilon - \varepsilon_{\mathscr{W}})$$

- $\rho \to 1$ ($\to \infty$), suff. cond. tight
- $\rho = 2$ ($= 1$), suff. cond. loose; computationally feasible
- How much ground lost by replacing l_∞ with l_1 ?
  - Not much: As kernels near bottom are extremely sparse

Complexity: Polynomial in size of $\mathscr{W}$?

Final Sufficiency Condition:

$$\max_{t \in sources(\mathscr{W}^c)} \left(\sum_{v \in D(t)} \left(\frac{\alpha_{\mathscr{W}}^\top \mathbf{K}_v \alpha_{\mathscr{W}}}{\left(\sum_{u \in A(v) \cap D(t)} d_u \right)^2} \right) \right)^{\frac{1}{}} \leq (\Omega(\mathbf{w}_{\mathscr{W}}))^2 + 2(\varepsilon - \varepsilon_{\mathscr{W}})$$

- $\rho \to 1$ ($\to \infty$), suff. cond. **tight**
- $\rho = 2$ ($= 1$), suff. cond. loose; computationally **feasible**
- How much ground lost by replacing l_∞ with l_1 ?

Not much: As kernels near bottom are extremely sparse!

Complexity: Polynomial in size of $\mathscr{W}$?

Final Sufficiency Condition:

$$\max_{t \in sources(\mathscr{W}^c)} \left(\sum_{v \in D(t)} \left(\frac{\alpha_{\mathscr{W}}^{\top} \mathbf{K}_v \alpha_{\mathscr{W}}}{\left(\sum_{u \in A(v) \cap D(t)} d_u \right)^2} \right) \right)^{\frac{1}{}} \leq (\Omega(\mathbf{w}_{\mathscr{W}}))^2 + 2(\varepsilon - \varepsilon_{\mathscr{W}})$$

- $\rho \to 1$ ($\to \infty$), suff. cond. **tight**
- $\rho = 2$ ($= 1$), suff. cond. loose; computationally **feasible**
- How much ground lost by replacing l_∞ with l_1 ?
  - Not much: As kernels near bottom are extremely sparse!

Complexity: Polynomial in size of $\mathscr{W}$?

Final Sufficiency Condition:

$$\max_{t \in sources(\mathscr{W}^c)} \left(\sum_{v \in D(t)} \left(\frac{\alpha_{\mathscr{W}}^{\top} \mathbf{K}_v \alpha_{\mathscr{W}}}{\left(\sum_{u \in A(v) \cap D(t)} d_u \right)^2} \right) \right)^{\frac{1}{}} \leq (\Omega(\mathbf{w}_{\mathscr{W}}))^2 + 2(\varepsilon - \varepsilon_{\mathscr{W}})$$

- $\rho \to 1$ $(\to \infty)$, suff. cond. **tight**
- $\rho = 2$ $(= 1)$, suff. cond. loose; computationally **feasible**
- How much ground lost by replacing l_∞ with l_1 ?

Complexity: Polynomial in size of $\mathscr{W}$?

Final Sufficiency Condition:

$$\max_{t\in sources(\mathscr{W}^c)}\left(\sum_{v\in D(t)}\left(\frac{\alpha_{\mathscr{W}}^{\top}\mathbf{K}_v\alpha_{\mathscr{W}}}{\left(\sum_{u\in A(v)\cap D(t)} d_u\right)^2}\right)\right)^{\frac{1}{\;}} \leq (\Omega(\mathbf{w}_{\mathscr{W}}))^2 + 2(\varepsilon - \varepsilon_{\mathscr{W}})$$

- $\rho \to 1$ $(\to \infty)$, suff. cond. **tight**
- $\rho = 2$ $(= 1)$, suff. cond. loose; computationally **feasible**
- How much ground lost by replacing l_∞ with l_1?
  - Not much: As kernels near bottom are extremely sparse!

Complexity: Polynomial in size of $\mathscr{W}$?

Final Sufficiency Condition:

$$\max_{t \in sources(\mathscr{W}^c)} \left(\sum_{v \in D(t)} \left(\frac{\alpha_{\mathscr{W}}^\top \mathbf{K}_v \alpha_{\mathscr{W}}}{\left(\sum_{u \in A(v) \cap D(t)} d_u \right)^2} \right) \right) \leq (\Omega(\mathbf{w}_{\mathscr{W}}))^2 + 2(\varepsilon - \varepsilon_{\mathscr{W}})$$

- $\rho \to 1$ ($\to \infty$), suff. cond. **tight**
- $\rho = 2$ ($= 1$), suff. cond. loose; computationally **feasible**
- How much ground lost by replacing l_∞ with l_1 ?
  - Not much: As kernels near bottom are extremely sparse!

Complexity: Polynomial in size of $\mathscr{W}$?

Final Sufficiency Condition:

$$\max_{t \in sources(\mathscr{W}^c)} \left(\Sigma_{v \in D(t)} \left(\frac{\alpha_{\mathscr{W}}^{\top} \mathbf{K}_v \alpha_{\mathscr{W}}}{\left(\Sigma_{u \in A(v) \cap D(t)} \, d_u \right)^2} \right) \right)^{\frac{1}{}} \leq (\Omega(\mathbf{w}_{\mathscr{W}}))^2 + 2(\varepsilon - \varepsilon_{\mathscr{W}})$$

- $\rho \to 1$ $(\to \infty)$, suff. cond. **tight**
- $\rho = 2$ $(= 1)$, suff. cond. loose; computationally **feasible**
- How much ground lost by replacing l_∞ with l_1 ?
  - **Not much**: As kernels near bottom are extremely sparse!

Complexity: Polynomial in size of $\mathscr{W}$?

Final Sufficiency Condition:

$$\max_{t \in sources(\mathscr{W}^c)} \left(\Sigma_{v \in D(t)} \left(\frac{\alpha_{\mathscr{W}}^\top \mathbf{K}_v \alpha_{\mathscr{W}}}{\left(\Sigma_{u \in A(v) \cap D(t)} d_u\right)^2} \right) \right)^{\frac{1}{}} \leq (\Omega(\mathbf{w}_{\mathscr{W}}))^2 + 2(\varepsilon - \varepsilon_{\mathscr{W}})$$

- $\rho \to 1$ $(\to \infty)$, suff. cond. **tight**
- $\rho = 2$ $(= 1)$, suff. cond. loose; computationally **feasible**
- How much ground lost by replacing l_∞ with l_1 ?
  - **Not much**: As kernels near bottom are extremely sparse!

Complexity: Polynomial in size of $\mathscr{W}$?

Final Sufficiency Condition:

$$\max_{t \in sources(\mathscr{W}^c)} \left(\Sigma_{v \in D(t)} \left(\frac{\alpha_{\mathscr{W}}^\top \mathbf{K}_v \alpha_{\mathscr{W}}}{\left(\Sigma_{u \in A(v) \cap D(t)}\, d_u\right)^2} \right) \right)^{\frac{1}{}} \leq (\Omega(\mathbf{w}_{\mathscr{W}}))^2 + 2(\varepsilon - \varepsilon_{\mathscr{W}})$$

- $\rho \to 1$ $(\to \infty)$, suff. cond. **tight**
- $\rho = 2$ $(= 1)$, suff. cond. loose; computationally **feasible**
- How much ground lost by replacing l_∞ with l_1 ?
  - **Not much**: As kernels near bottom are extremely sparse!

Complexity: Polynomial in size of $\mathscr{W}$?

Final Sufficiency Condition:

$$\max_{t\in sources(\mathscr{W}^c)}\left(\sum_{v\in D(t)}\left(\frac{\alpha_{\mathscr{W}}^{\top}\mathbf{K}_v\alpha_{\mathscr{W}}}{\left(\sum_{u\in A(v)\cap D(t)} d_u\right)^2}\right)\right)^{\frac{1}{}}\leq(\Omega(\mathbf{w}_{\mathscr{W}}))^2+2(\varepsilon-\varepsilon_{\mathscr{W}})$$

- $\rho \to 1$ $(\to \infty)$, suff. cond. **tight**
- $\rho = 2$ $(= 1)$, suff. cond. loose; computationally **feasible**
- How much ground lost by replacing l_∞ with l_1 ?
  - **Not much**: As kernels near bottom are extremely sparse!

Complexity: Polynomial in size of $\mathscr{W}$?

Final Sufficiency Condition:

$$\max_{t \in sources(\mathscr{W}^c)} \left(\Sigma_{v \in D(t)} \left(\frac{\alpha_{\mathscr{W}}^\top \mathbf{K}_v \alpha_{\mathscr{W}}}{\left(\Sigma_{u \in A(v) \cap D(t)}\, d_u\right)^2} \right) \right)^{\frac{1}{}} \leq (\Omega(\mathbf{w}_{\mathscr{W}}))^2 + 2(\varepsilon - \varepsilon_{\mathscr{W}})$$

- $\rho \to 1$ ($\to \infty$), suff. cond. **tight**
- $\rho = 2$ ($= 1$), suff. cond. loose; computationally **feasible**
- How much ground lost by replacing l_∞ with l_1 ?
  - **Not much**: As kernels near bottom are extremely sparse!

Performance Comparison

Dataset	RuleFit	SLI	ENDER	HKL	HKL$_{\rho=1.1}$
TIC-TAC-TOE	0.652 ± 0.068 (**40**, 2.51)	0.747 ± 0.026 (59, 2.35)	0.633 ± 0.011 (111, 2.46)	0.889 ± 0.029 (129, 1.85)	$\mathbf{0.935} \pm 0.043$ (79, **1.77**)
BLOOD TRANS.	0.549 ± 0.092 (18, 1.99)	0.559 ± 0.100 (**6**, **1.07**)	0.489 ± 0.054 (58, 1.5)	$\mathbf{0.594} \pm 0.009$ (242, 1.64)	0.593 ± 0.011 (7,1.40)
BALANCE	0.835 ± 0.034 (**17**, 2.18)	0.856 ± 0.027 (25, 1.88)	0.827 ± 0.013 (64, 1.99)	0.893 ± 0.027 (65, 1.65)	$\mathbf{0.899} \pm 0.023$ (28,**1.23**)
HABERMAN	0.512 ± 0.072 (**6**, 1.68)	0.565 ± 0.066 (8, **1.14**)	0.424 ± 0.000 (18, 1.87)	$\mathbf{0.594} \pm 0.056$ (32, 1.27)	$\mathbf{0.594} \pm 0.056$ (12,1.20)
CAR	0.913 ± 0.033 (**34**, 3.12)	0.895 ± 0.024 (141, 2.27)	0.755 ± 0.028 (80, 1.85)	**0.943** ± 0.024 (87, 1.78)	0.935 ± 0.036 (50,**1.68**)
CMC	0.632 ± 0.013 (39, 2.41)	0.601 ± 0.041 (**13**, 2.13)	0.644 ± 0.026 (74, 2.65)	0.656 ± 0.014 (127, 1.96)	$\mathbf{0.659} \pm 0.008$ (43,**1.70**)

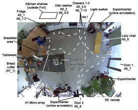
Conjunctive Feature Induction for Sequence Labeling

Sequence Labeling: Assign a label to each instance in a sequence of observations.

Ex: Identify the sequence of activities performed by an old age person in a home based on sensor observations.

Observation: Labels at successive time steps are dependent. Ex: Cooking followed by dinner.

[a]



[b]

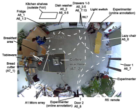
[a] http://depositphotos.com
[b] http://www.opportunity-project.eu

Sequence Labeling: Assign a label to each instance in a sequence of observations.

Ex: Identify the sequence of activities performed by an old age person in a home based on sensor observations.

Observation: Labels at successive time steps are dependent. Ex: Cooking followed by dinner.

[a]



[b]

---

[a] http://depositphotos.com

[b] http://www.opportunity-project.eu

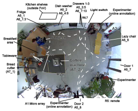
Conjunctive Feature Induction for Sequence Labeling

Sequence Labeling: Assign a label to each instance in a sequence of observations.

Ex: Identify the sequence of activities performed by an old age person in a home based on sensor observations.

Observation: Labels at successive time steps are dependent. Ex: Cooking followed by dinner.

[a]



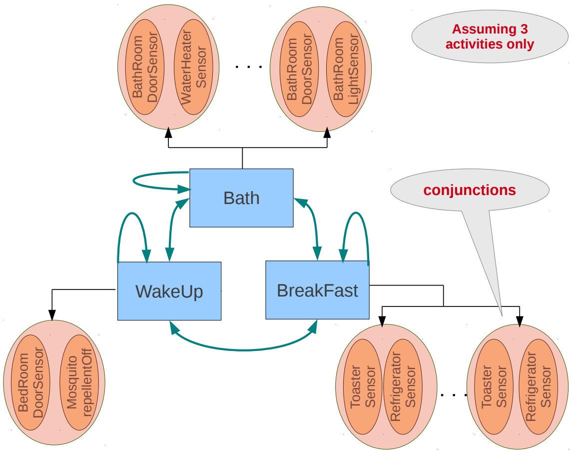
[b]

---

[a] http://depositphotos.com
[b] http://www.opportunity-project.eu

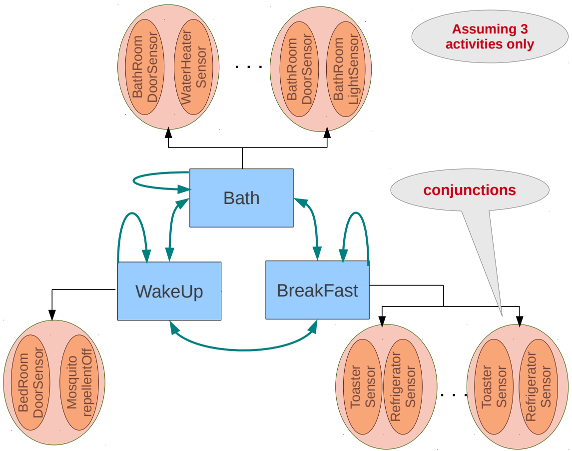
Feature Conjunctions for Sequence Labeling

- Objective is to learn emission features as conjunctions and combine them with all transition relations.
- Model desired



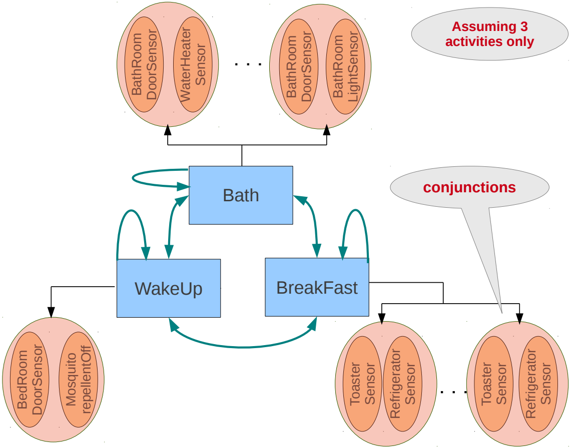
Feature Conjunctions for Sequence Labeling

- Objective is to learn emission features as conjunctions and combine them with all transition relations.



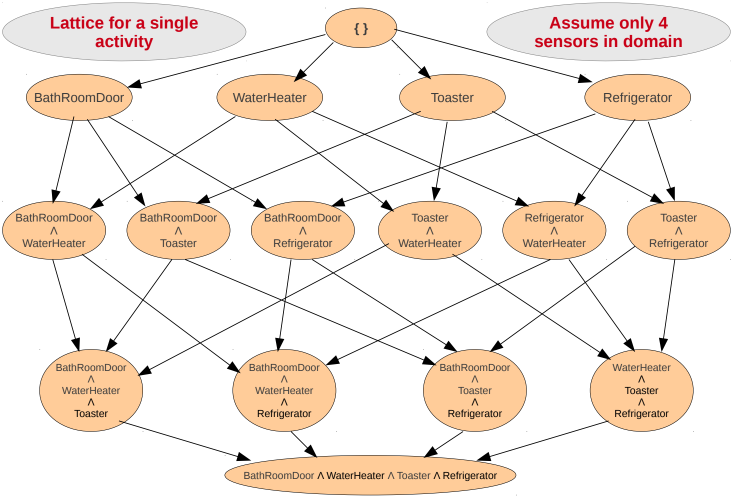
Feature Conjunctions for Sequence Labeling

- Objective is to learn emission features as conjunctions and combine them with all transition relations.
- Model desired

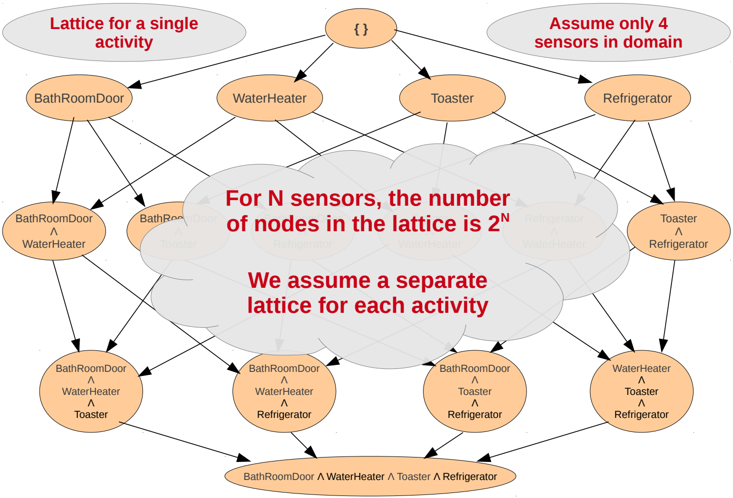


All possible features for a label can be ordered as a partially ordered set (Lattice).

All possible features for a label can be ordered as a partially ordered set (Lattice).

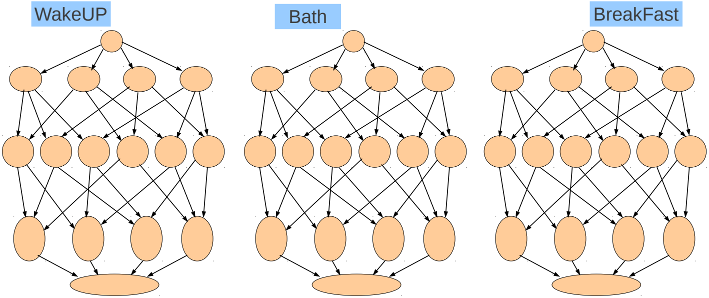


All possible features for a label can be ordered as a partially ordered set (Lattice).

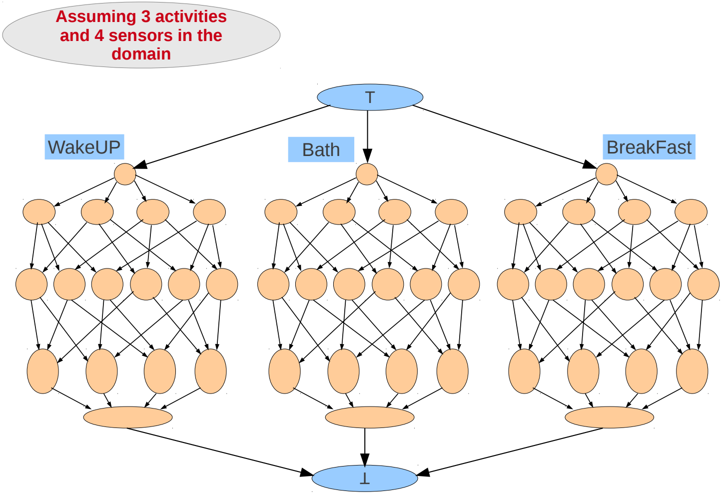


All possible features for a label can be ordered as a partially ordered set (Lattice).

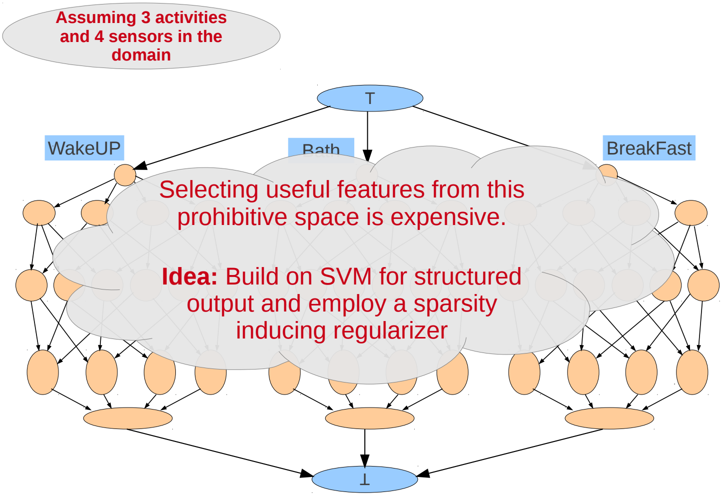
Assuming 3 activities and 4 sensors in the domain



All possible features for a label can be ordered as a partially ordered set (Lattice).



All possible features for a label can be ordered as a partially ordered set (Lattice).



Results - I

- Dataset - Kasteren *et al.* [vKNEK08]
- Activities - 8 ($sleeping$, $usingToilet$, $preparingDinner$, $preparingBreakfast$, $leavingOut$, etc.)
- No. of sensors - 14
- No. of instances - 40K

	Micro avg.	Macro avg.
Std. HMM	25.40 (±18.55)	21.75 (±12.12)
B&B HMM	29.54 (±20.70)	16.39 (±02.74)
Greedy FIHMM	58.08 (±10.14)	26.84 (±04.41)
StructSVM	58.02 (±11.87)	35.00 (±05.24)
CRF	48.49 (±05.02)	20.65 (±04.82)
FICRF	59.52 (±11.76)	33.60 (±07.38)
RELHKL	46.28 (±11.44)	23.11 (±07.46)
StructRELHKL	63.96 (±05.74)	32.01 (±03.04)

Table: Micro (Weighted Per-Class) average accuracy and macro(Simple Per-Class) average accuracy of classification on UA dataset.

Results - I

- Dataset - Kasteren *et al.* [vKNEK08]
- Activities - 8 $(sleeping, usingToilet, preparingDinner, preparingBreakfast, leavingOut,$ *etc.*)
- No. of sensors - 14
- No. of instances - 40K

	Micro avg.	Macro avg.
Std. HMM	25.40 (±18.55)	21.75 (±12.12)
B&B HMM	29.54 (±20.70)	16.39 (±02.74)
Greedy FIHMM	58.08 (±10.14)	26.84 (±04.41)
StructSVM	58.02 (±11.87)	35.00 (±05.24)
CRF	48.49 (±05.02)	20.65 (±04.82)
FICRF	59.52 (±11.76)	33.60 (±07.38)
RELHKL	46.28 (±11.44)	23.11 (±07.46)
StructRELHKL	63.96 (±05.74)	32.01 (±03.04)

Table: Micro (Weighted Per-Class) average accuracy and macro(Simple Per-Class) average accuracy of classification on UA dataset.

Results - I

- Dataset - Kasteren *et al.* [vKNEK08]
- Activities - 8 $(sleeping, usingToilet, preparingDinner, preparingBreakfast, leavingOut,$ *etc.*)
- No. of sensors - 14
- No. of instances - 40K

	Micro avg.	Macro avg.
Std. HMM	25.40 (±18.55)	21.75 (±12.12)
B&B HMM	29.54 (±20.70)	16.39 (±02.74)
Greedy FIHMM	58.08 (±10.14)	26.84 (±04.41)
StructSVM	58.02 (±11.87)	35.00 (±05.24)
CRF	48.49 (±05.02)	20.65 (±04.82)
FICRF	59.52 (±11.76)	33.60 (±07.38)
RELHKL	46.28 (±11.44)	23.11 (±07.46)
StructRELHKL	63.96 (±05.74)	32.01 (±03.04)

Table: Micro (Weighted Per-Class) average accuracy and macro(Simple Per-Class) average accuracy of classification on UA dataset.

Results - I

- Dataset - Kasteren *et al.* [vKNEK08]
- Activities - 8 ($sleeping$, $usingToilet$, $preparingDinner$, $preparingBreakfast$, $leavingOut$, *etc.*)
- No. of sensors - 14
- No. of instances - 40K

	Micro avg.	Macro avg.
Std. HMM	25.40 (±18.55)	21.75 (±12.12)
B&B HMM	29.54 (±20.70)	16.39 (±02.74)
Greedy FIHMM	58.08 (±10.14)	26.84 (±04.41)
StructSVM	58.02 (±11.87)	35.00 (±05.24)
CRF	48.49 (±05.02)	20.65 (±04.82)
FICRF	59.52 (±11.76)	33.60 (±07.38)
RELHKL	46.28 (±11.44)	23.11 (±07.46)
StructRELHKL	63.96 (±05.74)	32.01 (±03.04)

Table: Micro (Weighted Per-Class) average accuracy and macro(Simple Per-Class) average accuracy of classification on UA dataset.

Results - II

- Dataset - MIT PlaceLab [TIL04] on Subject1 and Subject2
- No. of sensors - 76 for Subject1 and 70 for Subject2
- No. of examples - 20K and 24K resp.

		Micro avg.	Macro avg.
Subject 1	StructSVM	75.03 (±04.51)	26.99 (±07.73)
	CRF	65.54 (±06.80)	31.19 (±07.39)
	FICRF	68.52 (±07.19)	29.77 (±03.59)
	StructRELHKL	82.88 (±0.43)	28.92 (±01.53)
Subject 2	StructSVM	63.49 (±02.75)	25.33 (±05.8)
	CRF	50.23 (±06.80)	27.42 (±07.65)
	FICRF	51.86 (±07.35)	26.11 (±05.89)
	StructRELHKL	67.16 (±08.64)	24.32 (±02.12)

Table: Micro average accuracy and macro average accuracy of classification on PlaceLab dataset.

Results - II

- Dataset - MIT PlaceLab [TIL04] on Subject1 and Subject2
- No. of sensors - 76 for Subject1 and 70 for Subject2
- No. of examples - 20K and 24K resp.

		Micro avg.	Macro avg.
Subject 1	StructSVM	75.03 ($\pm$04.51)	26.99 ($\pm$07.73)
	CRF	65.54 ($\pm$06.80)	31.19 ($\pm$07.39)
	FICRF	68.52 ($\pm$07.19)	29.77 ($\pm$03.59)
	StructRELHKL	82.88 ($\pm$0.43)	28.92 ($\pm$01.53)
Subject 2	StructSVM	63.49 ($\pm$02.75)	25.33 ($\pm$05.8)
	CRF	50.23 ($\pm$06.80)	27.42 ($\pm$07.65)
	FICRF	51.86 ($\pm$07.35)	26.11 ($\pm$05.89)
	StructRELHKL	67.16 ($\pm$08.64)	24.32 ($\pm$02.12)

Table: Micro average accuracy and macro average accuracy of classification on PlaceLab dataset.

Results - II

- Dataset - MIT PlaceLab [TIL04] on Subject1 and Subject2
- No. of sensors - 76 for Subject1 and 70 for Subject2
- No. of examples - 20K and 24K resp.

		Micro avg.	Macro avg.
Subject 1	StructSVM	75.03 (±04.51)	26.99 (±07.73)
	CRF	65.54 (±06.80)	31.19 (±07.39)
	FICRF	68.52 (±07.19)	29.77 (±03.59)
	StructRELHKL	82.88 (±0.43)	28.92 (±01.53)
Subject 2	StructSVM	63.49 (±02.75)	25.33 (±05.8)
	CRF	50.23 (±06.80)	27.42 (±07.65)
	FICRF	51.86 (±07.35)	26.11 (±05.89)
	StructRELHKL	67.16 (±08.64)	24.32 (±02.12)

Table: Micro average accuracy and macro average accuracy of classification on PlaceLab dataset.

Sample Rules/Features Induced

- $usingToilet \leftarrow bathroomDoor \wedge toiletFlush$
- $sleeping \leftarrow bedroomDoor \wedge toiletDoor \wedge bathroomDoor,$
- $preparingDinner \leftarrow groceries\ Cupboard$

Sample Rules/Features Induced

- $usingToilet \leftarrow bathroomDoor \wedge toiletFlush$
- $sleeping \leftarrow bedroomDoor \wedge toiletDoor \wedge bathroomDoor,$
- $preparingDinner \leftarrow groceries\ Cupboard$

Sample Rules/Features Induced

- $usingToilet \leftarrow bathroomDoor \wedge toiletFlush$
- $sleeping \leftarrow bedroomDoor \wedge toiletDoor \wedge bathroomDoor,$
- $preparingDinner \leftarrow groceries\ Cupboard$

Inducing Feature Disjunctions

- Feature Subset Selection
  - wrappers provided in weka [WFH11]
  - legacy systems: Relief, Focus
  - L1-SVM
- Feature Extraction
  - Latent Dirichlet Allocation [BNJ03] and variants
  - Discriminant Analysis [YJ] and variants
  - Principal Component Analysis [Jol86] and variants
  - Max-Margin Dimensionality Reduction Methods [LJZ03]

Inducing Feature Disjunctions

- Feature Subset Selection
  - wrappers provided in weka [WFH11]
  - legacy systems: Relief, Focus
  - L1-SVM
- Feature Extraction
  - Latent Dirichlet Allocation [BNJ03] and variants
  - Discriminant Analysis [YJ] and variants
  - Principal Component Analysis [Jol86] and variants
  - Max-Margin Dimensionality Reduction Methods [LJZ03]

Inducing Feature Disjunctions

- Feature Subset Selection
  - wrappers provided in weka [WFH11]
  - legacy systems: Relief, Focus
  - L1-SVM
- Feature Extraction
  - Latent Dirichlet Allocation [BNJ03] and variants
  - Discriminant Analysis [YJ] and variants
  - Principal Component Analysis [Jol86] and variants
  - Max-Margin Dimensionality Reduction Methods [LJZ03]

Inducing Feature Disjunctions

- Feature Subset Selection
  - wrappers provided in weka [WFH11]
  - legacy systems: Relief, Focus
  - L1-SVM
- Feature Extraction
  - Latent Dirichlet Allocation [BNJ03] and variants
  - Discriminant Analysis [YJ] and variants
  - Principal Component Analysis [Jol86] and variants
  - Max-Margin Dimensionality Reduction Methods [LJZ03]

Inducing Feature Disjunctions

- Feature Subset Selection
  - wrappers provided in weka [WFH11]
  - legacy systems: Relief, Focus
  - L1-SVM
- Feature Extraction
  - Latent Dirichlet Allocation [BNJ03] and variants
  - Discriminant Analysis [YJ] and variants
  - Principal Component Analysis [Jol86] and variants
  - Max-Margin Dimensionality Reduction Methods [LJZ03]

Inducing Feature Disjunctions

- Feature Subset Selection
  - wrappers provided in weka [WFH11]
  - legacy systems: Relief, Focus
  - L1-SVM
- Feature Extraction
  - Latent Dirichlet Allocation [BNJ03] and variants
  - Discriminant Analysis [YJ] and variants
  - Principal Component Analysis [Jol86] and variants
  - Max-Margin Dimensionality Reduction Methods [LJZ03]

Inducing Feature Disjunctions

- Feature Subset Selection
  - wrappers provided in weka [WFH11]
  - legacy systems: Relief, Focus
  - L1-SVM
- Feature Extraction
  - Latent Dirichlet Allocation [BNJ03] and variants
  - Discriminant Analysis [YJ] and variants
  - Principal Component Analysis [Jol86] and variants
  - Max-Margin Dimensionality Reduction Methods [LJZ03]

Inducing Feature Disjunctions

- Feature Subset Selection
  - wrappers provided in weka [WFH11]
  - legacy systems: Relief, Focus
  - L1-SVM
- Feature Extraction
  - Latent Dirichlet Allocation [BNJ03] and variants
  - Discriminant Analysis [YJ] and variants
  - Principal Component Analysis [Jol86] and variants
  - Max-Margin Dimensionality Reduction Methods [LJZ03]

Inducing Feature Disjunctions

- Feature Subset Selection
  - wrappers provided in weka [WFH11]
  - legacy systems: Relief, Focus
  - L1-SVM
- Feature Extraction
  - Latent Dirichlet Allocation [BNJ03] and variants
  - Discriminant Analysis [YJ] and variants
  - Principal Component Analysis [Jol86] and variants
  - Max-Margin Dimensionality Reduction Methods [LJZ03]

Main Concern Addressed

- No. of reduced dimensions taken as input?
- Dimension Reduction is not integrated with model building
- Hence no guarantee of optimality

Main Concern Addressed

- No. of reduced dimensions taken as input?
- Dimension Reduction is not integrated with model building
- Hence no guarantee of optimality

Main Concern Addressed

- No. of reduced dimensions taken as input?
- Dimension Reduction is not integrated with model building
- Hence no guarantee of optimality

Dimension Reduction Techniques

Method	Parameterized	Supervised	Integrated
Latent Dirichlet Allocation(LDA)	✓	✗	✗
Supervised LDA	✓	✓	✓
Labeled LDA	✓	MultiLabel Supervision	✓
Discriminative LDA	✓	✓	✓
Hierarchical Supervised LDA	✓	Hierarchical Supervision	✓
Kernel Dimension Reducion	✓	✓	✗
Hierarchical Dirichlet Processes(HDP)	✗	✗	✗
Hierarchical LDA	✗	✗	✗
Supervised HDP	✗	✓	✗
PCA, Kernel PCA, LSI, pLSI	✓	✓	✗
Random Projection	✓	✓	✗
Self Organizing Map	✓	✓	✗
Multidimensional Scaling	✓	✓	✗
Discriminant Analysis	✓	✓	✗

Dimension Reduction Techniques - Max. Margin

Method	Parameterized	Supervised	Integrated
Max. Margin Dimension −Reduction(MMDR)	✓	✓	✓
Linear MMDR	✓	✓	✓
medLDA	✓	✓	✓
mmPLSA	✓	✓	✓

- We propose an approach of Max-Margin Dimension Reduction
  - non-parametric
  - supervised
  - Integrated with Classifier Training
  - leads to optimum model building
  - can incorporate further Background Knowledge

Dimension Reduction Techniques - Max. Margin

Method	Parameterized	Supervised	Integrated
Max. Margin Dimension −Reduction(MMDR)	✓	✓	✓
Linear MMDR	✓	✓	✓
medLDA	✓	✓	✓
mmPLSA	✓	✓	✓

- We propose an approach of Max-Margin Dimension Reduction
  - non-parametric
  - supervised
  - Integrated with Classifier Training
  - leads to optimum model building
  - can incorporate further Background Knowledge

Dimension Reduction Techniques - Max. Margin

Method	Parameterized	Supervised	Integrated
Max. Margin Dimension −Reduction(MMDR)	✓	✓	✓
Linear MMDR	✓	✓	✓
medLDA	✓	✓	✓
mmPLSA	✓	✓	✓

- We propose an approach of Max-Margin Dimension Reduction
  - non-parametric
  - supervised
  - Integrated with Classifier Training
  - leads to optimum model building
  - can incorporate further Background Knowledge

Dimension Reduction Techniques - Max. Margin

Method	Parameterized	Supervised	Integrated
Max. Margin Dimension −Reduction(MMDR)	✓	✓	✓
Linear MMDR	✓	✓	✓
medLDA	✓	✓	✓
mmPLSA	✓	✓	✓

- We propose an approach of Max-Margin Dimension Reduction
  - non-parametric
  - supervised
  - Integrated with Classifier Training
  - leads to optimum model building
  - can incorporate further Background Knowledge

Dimension Reduction Techniques - Max. Margin

Method	Parameterized	Supervised	Integrated
Max. Margin Dimension −Reduction(MMDR)	✓	✓	✓
Linear MMDR	✓	✓	✓
medLDA	✓	✓	✓
mmPLSA	✓	✓	✓

- We propose an approach of Max-Margin Dimension Reduction
  - non-parametric
  - supervised
  - Integrated with Classifier Training
  - leads to optimum model building
  - can incorporate further Background Knowledge

Dimension Reduction Techniques - Max. Margin

Method	Parameterized	Supervised	Integrated
Max. Margin Dimension −Reduction(MMDR)	✓	✓	✓
Linear MMDR	✓	✓	✓
medLDA	✓	✓	✓
mmPLSA	✓	✓	✓

- We propose an approach of Max-Margin Dimension Reduction
  - non-parametric
  - supervised
  - Integrated with Classifier Training
  - leads to optimum model building
  - can incorporate further Background Knowledge

Dimension Reduction by Disjunctions

- Discovering **Small Set** of **Good** & **Maximal Disjunctive Projections**
  - maintains synonymity: $\{elegant, exquisite\}$ ✓
  - relevant: $\{method, algorithm\}$ in Sentiment Analysis ✗
  - maximal: $\{elegant, exquisite, crude\}$ ✗
- Structure Induced
  - if $\{elegant, exquisite, stately\}$ is a good Disjunctive Projection so is its subsets
  - if $\{elegant, crude\}$ is not a good Disjunctive Projection, so isnt its supersets

Dimension Reduction by Disjunctions

- Discovering **Small Set** of **Good** & **Maximal Disjunctive Projections**
  - maintains synonymity: $\{elegant, exquisite\}$ ✓
  - relevant: $\{method, algorithm\}$ in Sentiment Analysis ✗
  - maximal: $\{elegant, exquisite, crude\}$ ✗
- Structure Induced
  - if $\{elegant, exquisite, stately\}$ is a good Disjunctive Projection so is its subsets
  - if $\{elegant, crude\}$ is not a good Disjunctive Projection, so isnt its supersets

Dimension Reduction by Disjunctions

- Discovering **Small Set** of **Good** & **Maximal Disjunctive Projections**
  - maintains synonymity: $\{elegant, exquisite\}$ ✓
  - relevant: $\{method, algorithm\}$ in Sentiment Analysis ✗
  - maximal: $\{elegant, exquisite, crude\}$ ✗
- Structure Induced
  - if $\{elegant, exquisite, stately\}$ is a good Disjunctive Projection so is its subsets
  - if $\{elegant, crude\}$ is not a good Disjunctive Projection, so isnt its supersets

Dimension Reduction by Disjunctions

- Discovering **Small Set** of **Good** & **Maximal Disjunctive Projections**
  - maintains synonymity: $\{elegant, exquisite\}$ ✓
  - relevant: $\{method, algorithm\}$ in Sentiment Analysis ✗
  - maximal: $\{elegant, exquisite, crude\}$ ✗
- Structure Induced
  - if $\{elegant, exquisite, stately\}$ is a good Disjunctive Projection so is its subsets
  - if $\{elegant, crude\}$ is not a good Disjunctive Projection, so isnt its supersets

Dimension Reduction by Disjunctions

- Discovering **Small Set** of **Good** & **Maximal Disjunctive Projections**
  - maintains synonymity: $\{elegant, exquisite\}$ ✓
  - relevant: $\{method, algorithm\}$ in Sentiment Analysis ✗
  - maximal: $\{elegant, exquisite, crude\}$ ✗
- Structure Induced
  - if $\{elegant, exquisite, stately\}$ is a good Disjunctive Projection so is its subsets
  - if $\{elegant, crude\}$ is not a good Disjunctive Projection, so isnt its supersets

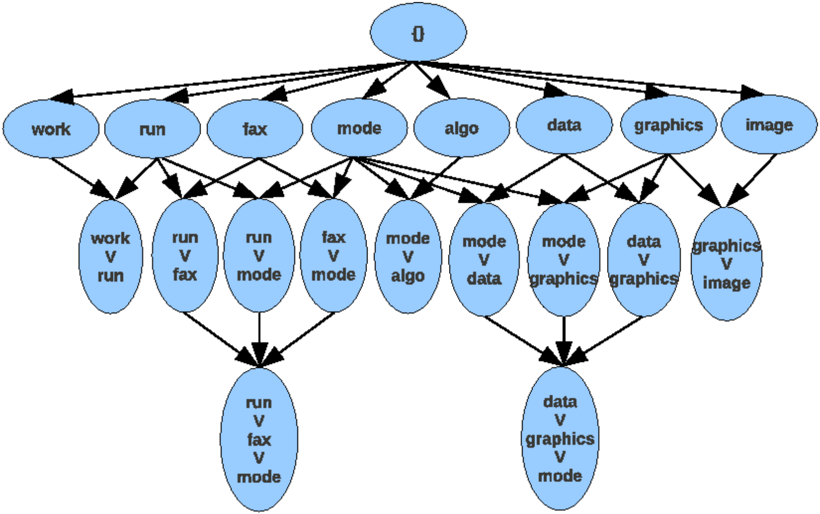
Dimension Reduction by Disjunctions

- Discovering **Small Set** of **Good** & **Maximal Disjunctive Projections**
  - maintains synonymity: $\{elegant, exquisite\}$ ✓
  - relevant: $\{method, algorithm\}$ in Sentiment Analysis ✗
  - maximal: $\{elegant, exquisite, crude\}$ ✗
- Structure Induced
  - if $\{elegant, exquisite, stately\}$ is a good Disjunctive Projection so is its subsets
  - if $\{elegant, crude\}$ is not a good Disjunctive Projection, so isnt its supersets

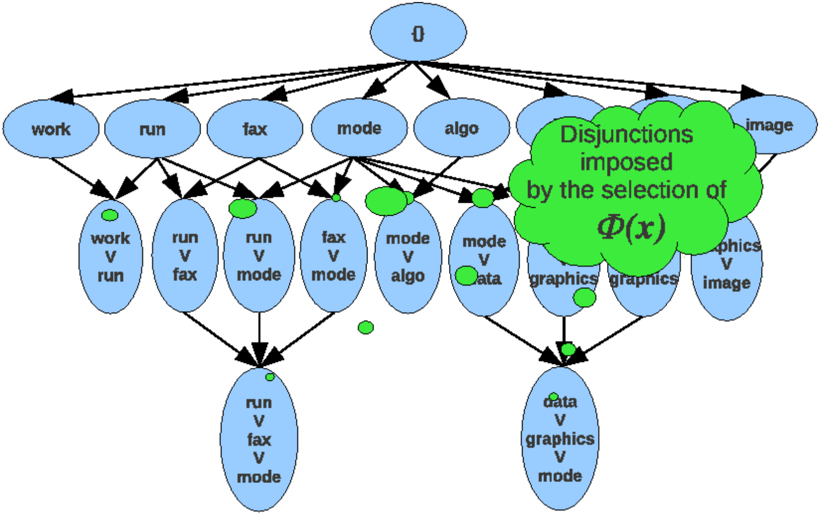
Dimension Reduction by Disjunctions

- Discovering **Small Set** of **Good** & **Maximal Disjunctive Projections**
  - maintains synonymity: $\{elegant, exquisite\}$ ✓
  - relevant: $\{method, algorithm\}$ in Sentiment Analysis ✗
  - maximal: $\{elegant, exquisite, crude\}$ ✗
- Structure Induced
  - if $\{elegant, exquisite, stately\}$ is a good Disjunctive Projection so is its subsets
  - if $\{elegant, crude\}$ is not a good Disjunctive Projection, so isnt its supersets

Lattice Structure of Disjunctive Projections



Lattice Structure of Disjunctive Projections



Hierarchical Kernel Learning Setting

- For such structure : group Norm on Descendant Sets in HKL framework
- Sparse number of disjunctions : ρ-norm
- Efficient solution : Active Set Algorithm

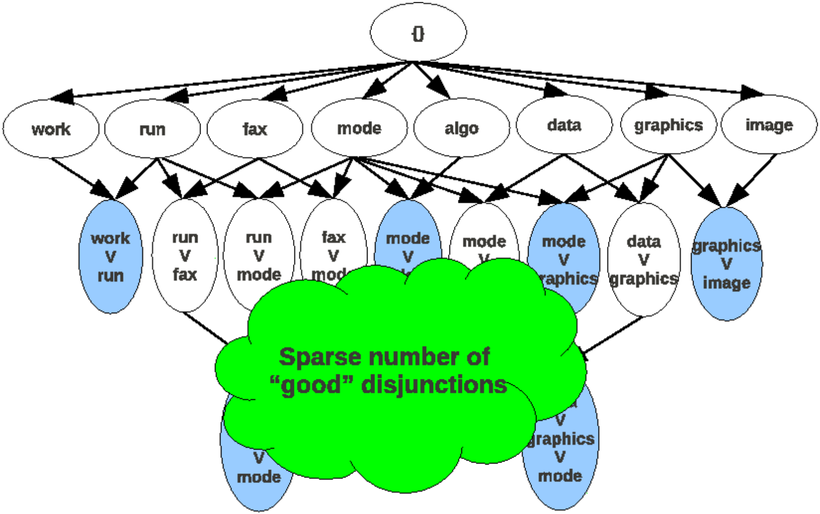
Hierarchical Kernel Learning Setting

- For such structure : group Norm on Descendant Sets in HKL framework
- Sparse number of disjunctions : ρ-norm
- Efficient solution : Active Set Algorithm

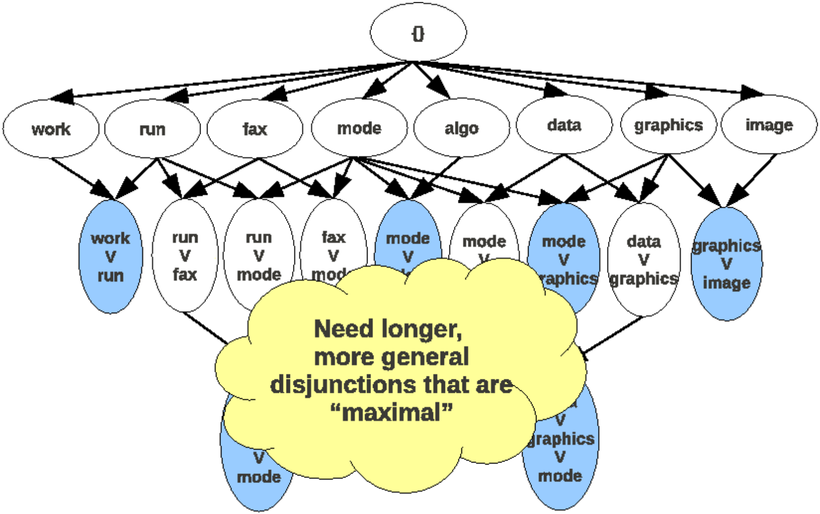
Hierarchical Kernel Learning Setting

- For such structure : group Norm on Descendant Sets in HKL framework
- Sparse number of disjunctions : ρ-norm
- Efficient solution : Active Set Algorithm

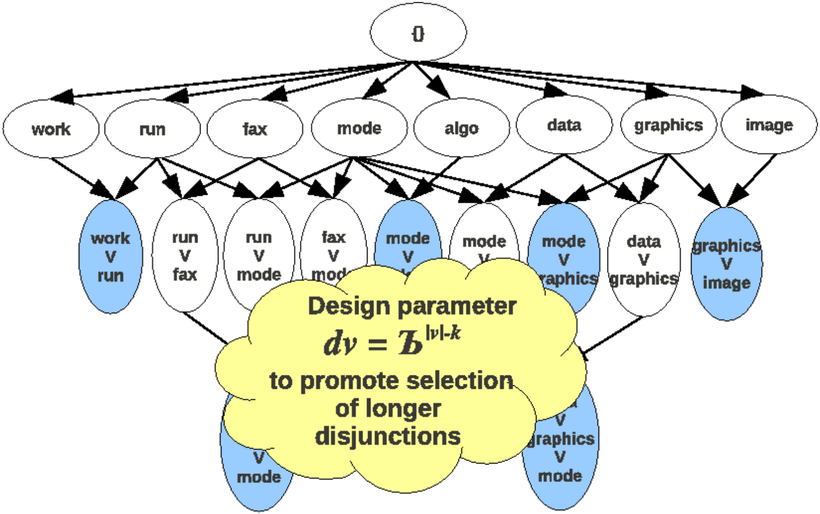
Lattice Structure of Disjunctive Projections



Lattice Structure of Disjunctive Projections



Lattice Structure of Disjunctive Projections



Max-Margin Objective

- Max Margin Objective

$$\min_{\mathbf{w},b,\xi} \; \frac{1}{2}\left(\sum_{v\in\mathcal{V}} \delta_v \parallel \mathbf{w}_{D(v)} \parallel_\rho\right)^2 + C\, \mathbf{1}^\top \xi$$

$$s.t.\ \forall i:\ y_i\left(\sum_{v\in\mathcal{V}} \langle w_v, \phi_v(\mathbf{x}_i)\rangle - b\right) \geq 1 - \xi_i,\, ,\, \xi \geq 0$$

- decision function is $\sum_{v\in\mathcal{V}} w_v \phi_v(x) - b$
- where $\phi_v(.) = \vee_{\hat{v}\in v}\, \phi_{\hat{v}}(.) = 1 - \prod_{\hat{v}\in v} \overline{\phi_{\hat{v}}}(.)$

Max-Margin Objective

- Max Margin Objective

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\Big(\sum_{v\in\mathcal{V}} \delta_v \parallel \mathbf{w}_{D(v)} \parallel_\rho \Big)^2 + C\,\mathbf{1}^\top \xi$$

$$s.t.\ \forall i:\ y_i\Big(\sum_{v\in\mathcal{V}} \langle w_v, \phi_v(\mathbf{x}_i)\rangle - b\Big) \geq 1-\xi_i,, \xi \geq 0$$

- decision function is $\sum_{v\in\mathcal{V}} w_v\phi_v(x) - b$
- where $\phi_v(.) = \vee_{\hat{v}\in v}\ \phi_{\hat{v}}(.) = 1 - \prod_{\hat{v}\in v} \overline{\phi_{\hat{v}}}(.)$

Max-Margin Objective

- Max Margin Objective

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\left(\sum_{v\in\mathcal{V}} \delta_v \parallel \mathbf{w}_{D(v)} \parallel_\rho\right)^2 + C\,\mathbf{1}^\top \xi$$

$$s.t.\ \forall i:\ y_i\left(\sum_{v\in\mathcal{V}} \langle w_v, \phi_v(\mathbf{x}_i)\rangle - b\right) \geq 1 - \xi_i, , \xi \geq 0$$

- decision function is $\sum_{v\in\mathcal{V}} w_v \phi_v(x) - b$
- where $\phi_v(.) = \vee_{\hat{v}\in v}\ \phi_{\hat{v}}(.) = 1 - \prod_{\hat{v}\in v} \overline{\phi_{\hat{v}}}(.)$

Experiment-1

- Dataset - 20 Newsgroups Posting on '*alt.atheism*' and '*talk.religion.misc*'
- No of words/Features - 18826
- No of instances - 856

Approach	Accuracy	No. of Topics/Disjunctions
MMpLSA [Xu10]	84.7%	3
MedLDA [ZAX10]	73.12%[2]	20
DiscLDA [LJSJ08]	83.0%	60
Integ. Dim. Red.	94.55%	170

Table: Comparison of accuracies of different approaches on 20 *Newsgroups* [Lan] dataset.

Experiment-1

- Dataset - 20 Newsgroups Posting on '*alt.atheism*' and '*talk.religion.misc*'
- No of words/Features - 18826
- No of instances - 856

Approach	Accuracy	No. of Topics/Disjunctions
MMpLSA [Xu10]	84.7%	3
MedLDA [ZAX10]	73.12%[2]	20
DiscLDA [LJSJ08]	83.0%	60
Integ. Dim. Red.	94.55%	170

Table: Comparison of accuracies of different approaches on 20 *Newsgroups* [Lan] dataset.

Experiment-1

- Dataset - 20 Newsgroups Posting on '*alt.atheism*' and '*talk.religion.misc*'
- No of words/Features - 18826
- No of instances - 856

Approach	Accuracy	No. of Topics/Disjunctions
MMpLSA [Xu10]	84.7%	3
MedLDA [ZAX10]	73.12%[2]	20
DiscLDA [LJSJ08]	83.0%	60
Integ. Dim. Red.	94.55%	170

Table: Comparison of accuracies of different approaches on 20 *Newsgroups* [Lan] dataset.

Experiments-2

		Breast-cancer		Wisconsin		Hepatitis		20Newsgroups	
Subset Evaluator	Search Method	$\mathscr{L}_2$	$\mathscr{L}_1$	$\mathscr{L}_2$	$\mathscr{L}_1$	$\mathscr{L}_2$	$\mathscr{L}_1$	$\mathscr{L}_2$	$\mathscr{L}_1$
Correlation	BestFirst	74.64	70.65	93.84	94.72	95.0	93.75	93.67	89.10
	GreedyStep	74.64	67.75	93.84	94.72	95.0	93.75	93.76	89.10
	LinearFwd	74.64	67.75	93.84	94.72	95.0	93.75	92.09	89.98
	Rank	74.64	70.65	93.84	94.72	94.15	93.75	92.26	91.38
	SubsetSizeFwd	74.64	70.65	93.84	94.72	94.15	93.75	92.09	89.98
Consistent	BestFirst	67.39	72.46	95.31	95.89	88.75	86.25	89.98	92.97
	GreedyStep	70.29	67.75	95.75	95.01	87.0	86.25	-	-
	LinearFwd	70.65	71.01	94.72	94.43	88.75	87.5	87.34	89.28
	Rank	68.48	68.48	94.57	92.08	92.5	91.25	93.67	91.91
	SubsetSizeFwd	70.65	71.01	94.72	94.43	88.75	88.75	87.34	89.28
Filtered	BestFirst	77.54	70.29	94.43	94.14	91.25	91.25	93.14	91.56
	GreedyStep	77.54	70.29	94.43	94.14	91.25	91.25	93.14	91.56
	LinearFwd	77.54	70.29	94.43	94.14	91.25	91.25	90.51	87.34
	Rank	77.14	70.29	94.43	94.14	88.75	92.5	84.88	85.86
	SubsetSizeFwd	77.89	70.29	94.43	94.14	91.25	91.25	90.51	87.34
Integ. Dim. Red.		75.36±0.49		96.34±0.19		91.25±0.29		94.55±0.23	

Table: Comparison with Dimension Reduction Weka's Feature Selection Methods followed by SVM-2norm ($\mathscr{L}_2$) and SVM-1norm ($\mathscr{L}_1$)

Experiments

		Transfusion		Vote		Tic-Tac-Toe	
Subset Evaluator	Search Method	$\mathscr{L}_2$	$\mathscr{L}_1$	$\mathscr{L}_2$	$\mathscr{L}_1$	$\mathscr{L}_2$	$\mathscr{L}_1$
Correlation	BestFirst	91.04	91.04	96.28	96.28	80.88	73.88
	GreedyStep	91.04	91.04	96.28	96.28	80.88	73.88
	LinearFwd	91.04	91.04	96.28	96.28	80.88	73.88
	Rank	91.04	91.04	96.28	96.28	80.88	73.88
	SubsetSizeFwd	91.04	91.04	96.28	96.28	80.88	73.88
Consistent	BestFirst	92.37	90.78	95.26	96.28	100.0	76.49
	GreedyStep	91.18	71.92	94.40	93.10	99.58	98.33
	LinearFwd	91.44	90.78	97.41	95.69	99.68	75.44
	Rank	91.57	89.84	94.6	93.1	99.68	80.45
	SubsetSizeFwd	91.04	91.04	96.28	96.28	99.68	75.44
Filtered	BestFirst	90.1	90.1	96.28	96.28	70.01	70.01
	GreedyStep	90.1	90.1	96.28	96.28	70.01	70.01
	LinearFwd	90.1	90.1	96.28	96.28	70.01	70.01
	Rank	91.04	91.04	96.28	96.28	70.01	70.01
	SubsetSizeFwd	90.1	90.1	96.28	96.28	70.01	70.01
Integ. Dim. Red.		91.04±0.30		96.28±0.17		100.0±0.0	

Table: Comparison with Dimension Reduction Weka's Feature Selection Methods followed by SVM-2norm ($\mathscr{L}_2$) and SVM-1norm ($\mathscr{L}_1$)

Experiments

Subset Evaluator	Search Method	Monk-1 $\mathscr{L}_2$	Monk-1 $\mathscr{L}_1$	Monk-2 $\mathscr{L}_2$	Monk-2 $\mathscr{L}_1$	Monk-3 $\mathscr{L}_2$	Monk-3 $\mathscr{L}_1$
Correlation	BestFirst	69.37	74.94	63.34	59.63	97.22	97.22
	GreedyStep	69.37	74.94	63.34	59.63	97.22	97.22
	LinearFwd	69.37	74.94	63.34	59.63	97.22	97.22
	Rank	69.37	74.94	63.34	59.63	97.22	97.22
	SubsetSizeFwd	69.37	74.94	63.34	59.63	97.22	97.22
Consistent	BestFirst	100.0	83.29	93.27	59.63	93.03	97.22
	GreedyStep	100.0	74.94	87.93	59.63	93.5	97.22
	LinearFwd	100.0	83.3	93.27	59.63	97.22	97.22
	Rank	100.0	74.94	91.87	59.63	93.27	97.22
	SubsetSizeFwd	83.3	66.59	93.27	59.63	97.22	97.22
Filtered	BestFirst	74.94	74.94	62.41	59.63	97.22	97.22
	GreedyStep	74.94	74.94	62.41	59.63	97.22	97.22
	LinearFwd	74.94	74.94	62.41	59.63	97.22	97.22
	Rank	74.94	74.94	62.41	59.63	97.22	97.22
	SubsetSizeFwd	74.94	74.94	62.41	59.63	97.22	97.22
Integ. Dim. Red.		100.0±0.0		85.15±0.38		97.22±0.16	

Table: Comparison with Dimension Reduction Weka's Feature Selection Methods followed by SVM-2norm ($\mathscr{L}_2$) and SVM-1norm ($\mathscr{L}_1$)

Learning First Order Features

- Statistical Learner constructs model from features identified by a relational learner
- Relational Features $\leftrightarrow$ First Order Logic Clauses

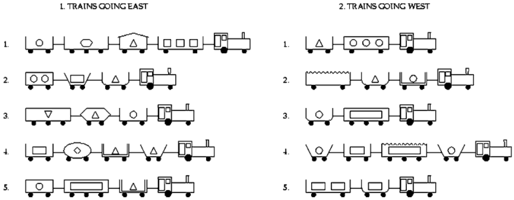


Figure: Reproduced from $Michalski's$ famous trains example

Learning First Order Features

- Statistical Learner constructs model from features identified by a relational learner
- Relational Features $\leftrightarrow$ First Order Logic Clauses

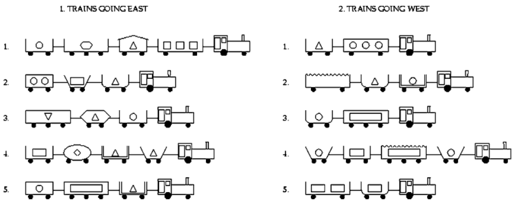


Figure: Reproduced from *Michalski's* famous trains example

What kind of features are useful?

- $eastbound(A) \leftarrow hasCar(A,B),\ hasCar(A,C),\ short(B),\ closed(C)$
- Is there some redundancy?
- Does this feature need to be learnt explicitly by the relational learner?

What kind of features are useful?

- $eastbound(A) \leftarrow hasCar(A,B),\ hasCar(A,C),\ short(B),\ closed(C)$
- Is there some redundancy?
- Does this feature need to be learnt explicitly by the relational learner?

What kind of features are useful?

- $eastbound(A) \leftarrow hasCar(A,B),\ hasCar(A,C),\ short(B),\ closed(C)$

- Is there some redundancy?
- Does this feature need to be learnt explicitly by the relational learner?

Feature Classes: Definite Feature

- *Definite Features* (F_d) [SMSK96]
  - *From* Definite Clauses *i.e.* **Clauses with non-empty head**

Feature Classes: Definite Feature

- *Definite Features* (F_d) [SMSK96]
  - *From* Definite Clauses *i.e.* **Clauses with non-empty head**

Feature Classes: Independent Feature

- *Independent Features* (F_i) [CSC+02]
  - Definite Clauses consisting of exactly **1 independent component**
  - $eastbound(A) \leftarrow hasCar(A,B), hasCar(A,C), short(B), closed(C)$
  - Independent Clauses
    - $1 : (eastbound(A) \leftarrow hasCar(A,B),\ short(B))$
    - $2 : (eastbound(A) \leftarrow hasCar(A,C),\ closed(C))$

Feature Classes: Independent Feature

- *Independent Features* (F_i) [CSC+02]
  - Definite Clauses consisting of exactly **1 independent component**
  - $eastbound(A) \leftarrow hasCar(A,B), hasCar(A,C), short(B), closed(C)$
  - Independent Clauses
    - $1 : (eastbound(A) \leftarrow hasCar(A,B),\ short(B))$
    - $2 : (eastbound(A) \leftarrow hasCar(A,C),\ closed(C))$

Feature Classes: Independent Feature

- $Independent\ Features\ (F_i)$ [CSC+02]
  - Definite Clauses consisting of exactly **1 independent component**
  - $eastbound(A) \leftarrow hasCar(A,B), hasCar(A,C), short(B), closed(C)$
  - Independent Clauses
    - $1 : (eastbound(A) \leftarrow hasCar(A,B),\ short(B))$
    - $2 : (eastbound(A) \leftarrow hasCar(A,C),\ closed(C))$

Feature Classes: Independent Feature

- *Independent Features* (F_i) [CSC+02]
  - Definite Clauses consisting of exactly **1 independent component**
  - $eastbound(A) \leftarrow hasCar(A,B), hasCar(A,C), short(B), closed(C)$
  - Independent Clauses
    - $1 : (eastbound(A) \leftarrow hasCar(A,B),\ short(B))$
    - $2 : (eastbound(A) \leftarrow hasCar(A,C),\ closed(C))$

Feature Classes: Independent Feature

- *Independent Features* (F_i) [CSC+02]
  - Definite Clauses consisting of exactly **1 independent component**
  - $eastbound(A) \leftarrow hasCar(A,B), hasCar(A,C), short(B), closed(C)$
  - Independent Clauses
    - $1 : (eastbound(A) \leftarrow hasCar(A,B),\ short(B))$
    - $2 : (eastbound(A) \leftarrow hasCar(A,C),\ closed(C))$

Feature Classes: Independent Feature

- *Independent Features* (F_i) [CSC+02]
  - Definite Clauses consisting of exactly **1 independent component**
  - $eastbound(A) \leftarrow hasCar(A,B), hasCar(A,C), short(B), closed(C)$
  - Independent Clauses
    - $1 : (eastbound(A) \leftarrow hasCar(A,B),\ short(B))$
    - $2 : (eastbound(A) \leftarrow hasCar(A,C),\ closed(C))$

Feature Classes: Relational Subgroup Discovery Feature

- *RSDFeatures* (F_r) [LZF02]
  - Independent Features with **no unused variable** in clause body
  - $eastbound(A) \leftarrow hasCar(A,B)$ is **independent but not RSD feature**

Feature Classes: Relational Subgroup Discovery Feature

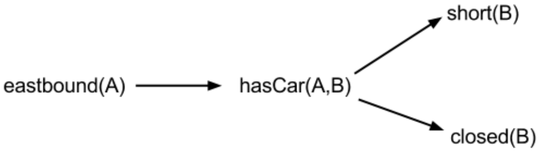
- $RSDFeatures\ (F_r)$ [LZF02]
  - Independent Features with **no unused variable** in clause body
  - $eastbound(A) \leftarrow hasCar(A,B)$ is **independent but not RSD feature**

Feature Classes: Relational Subgroup Discovery Feature

- $RSDFeatures\ (F_r)$ [LZF02]
    - Independent Features with **no unused variable** in clause body
    - $eastbound(A) \leftarrow hasCar(A,B)$ is **independent but not RSD feature**

Feature Classes : Simple Feature

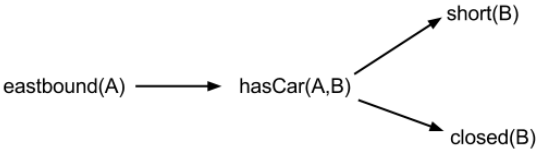
- *Simple Features* (F_s) [MS98]
  - Independent Features with one sink in the variable dependency graph



  - $eastbound(A) \leftarrow hasCar(A,B), short(B), closed(B)$ is **Independent but not simple**
  - Simple Features
    - $1 : eastbound(A) \leftarrow hasCar(A,B), short(B)$
    - $2 : eastbound(A) \leftarrow hasCar(A,B), closed(B)$

Feature Classes : Simple Feature

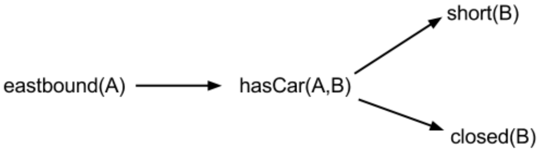
- *Simple Features* (F_s) [MS98]
  - Independent Features with one sink in the variable dependency graph



  - $eastbound(A) \leftarrow hasCar(A,B), short(B), closed(B)$ is **Independent but not simple**
  - Simple Features
    - $1 : eastbound(A) \leftarrow hasCar(A,B), short(B)$
    - $2 : eastbound(A) \leftarrow hasCar(A,B), closed(B)$

Feature Classes : Simple Feature

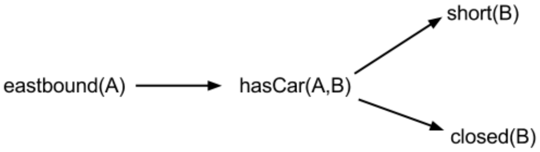
- *Simple Features* (F_s) [MS98]
  - Independent Features with one sink in the variable dependency graph



  - $eastbound(A) \leftarrow hasCar(A,B), short(B), closed(B)$ is **Independent but not simple**
  - Simple Features
    - $1 : eastbound(A) \leftarrow hasCar(A,B), short(B)$
    - $2 : eastbound(A) \leftarrow hasCar(A,B), closed(B)$

Feature Classes : Simple Feature

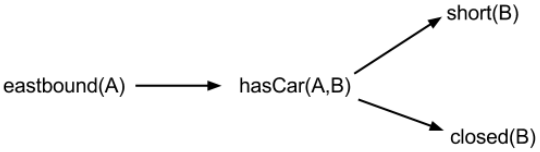
- *Simple Features* (F_s) [MS98]
  - Independent Features with one sink in the variable dependency graph



  - $eastbound(A) \leftarrow hasCar(A,B), short(B), closed(B)$ is **Independent but not simple**
  - Simple Features
    - $1 : eastbound(A) \leftarrow hasCar(A,B), short(B)$
    - $2 : eastbound(A) \leftarrow hasCar(A,B), closed(B)$

Feature Classes : Simple Feature

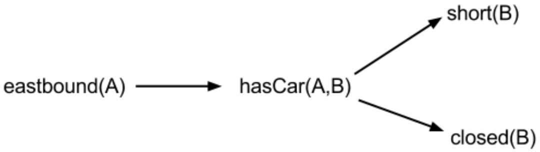
- *Simple Features* (F_s) [MS98]
  - Independent Features with one sink in the variable dependency graph



  - $eastbound(A) \leftarrow hasCar(A,B), short(B), closed(B)$ is **Independent but not simple**
  - Simple Features
    - $1 : eastbound(A) \leftarrow hasCar(A,B), short(B)$
    - $2 : eastbound(A) \leftarrow hasCar(A,B), closed(B)$

Feature Classes : Simple Feature

- *Simple Features* (F_s) [MS98]
  - Independent Features with one sink in the variable dependency graph



  - $eastbound(A) \leftarrow hasCar(A,B), short(B), closed(B)$ is **Independent but not simple**
  - Simple Features
    - $1 : eastbound(A) \leftarrow hasCar(A,B), short(B)$
    - $2 : eastbound(A) \leftarrow hasCar(A,B), closed(B)$

Feature Classes: Elementary Feature

- Elementary Feature [FL00]
  - Simple Features with **no unused variable** in body
    - $eastbound(A) \leftarrow hasCar(A,B)$ is **simple but not elementary feature**

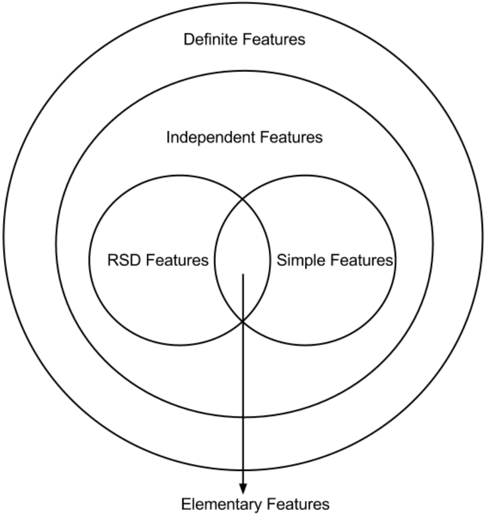
Feature Classes: Elementary Feature

- Elementary Feature [FL00]
  - Simple Features with **no unused variable** in body
    - $eastbound(A) \leftarrow hasCar(A,B)$ is **simple but not elementary feature**

Feature Classes: Elementary Feature

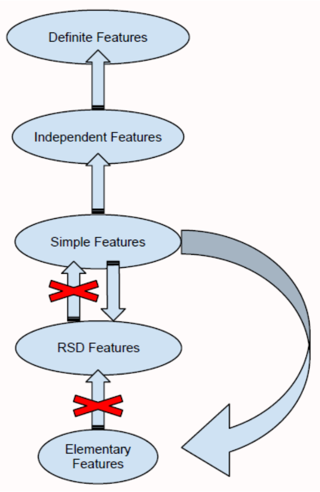
- Elementary Feature [FL00]
  - Simple Features with **no unused variable** in body
    - $eastbound(A) \leftarrow hasCar(A,B)$ is **simple but not elementary feature**

Subset Relation between Feature Classes



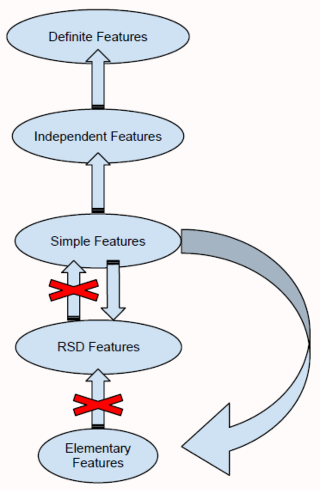
Reconstruction Property of Feature Classes

- Reconstruction by statistical learners using
  - exact logical operations (conjunctions)
  - approximation by weighted linear combinations



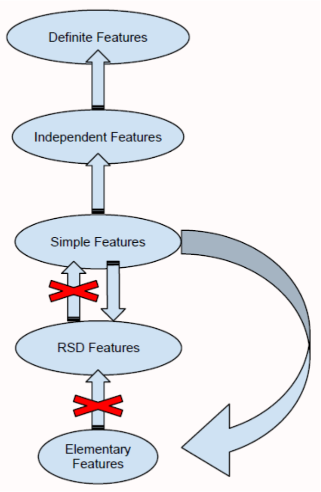
Reconstruction Property of Feature Classes

- Reconstruction by statistical learners using
  - exact logical operations (conjunctions)
  - approximation by weighted linear combinations



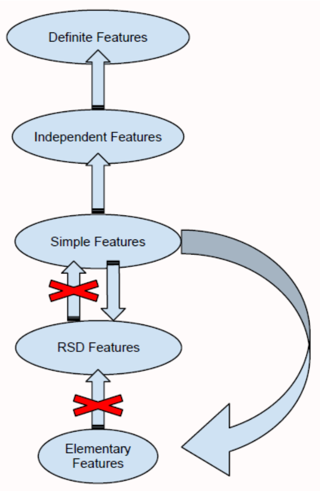
Reconstruction Property of Feature Classes

- Reconstruction by statistical learners using
  - exact logical operations (conjunctions)
  - approximation by weighted linear combinations



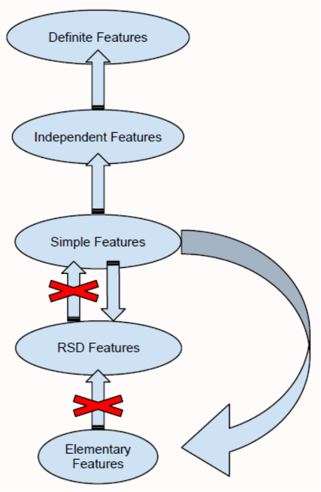
Reconstruction Property of Feature Classes

- Reconstruction by statistical learners using
  - exact logical operations (conjunctions)
  - approximation by weighted linear combinations



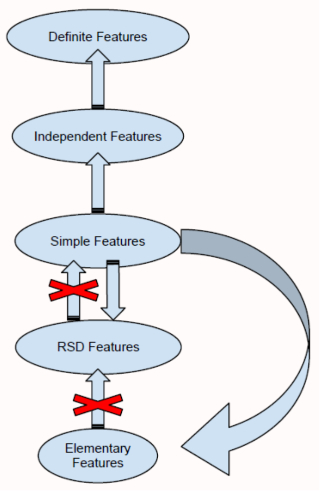
Reconstruction Property of Feature Classes

- Reconstruction by statistical learners using
  - exact logical operations (conjunctions)
  - approximation by weighted linear combinations



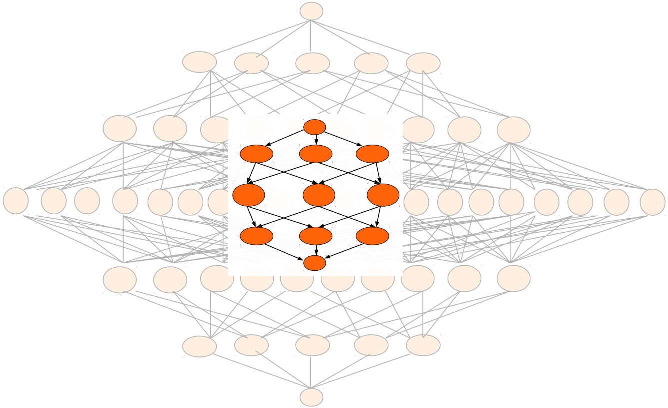
Reconstruction Property of Feature Classes

- Reconstruction by statistical learners using
  - exact logical operations (conjunctions)
  - approximation by weighted linear combinations



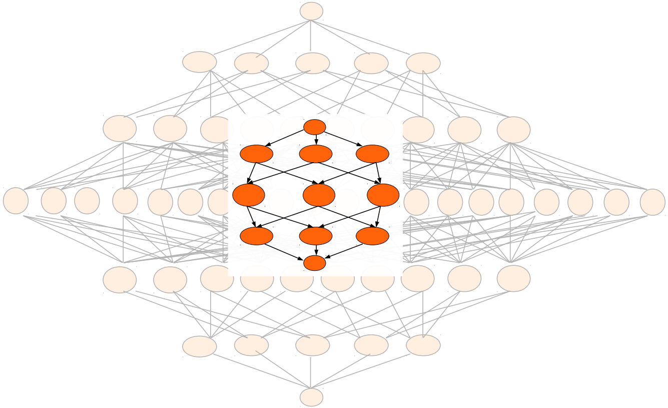
Goal: Discovering most effective subclass

- The subclass that is yields most accurate statistical models
- The subclass that can be logically composed to yield other powerful subclasses



Goal: Discovering most effective subclass

- The subclass that is yields most accurate statistical models
- The subclass that can be logically composed to yield other powerful subclasses



Experiment-I: SVM as Model Builder

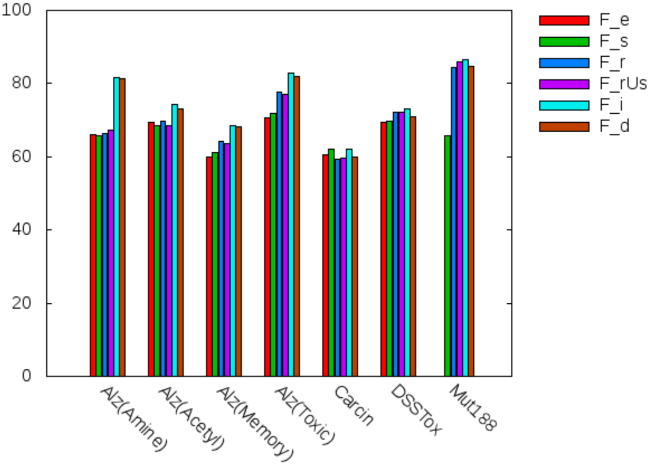


Figure: Accuracy of the Models Built by SVM-2 norm regularizer(LibSVM) [CL11] on the feature classes F_e, F_s, F_r, $F_{r\cup s}$, F_i, F_d

Experiment-I: SVM as Model Builder

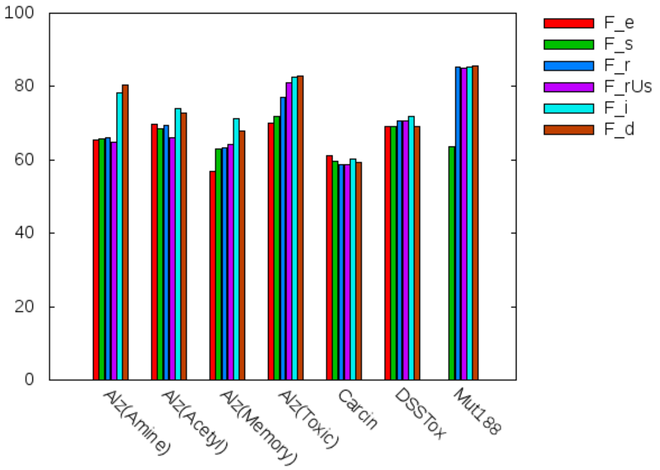


Figure: Accuracy of the Models Built by SVM-1 norm regularizer(LibLinear) [FCH+08] on the feature classes F_e, F_s, F_r, $F_{r \cup s}$, F_i, F_d

Experiment-II: LR as Model Builder

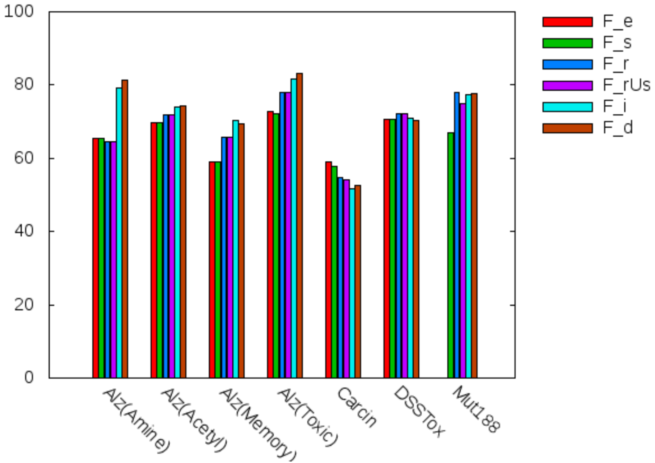


Figure: Accuracy of the Models Built by Logistic Regression on the feature classes F_e, F_s, F_r, $F_{r \cup s}$, F_i, F_d

Experiment-II: LR as Model Builder

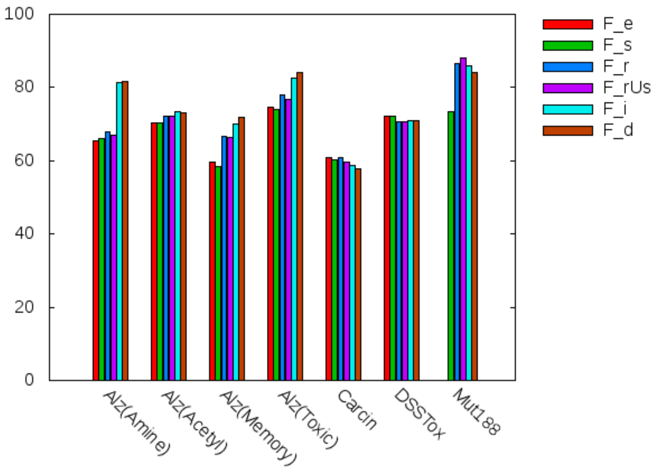


Figure: Accuracy of the Models Built by Sparse Multinomial Logistic Regression [KCFH05] on the feature classes F_e, F_s, F_r, $F_{r\cup s}$, F_i, F_d

Experiment-III: Rule Ensemble learner MLRules

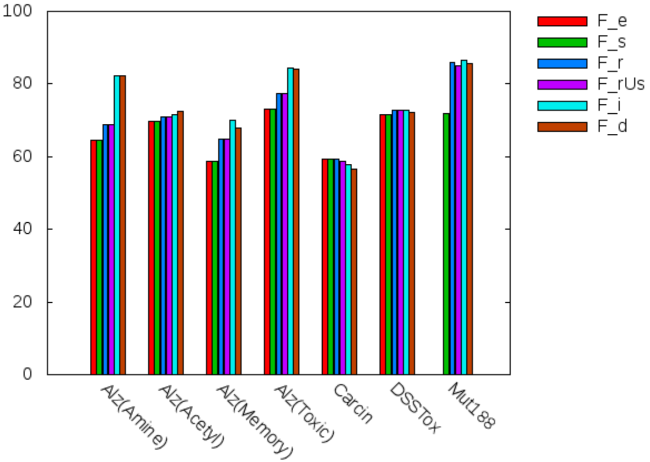


Figure: Accuracy of Model Built by Maximum Likelihood Rule Ensembles [DKS08] on the feature classes F_e, F_s, F_r, $F_{r\cup s}$, F_i, F_d

How significantly different?

Feature Class	Number of Wins					Total Wins
	L1-SVM	L2-SVM	LR	SMLR	MLRules	
F_e	1	0	1	1	1	4/35
F_s	0	1	0	1	0	2/35
F_r	0	1	1	1	0	3/35
$F_{r\cup s}$	0	0	1	1	1	3/35
F_i	3	6	1	1	4	15/35
F_d	3	0	4	3	1	11/35

Figure: Number of outright wins for a feature class. This is number of occasions out of the total number of possible occasions (*i.e.* 35) on which a statistical learners achieves the highest mean predictive accuracy using features from that class.

How significantly different?

Feature Class	Number of Good Enough Models					Total No. of Good Models
	L1-SVM	L2-SVM	LR	SMLR	MLRules	
F_e	2	1	2	2	2	9/35
F_s	2	1	2	2	2	9/35
F_r	2	2	1	4	4	13/35
$F_{r \cup s}$	3	2	1	4	4	14/35
F_i	7	7	6	7	7	34/35
F_d	4	5	6	7	7	25/35

Figure: Number of good enough models (out of all possible models *i.e.* 35), using a feature class. A model is taken to be good enough if its predictive accuracy is not statistically different to the model with the highest predictive accuracy.

Comparison with Parameter-tuned ILP models

Data	Statistical Model	ILP Model With Parameter Selection & Optimization
Alz (Amine)	82.32±1.18	80.20
Alz (Acetyl)	74.16±0.24	77.40
Alz (Memory)	71.83±1.67	67.40
Alz (Toxic)	84.50±0.44	87.20
Carcin	62.15±1.75	59.10
DSSTox	73.12±0.94	73.10
Mut(188)	88.06±1.57	88.30

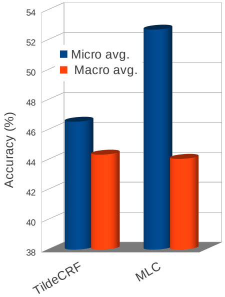
Figure: Comparison of mean predictive accuracies of statistical models against the ILP models constructed with parameter selection and optimisation (see [SR11]).

Results for Sequence Labeling

Katholieke Universiteit Data [Landwehr *et al.*, 2009] (KU Data)

	Micro avg.(%)	Macro avg.(%)	F score
TildeCRF: Baseline	46.54	44.34	0.51
Independent Features (F_i)	52.69	44.06	0.57

- 25 sensors
- 19 activities
- 20 sequences of around 250 time steps

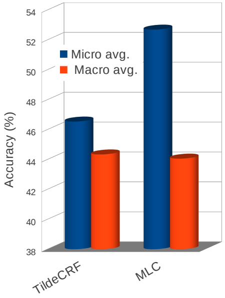


Results for Sequence Labeling

Katholieke Universiteit Data [Landwehr *et al.*, 2009] (KU Data)

	Micro avg.(%)	Macro avg.(%)	F score
TildeCRF: Baseline	46.54	44.34	0.51
Independent Features (F_i)	52.69	44.06	0.57

- 25 sensors
- 19 activities
- 20 sequences of around 250 time steps

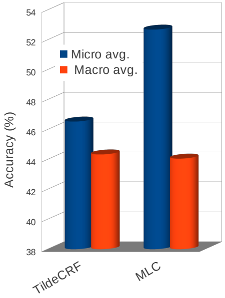


Results for Sequence Labeling

Katholieke Universiteit Data [Landwehr *et al.*, 2009] (KU Data)

	Micro avg.(%)	Macro avg.(%)	F score
TildeCRF: Baseline	46.54	44.34	0.51
Independent Features (F_i)	52.69	44.06	0.57

- 25 sensors
- 19 activities
- 20 sequences of around 250 time steps

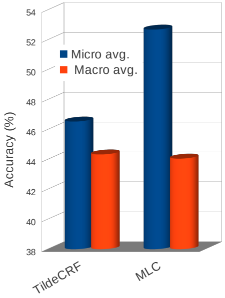


Results for Sequence Labeling

Katholieke Universiteit Data [Landwehr *et al.*, 2009] (KU Data)

	Micro avg.(%)	Macro avg.(%)	F score
TildeCRF: Baseline	46.54	44.34	0.51
Independent Features (F_i)	52.69	44.06	0.57

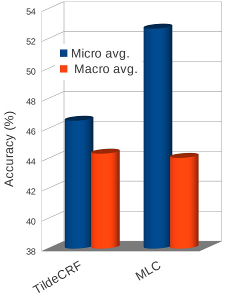
- 25 sensors
- 19 activities
- 20 sequences of around 250 time steps



Katholieke Universiteit Data [Landwehr *et al.*, 2009] (KU Data)

	Micro avg.(%)	Macro avg.(%)	F score
TildeCRF: Baseline	46.54	44.34	0.51
Independent Features (F_i)	52.69	44.06	0.57

- 25 sensors
- 19 activities
- 20 sequences of around 250 time steps

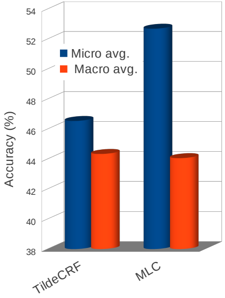


Results for Sequence Labeling

Katholieke Universiteit Data [Landwehr *et al.*, 2009] (KU Data)

	Micro avg.(%)	Macro avg.(%)	F score
TildeCRF: Baseline	46.54	44.34	0.51
Independent Features (F_i)	52.69	44.06	0.57

- 25 sensors
- 19 activities
- 20 sequences of around 250 time steps



Summary and Conclusion

- Presented methods for learning statistically optimal features
  - Improved generalization
  - **Bridged gap** between statisical and rule learning communities
  - Investigated feature in which class are worth learning
- Two-fold objective: Learning short and few features
  - Applicable elsewhere
- Efficient mirror-descent based active set method

Summary and Conclusion

- Presented methods for learning statistically optimal features
  - Improved generalization
  - **Bridged gap** between statisical and rule learning communities
  - Investigated feature in which class are worth learning
- Two-fold objective: Learning short and few features
  - Generalizes well while learning compact ruleset
  - Sometimes 25% improvement in generalization
  - Applicable elsewhere
- Efficient mirror-descent based active set method
  - Complexity: polynomial in active set size
  - Searched rule space size

Summary and Conclusion

- Presented methods for learning statistically optimal features
  - Improved generalization
  - **Bridged gap** between statisical and rule learning communities
  - Investigated feature in which class are worth learning
- Two-fold objective: Learning short and few features
- Efficient mirror-descent based active set method

Summary and Conclusion

- Presented methods for learning statistically optimal features
  - Improved generalization
  - **Bridged gap** between statisical and rule learning communities
  - Investigated feature in which class are worth learning
- Two-fold objective: Learning short and few features
  - Applicable elsewhere
- Efficient mirror-descent based active set method

Summary and Conclusion

- Presented methods for learning statistically optimal features
  - Improved generalization
  - **Bridged gap** between statisical and rule learning communities
  - Investigated feature in which class are worth learning
- Two-fold objective: Learning short and few features
  - Generalizes well while learning compact ruleset
  - Sometimes 25% improvement in generalization
  - Applicable elsewhere
- Efficient mirror-descent based active set method
  - Complexity: polynomial in active set size ($\ll O(2^n)$)
  - Searched rule space size $\sim 2^{50}$ in ~ 10 min.

Summary and Conclusion

- Presented methods for learning statistically optimal features
  - Improved generalization
  - **Bridged gap** between statisical and rule learning communities
  - Investigated feature in which class are worth learning
- Two-fold objective: Learning short and few features
  - Generalizes well while learning compact ruleset
  - Sometimes **25%** improvement in generalization
  - Applicable elsewhere
- Efficient mirror-descent based active set method
  - Complexity: polynomial in active set size ($\ll O(2^n)$)
  - Searched rule space size $\sim 2^{50}$ in $\sim$ 10 min.

Summary and Conclusion

- Presented methods for learning statistically optimal features
  - Improved generalization
  - **Bridged gap** between statisical and rule learning communities
  - Investigated feature in which class are worth learning
- Two-fold objective: Learning short and few features
  - Generalizes well while learning compact ruleset
  - Sometimes **25%** improvement in generalization
  - Applicable elsewhere
- Efficient mirror-descent based active set method

Summary and Conclusion

- Presented methods for learning statistically optimal features
  - Improved generalization
  - **Bridged gap** between statisical and rule learning communities
  - Investigated feature in which class are worth learning
- Two-fold objective: Learning short and few features
  - Generalizes well while learning compact ruleset
  - Sometimes **25%** improvement in generalization
  - Applicable elsewhere
- Efficient mirror-descent based active set method
  - Complexity: polynomial in active set size ($\ll O(2^n)$)
  - Searched rule space size $\sim 2^{50}$ in <10 min.

Summary and Conclusion

- Presented methods for learning statistically optimal features
  - Improved generalization
  - **Bridged gap** between statisical and rule learning communities
  - Investigated feature in which class are worth learning
- Two-fold objective: Learning short and few features
  - Generalizes well while learning compact ruleset
  - Sometimes **25%** improvement in generalization
  - Applicable elsewhere
- Efficient mirror-descent based active set method

Summary and Conclusion

- Presented methods for learning statistically optimal features
  - Improved generalization
  - **Bridged gap** between statisical and rule learning communities
  - Investigated feature in which class are worth learning
- Two-fold objective: Learning short and few features
  - Generalizes well while learning compact ruleset
  - Sometimes **25%** improvement in generalization
  - Applicable elsewhere
- Efficient mirror-descent based active set method

Summary and Conclusion

- Presented methods for learning statistically optimal features
    - Improved generalization
    - **Bridged gap** between statisical and rule learning communities
    - Investigated feature in which class are worth learning
- Two-fold objective: Learning short and few features
    - Generalizes well while learning compact ruleset
    - Sometimes **25%** improvement in generalization
    - Applicable elsewhere
- Efficient mirror-descent based active set method
    - Complexity: polynomial in active set size ($\ll O(2^n)$)
    - Searched rule space size $\sim 2^{50}$ **in** ~ 10 **min.**

Summary and Conclusion

- Presented methods for learning statistically optimal features
  - Improved generalization
  - **Bridged gap** between statisical and rule learning communities
  - Investigated feature in which class are worth learning
- Two-fold objective: Learning short and few features
  - Generalizes well while learning compact ruleset
  - Sometimes **25%** improvement in generalization
  - Applicable elsewhere
- Efficient mirror-descent based active set method
  - Complexity: polynomial in active set size ($\ll O(2^n)$)
  - Searched rule space size $\sim 2^{50}$ **in** ~ 10 **min.**

Summary and Conclusion

- Presented methods for learning statistically optimal features
  - Improved generalization
  - **Bridged gap** between statisical and rule learning communities
  - Investigated feature in which class are worth learning
- Two-fold objective: Learning short and few features
  - Generalizes well while learning compact ruleset
  - Sometimes **25%** improvement in generalization
  - Applicable elsewhere
- Efficient mirror-descent based active set method
  - Complexity: polynomial in active set size ($\ll O(2^n)$)
  - Searched rule space size $\sim 2^{50}$ **in** ~ 10 **min.**

THANK YOU

Bibliography

- Francis Bach. *High-dimensional non-linear variable selection through hierarchical kernel learning.* CoRR, abs/0909.0844, 2009.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. *Latent dirichlet allocation.* J. Mach. Learn. Res., 3:993–1022, March 2003.
- *Chih-Chung Chang and Chih-Jen Lin.* LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology,
- Vitor Santos Costa, Ashwin Srinivasan, Rui Camacho, Hendrik Blockeel, Bart Demoen, Gerda Janssens, Wim Van Laer, James Cussens, and Alan Frisch.*Query transformations for improving the efficiency of ilp systems.* Journal of Machine Learning Research, 4:491, 2002.
- Krzysztof Dembczynski, Wojciech Kotlowski, and Roman Slowinski.*Maximum likelihood rule ensembles.* In ICML, pages 224–231, 2008.

Bibliography

Francis Bach. *High-dimensional non-linear variable selection through hierarchical kernel learning.* CoRR, abs/0909.0844, 2009.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. *Latent dirichlet allocation*. J. Mach. Learn. Res., 3:993–1022, March 2003.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology,

Vitor Santos Costa, Ashwin Srinivasan, Rui Camacho, Hendrik Blockeel, Bart Demoen, Gerda Janssens, Wim Van Laer, James Cussens, and Alan Frisch.*Query transformations for improving the efficiency of ilp systems*. Journal of Machine Learning Research, 4:491, 2002.

Krzysztof Dembczynski, Wojciech Kotlowski, and Roman Slowinski.*Maximum likelihood rule ensembles*. In ICML, pages 224–231, 2008.

Bibliography

Francis Bach. *High-dimensional non-linear variable selection through hierarchical kernel learning.* CoRR, abs/0909.0844, 2009.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. *Latent dirichlet allocation*. J. Mach. Learn. Res., 3:993–1022, March 2003.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology,

Vitor Santos Costa, Ashwin Srinivasan, Rui Camacho, Hendrik Blockeel, Bart Demoen, Gerda Janssens, Wim Van Laer, James Cussens, and Alan Frisch.*Query transformations for improving the efficiency of ilp systems*. Journal of Machine Learning Research, 4:491, 2002.

Krzysztof Dembczynski, Wojciech Kotlowski, and Roman Slowinski.*Maximum likelihood rule ensembles*. In ICML, pages 224–231, 2008.

Bibliography

Francis Bach. *High-dimensional non-linear variable selection through hierarchical kernel learning.* CoRR, abs/0909.0844, 2009.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. *Latent dirichlet allocation*. J. Mach. Learn. Res., 3:993–1022, March 2003.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology,

Vitor Santos Costa, Ashwin Srinivasan, Rui Camacho, Hendrik Blockeel, Bart Demoen, Gerda Janssens, Wim Van Laer, James Cussens, and Alan Frisch.*Query transformations for improving the efficiency of ilp systems.* Journal of Machine Learning Research, 4:491, 2002.

Krzysztof Dembczynski, Wojciech Kotlowski, and Roman Slowinski.*Maximum likelihood rule ensembles.* In ICML, pages 224–231, 2008.

Bibliography

Francis Bach. *High-dimensional non-linear variable selection through hierarchical kernel learning.* CoRR, abs/0909.0844, 2009.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. *Latent dirichlet allocation*. J. Mach. Learn. Res., 3:993–1022, March 2003.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology,

Vitor Santos Costa, Ashwin Srinivasan, Rui Camacho, Hendrik Blockeel, Bart Demoen, Gerda Janssens, Wim Van Laer, James Cussens, and Alan Frisch.*Query transformations for improving the efficiency of ilp systems.* Journal of Machine Learning Research, 4:491, 2002.

Krzysztof Dembczynski, Wojciech Kotlowski, and Roman Slowinski.*Maximum likelihood rule ensembles.* In ICML, pages 224–231, 2008.

Bibliography

- Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin.*LIBLINEAR: A library for large linear classification.* Journal of Machine Learning Research, 9:1871–1874, 2008.
- Peter A. Flach and Nicolas Lachiche.-order bayesian classification with 1bc. 2000.
- Pratik Jawanpuria, Jagarlapudi Saketha Nath, and Ganesh Ramakrishnan.*Efficient rule ensemble learning using hierarchical kernels.* In ICML, pages 161–168, 2011.
- I. T. Jolliffe. *Principal Component Analysis.*Springer-Verlag, 1986.
- Balaji Krishnapuram, Lawrence Carin, Mario A. T. Figueiredo, and Alexander J. Hartemink.*Sparse multinomial logistic regression: Fast algorithms and generalization bounds.* IEEE Trans. Pattern Anal. Mach. Intell., 27(6):957–968, June 2005.

Bibliography

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. *LIBLINEAR: A library for large linear classification.* Journal of Machine Learning Research, 9:1871–1874, 2008.

Peter A. Flach and Nicolas Lachiche.-order bayesian classification with 1bc. 2000.

Pratik Jawanpuria, Jagarlapudi Saketha Nath, and Ganesh Ramakrishnan. *Efficient rule ensemble learning using hierarchical kernels.* In ICML, pages 161–168, 2011.

I. T. Jolliffe. *Principal Component Analysis.* Springer-Verlag, 1986.

Balaji Krishnapuram, Lawrence Carin, Mario A. T. Figueiredo, and Alexander J. Hartemink. *Sparse multinomial logistic regression: Fast algorithms and generalization bounds.* IEEE Trans. Pattern Anal. Mach. Intell., 27(6):957–968, June 2005.

Bibliography

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin.*LIBLINEAR: A library for large linear classification.* Journal of Machine Learning Research, 9:1871–1874, 2008.

Peter A. Flach and Nicolas Lachiche.-order bayesian classification with 1bc. 2000.

Pratik Jawanpuria, Jagarlapudi Saketha Nath, and Ganesh Ramakrishnan.*Efficient rule ensemble learning using hierarchical kernels.* In ICML, pages 161–168, 2011.

I. T. Jolliffe. *Principal Component Analysis.*Springer-Verlag, 1986.

Balaji Krishnapuram, Lawrence Carin, Mario A. T. Figueiredo, and Alexander J. Hartemink.*Sparse multinomial logistic regression: Fast algorithms and generalization bounds.* IEEE Trans. Pattern Anal. Mach. Intell., 27(6):957–968, June 2005.

Bibliography

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin.*LIBLINEAR: A library for large linear classification.* Journal of Machine Learning Research, 9:1871–1874, 2008.

Peter A. Flach and Nicolas Lachiche.-order bayesian classification with 1bc. 2000.

Pratik Jawanpuria, Jagarlapudi Saketha Nath, and Ganesh Ramakrishnan.*Efficient rule ensemble learning using hierarchical kernels.* In ICML, pages 161–168, 2011.

I. T. Jolliffe. *Principal Component Analysis.*Springer-Verlag, 1986.

Balaji Krishnapuram, Lawrence Carin, Mario A. T. Figueiredo, and Alexander J. Hartemink.*Sparse multinomial logistic regression: Fast algorithms and generalization bounds.* IEEE Trans. Pattern Anal. Mach. Intell., 27(6):957–968, June 2005.

Bibliography

- Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin.*LIBLINEAR: A library for large linear classification.* Journal of Machine Learning Research, 9:1871–1874, 2008.
- Peter A. Flach and Nicolas Lachiche.-order bayesian classification with 1bc. 2000.
- Pratik Jawanpuria, Jagarlapudi Saketha Nath, and Ganesh Ramakrishnan.*Efficient rule ensemble learning using hierarchical kernels.* In ICML, pages 161–168, 2011.
- I. T. Jolliffe. *Principal Component Analysis.*Springer-Verlag, 1986.
- Balaji Krishnapuram, Lawrence Carin, Mario A. T. Figueiredo, and Alexander J. Hartemink.*Sparse multinomial logistic regression: Fast algorithms and generalization bounds.* IEEE Trans. Pattern Anal. Mach. Intell., 27(6):957–968, June 2005.

Bibliography

K. Lang. *20 newsgroups data set.*

Niels Landwehr, Bernd Gutmann, Ingo Thon, Luc De Raedt, and Matthai Philipose. *Relational transformation-based tagging for activity recognition.* Progress on Multi-Relational Data Mining, 89(1):111–129, 2009.

Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. *Disclda: Discriminative learning for dimensionality reduction and classification.* In NIPS, pages 897–904, 2008.

Haifeng Li, Tao Jiang, and Keshu Zhang. *Efficient and robust feature extraction by maximum margin criterion.* In In Advances in Neural Information Processing Systems 16, pages 157–165. MIT Press, 2003.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

Bibliography

K. Lang.*20 newsgroups data set.*

Niels Landwehr, Bernd Gutmann, Ingo Thon, Luc De Raedt, and Matthai Philipose.*Relational transformation-based tagging for activity recognition.* Progress on Multi-Relational Data Mining, 89(1):111–129, 2009.

Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan.*Disclda: Discriminative learning for dimensionality reduction and classification.* In NIPS, pages 897–904, 2008.

Haifeng Li, Tao Jiang, and Keshu Zhang. *Efficient and robust feature extraction by maximum margin criterion.* In In Advances in Neural Information Processing Systems 16, pages 157–165. MIT Press, 2003.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira.*Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

Bibliography

K. Lang.*20 newsgroups data set.*

Niels Landwehr, Bernd Gutmann, Ingo Thon, Luc De Raedt, and Matthai Philipose.*Relational transformation-based tagging for activity recognition.* Progress on Multi-Relational Data Mining, 89(1):111–129, 2009.

Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan.*Disclda: Discriminative learning for dimensionality reduction and classification.* In NIPS, pages 897–904, 2008.

Haifeng Li, Tao Jiang, and Keshu Zhang. *Efficient and robust feature extraction by maximum margin criterion.* In In Advances in Neural Information Processing Systems 16, pages 157–165. MIT Press, 2003.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira.*Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

Bibliography

K. Lang. *20 newsgroups data set.*

Niels Landwehr, Bernd Gutmann, Ingo Thon, Luc De Raedt, and Matthai Philipose. *Relational transformation-based tagging for activity recognition.* Progress on Multi-Relational Data Mining, 89(1):111–129, 2009.

Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. *Disclda: Discriminative learning for dimensionality reduction and classification.* In NIPS, pages 897–904, 2008.

Haifeng Li, Tao Jiang, and Keshu Zhang. *Efficient and robust feature extraction by maximum margin criterion.* In In Advances in Neural Information Processing Systems 16, pages 157–165. MIT Press, 2003.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

Bibliography

K. Lang.*20 newsgroups data set.*

Niels Landwehr, Bernd Gutmann, Ingo Thon, Luc De Raedt, and Matthai Philipose.*Relational transformation-based tagging for activity recognition.* Progress on Multi-Relational Data Mining, 89(1):111–129, 2009.

Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan.*Disclda: Discriminative learning for dimensionality reduction and classification.* In NIPS, pages 897–904, 2008.

Haifeng Li, Tao Jiang, and Keshu Zhang. *Efficient and robust feature extraction by maximum margin criterion.* In In Advances in Neural Information Processing Systems 16, pages 157–165. MIT Press, 2003.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira.*Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

Bibliography

- Nada Lavrac, Filip Zelezny, and Peter A/ Flach. *Rsd: Relational subgroup discovery through first-order feature construction.* pages 149–165, July 2002.
- Eric McCreath and Arun Sharma.*Lime: A system for learning relations.* pages 336–374, 1998.
- Sriraam Natarajan, Hung H. Bui, Prasad Tadepalli, Kristian Kersting, and Weng keen Wong.*Logical hierarchical hidden markov models for modeling user activities.* In Proc. of ILP-08, 2008.
- Naveen Nair, Ganesh Ramakrishnan, and Shonali Krishnaswamy.*Enhancing activity recognition in smart homes using feature induction.* In Proceedings of the 13th international conference on Data warehousing and knowledge discovery, DaWaK'11, pages 406–418, Berlin, Heidelberg, 2011. Springer-Verlag.
- Naveen Nair, Amrita Saha, Ganesh Ramakrishnan, and Shonali Krishnaswamy.*Rule Ensemble Learning In Structured Output Spaces.* In Proceedings of American Assocition for Artificial Intelligence, AAAI'12, Toronto, Canada, 2012.
- L.R. Rabiner.*A tutorial on hidden markov models and selected applications in speech recognition.* Proceedings of the IEEE, 77(2):257–286, feb 1989.

Bibliography

- Nada Lavrac, Filip Zelezny, and Peter A/ Flach. *Rsd: Relational subgroup discovery through first-order feature construction.* pages 149–165, July 2002.
- Eric McCreath and Arun Sharma.*Lime: A system for learning relations.* pages 336–374, 1998.
- Sriraam Natarajan, Hung H. Bui, Prasad Tadepalli, Kristian Kersting, and Weng keen Wong.*Logical hierarchical hidden markov models for modeling user activities.* In Proc. of ILP-08, 2008.
- Naveen Nair, Ganesh Ramakrishnan, and Shonali Krishnaswamy.*Enhancing activity recognition in smart homes using feature induction.* In Proceedings of the 13th international conference on Data warehousing and knowledge discovery, DaWaK'11, pages 406–418, Berlin, Heidelberg, 2011. Springer-Verlag.
- Naveen Nair, Amrita Saha, Ganesh Ramakrishnan, and Shonali Krishnaswamy.*Rule Ensemble Learning In Structured Output Spaces.* In Proceedings of American Assocition for Artificial Intelligence, AAAI'12, Toronto, Canada, 2012.
- L.R. Rabiner.*A tutorial on hidden markov models and selected applications in speech recognition.* Proceedings of the IEEE, 77(2):257–286, feb 1989.

Bibliography

- Nada Lavrac, Filip Zelezny, and Peter A/ Flach. *Rsd: Relational subgroup discovery through first-order feature construction.* pages 149–165, July 2002.
- Eric McCreath and Arun Sharma.*Lime: A system for learning relations.* pages 336–374, 1998.
- Sriraam Natarajan, Hung H. Bui, Prasad Tadepalli, Kristian Kersting, and Weng keen Wong.*Logical hierarchical hidden markov models for modeling user activities.* In Proc. of ILP-08, 2008.
- Naveen Nair, Ganesh Ramakrishnan, and Shonali Krishnaswamy.*Enhancing activity recognition in smart homes using feature induction.* In Proceedings of the 13th international conference on Data warehousing and knowledge discovery, DaWaK'11, pages 406–418, Berlin, Heidelberg, 2011. Springer-Verlag.
- Naveen Nair, Amrita Saha, Ganesh Ramakrishnan, and Shonali Krishnaswamy.*Rule Ensemble Learning In Structured Output Spaces.* In Proceedings of American Assocition for Artificial Intelligence, AAAI'12, Toronto, Canada, 2012.
- L.R. Rabiner.*A tutorial on hidden markov models and selected applications in speech recognition.* Proceedings of the IEEE, 77(2):257–286, feb 1989.

Bibliography

Nada Lavrac, Filip Zelezny, and Peter A/ Flach. *Rsd: Relational subgroup discovery through first-order feature construction.* pages 149–165, July 2002.

Eric McCreath and Arun Sharma.*Lime: A system for learning relations.* pages 336–374, 1998.

Sriraam Natarajan, Hung H. Bui, Prasad Tadepalli, Kristian Kersting, and Weng keen Wong.*Logical hierarchical hidden markov models for modeling user activities.* In Proc. of ILP-08, 2008.

Naveen Nair, Ganesh Ramakrishnan, and Shonali Krishnaswamy.*Enhancing activity recognition in smart homes using feature induction.* In Proceedings of the 13th international conference on Data warehousing and knowledge discovery, DaWaK'11, pages 406–418, Berlin, Heidelberg, 2011. Springer-Verlag.

Naveen Nair, Amrita Saha, Ganesh Ramakrishnan, and Shonali Krishnaswamy.*Rule Ensemble Learning In Structured Output Spaces.* In Proceedings of American Assocition for Artificial Intelligence, AAAI'12, Toronto, Canada, 2012.

L.R. Rabiner.*A tutorial on hidden markov models and selected applications in speech recognition.* Proceedings of the IEEE, 77(2):257–286, feb 1989.

Bibliography

Nada Lavrac, Filip Zelezny, and Peter A/ Flach. *Rsd: Relational subgroup discovery through first-order feature construction.* pages 149–165, July 2002.

Eric McCreath and Arun Sharma.*Lime: A system for learning relations.* pages 336–374, 1998.

Sriraam Natarajan, Hung H. Bui, Prasad Tadepalli, Kristian Kersting, and Weng keen Wong.*Logical hierarchical hidden markov models for modeling user activities.* In Proc. of ILP-08, 2008.

Naveen Nair, Ganesh Ramakrishnan, and Shonali Krishnaswamy.*Enhancing activity recognition in smart homes using feature induction.* In Proceedings of the 13th international conference on Data warehousing and knowledge discovery, DaWaK'11, pages 406–418, Berlin, Heidelberg, 2011. Springer-Verlag.

Naveen Nair, Amrita Saha, Ganesh Ramakrishnan, and Shonali Krishnaswamy.*Rule Ensemble Learning In Structured Output Spaces.* In Proceedings of American Assocition for Artificial Intelligence, AAAI'12, Toronto, Canada, 2012.

L.R. Rabiner.*A tutorial on hidden markov models and selected applications in speech recognition.* Proceedings of the IEEE, 77(2):257–286, feb 1989.

Bibliography

Nada Lavrac, Filip Zelezny, and Peter A/ Flach. *Rsd: Relational subgroup discovery through first-order feature construction.* pages 149–165, July 2002.

Eric McCreath and Arun Sharma.*Lime: A system for learning relations.* pages 336–374, 1998.

Sriraam Natarajan, Hung H. Bui, Prasad Tadepalli, Kristian Kersting, and Weng keen Wong.*Logical hierarchical hidden markov models for modeling user activities.* In Proc. of ILP-08, 2008.

Naveen Nair, Ganesh Ramakrishnan, and Shonali Krishnaswamy.*Enhancing activity recognition in smart homes using feature induction.* In Proceedings of the 13th international conference on Data warehousing and knowledge discovery, DaWaK'11, pages 406–418, Berlin, Heidelberg, 2011. Springer-Verlag.

Naveen Nair, Amrita Saha, Ganesh Ramakrishnan, and Shonali Krishnaswamy.*Rule Ensemble Learning In Structured Output Spaces.* In Proceedings of American Assocition for Artificial Intelligence, AAAI'12, Toronto, Canada, 2012.

L.R. Rabiner.*A tutorial on hidden markov models and selected applications in speech recognition.* Proceedings of the IEEE, 77(2):257–286, feb 1989.

Bibliography

Marie Szafranski, Yves Grandvalet, and Alain Rakotomamonjy.*Composite kernel learning.* Machine Learning, 79(1-2):73–103, 2010.

Ashwin Srinivasan, Stephen Muggleton, Michael J. E. Sternberg, and Ross D. King.for mutagenicity: A study in first-order and feature-based induction. Artif. Intell., 85(1-2):277–299, 1996.

Ashwin Srinivasan and Ganesh Ramakrishnan.screening and optimisation for ilp using designed experiments. Journal of Machine Learning Research, 12:627–662, 2011.

Ashwin Srinivasan. *The aleph manual.* 1999.

Lucia Specia, Ashwin Srinivasan, Sachindra Joshi, Ganesh Ramakrishnan, and Maria das Graças Volpe Nunes.*An investigation into feature construction to assist word sense disambiguation.* Machine Learning, 76(1):109–136, 2009.

Bibliography

Marie Szafranski, Yves Grandvalet, and Alain Rakotomamonjy.*Composite kernel learning.* Machine Learning, 79(1-2):73–103, 2010.

Ashwin Srinivasan, Stephen Muggleton, Michael J. E. Sternberg, and Ross D. King.for mutagenicity: A study in first-order and feature-based induction. Artif. Intell., 85(1-2):277–299, 1996.

Ashwin Srinivasan and Ganesh Ramakrishnan.screening and optimisation for ilp using designed experiments. Journal of Machine Learning Research, 12:627–662, 2011.

Ashwin Srinivasan. *The aleph manual*. 1999.

Lucia Specia, Ashwin Srinivasan, Sachindra Joshi, Ganesh Ramakrishnan, and Maria das Graças Volpe Nunes.*An investigation into feature construction to assist word sense disambiguation.* Machine Learning, 76(1):109–136, 2009.

Bibliography

- Marie Szafranski, Yves Grandvalet, and Alain Rakotomamonjy.*Composite kernel learning.* Machine Learning, 79(1-2):73–103, 2010.
- Ashwin Srinivasan, Stephen Muggleton, Michael J. E. Sternberg, and Ross D. King.for mutagenicity: A study in first-order and feature-based induction. Artif. Intell., 85(1-2):277–299, 1996.
- Ashwin Srinivasan and Ganesh Ramakrishnan.screening and optimisation for ilp using designed experiments. Journal of Machine Learning Research, 12:627–662, 2011.
- Ashwin Srinivasan. *The aleph manual*. 1999.
- Lucia Specia, Ashwin Srinivasan, Sachindra Joshi, Ganesh Ramakrishnan, and Maria das Graças Volpe Nunes.*An investigation into feature construction to assist word sense disambiguation.* Machine Learning, 76(1):109–136, 2009.

Bibliography

Marie Szafranski, Yves Grandvalet, and Alain Rakotomamonjy.*Composite kernel learning.* Machine Learning, 79(1-2):73–103, 2010.

Ashwin Srinivasan, Stephen Muggleton, Michael J. E. Sternberg, and Ross D. King.for mutagenicity: A study in first-order and feature-based induction. Artif. Intell., 85(1-2):277–299, 1996.

Ashwin Srinivasan and Ganesh Ramakrishnan.screening and optimisation for ilp using designed experiments. Journal of Machine Learning Research, 12:627–662, 2011.

Ashwin Srinivasan. *The aleph manual*. 1999.

Lucia Specia, Ashwin Srinivasan, Sachindra Joshi, Ganesh Ramakrishnan, and Maria das Graças Volpe Nunes.*An investigation into feature construction to assist word sense disambiguation.* Machine Learning, 76(1):109–136, 2009.

Bibliography

Marie Szafranski, Yves Grandvalet, and Alain Rakotomamonjy.*Composite kernel learning.* Machine Learning, 79(1-2):73–103, 2010.

Ashwin Srinivasan, Stephen Muggleton, Michael J. E. Sternberg, and Ross D. King.for mutagenicity: A study in first-order and feature-based induction. Artif. Intell., 85(1-2):277–299, 1996.

Ashwin Srinivasan and Ganesh Ramakrishnan.screening and optimisation for ilp using designed experiments. Journal of Machine Learning Research, 12:627–662, 2011.

Ashwin Srinivasan. *The aleph manual*. 1999.

Lucia Specia, Ashwin Srinivasan, Sachindra Joshi, Ganesh Ramakrishnan, and Maria das Graças Volpe Nunes.*An investigation into feature construction to assist word sense disambiguation.* Machine Learning, 76(1):109–136, 2009.

Bibliography

Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun.*Support vector machine learning for interdependent and structured output spaces.* In Proceedings of the twenty-first international conference on Machine learning, ICML '04, pages 104–, New York, NY, USA, 2004. ACM.

Emmanuel Tapia, Stephen Intille, and Kent Larson. *Activity Recognition in the Home Using Simple and Ubiquitous Sensors Pervasive Computing.* In Alois Ferscha and Friedemann Mattern, editors, Pervasive Computing, volume 3001 of *Lecture Notes in Computer Science*, chapter 10, pages 158–175. Springer Berlin / Heidelberg.

Tim van Kasteren, Athanasios Noulas, Gwenn Englebienne, and Ben Kröse.*Accurate activity recognition in a home setting.* In Proceedings of the 10th international conference on Ubiquitous computing, UbiComp '08, pages 1–9, New York, NY, USA, 2008. ACM.

Ian H. Witten, Eibe Frank, and Mark A. Hall.*Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, Amsterdam, 3rd edition, 2011.

Bibliography

Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun.*Support vector machine learning for interdependent and structured output spaces.* In Proceedings of the twenty-first international conference on Machine learning, ICML '04, pages 104–, New York, NY, USA, 2004. ACM.

Emmanuel Tapia, Stephen Intille, and Kent Larson. *Activity Recognition in the Home Using Simple and Ubiquitous Sensors Pervasive Computing*. In Alois Ferscha and Friedemann Mattern, editors, Pervasive Computing, volume 3001 of *Lecture Notes in Computer Science*, chapter 10, pages 158–175. Springer Berlin / Heidelberg.

Tim van Kasteren, Athanasios Noulas, Gwenn Englebienne, and Ben Kröse.*Accurate activity recognition in a home setting.* In Proceedings of the 10th international conference on Ubiquitous computing, UbiComp '08, pages 1–9, New York, NY, USA, 2008. ACM.

Ian H. Witten, Eibe Frank, and Mark A. Hall.*Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, Amsterdam, 3rd edition, 2011.

Bibliography

Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. *Support vector machine learning for interdependent and structured output spaces.* In Proceedings of the twenty-first international conference on Machine learning, ICML '04, pages 104–, New York, NY, USA, 2004. ACM.

Emmanuel Tapia, Stephen Intille, and Kent Larson. *Activity Recognition in the Home Using Simple and Ubiquitous Sensors Pervasive Computing*. In Alois Ferscha and Friedemann Mattern, editors, Pervasive Computing, volume 3001 of *Lecture Notes in Computer Science*, chapter 10, pages 158–175. Springer Berlin / Heidelberg.

Tim van Kasteren, Athanasios Noulas, Gwenn Englebienne, and Ben Kröse. *Accurate activity recognition in a home setting.* In Proceedings of the 10th international conference on Ubiquitous computing, UbiComp '08, pages 1–9, New York, NY, USA, 2008. ACM.

Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, Amsterdam, 3rd edition, 2011.

Bibliography

Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun.*Support vector machine learning for interdependent and structured output spaces.* In Proceedings of the twenty-first international conference on Machine learning, ICML '04, pages 104–, New York, NY, USA, 2004. ACM.

Emmanuel Tapia, Stephen Intille, and Kent Larson. *Activity Recognition in the Home Using Simple and Ubiquitous Sensors Pervasive Computing*. In Alois Ferscha and Friedemann Mattern, editors, Pervasive Computing, volume 3001 of *Lecture Notes in Computer Science*, chapter 10, pages 158–175. Springer Berlin / Heidelberg.

Tim van Kasteren, Athanasios Noulas, Gwenn Englebienne, and Ben Kröse.*Accurate activity recognition in a home setting.* In Proceedings of the 10th international conference on Ubiquitous computing, UbiComp '08, pages 1–9, New York, NY, USA, 2008. ACM.

Ian H. Witten, Eibe Frank, and Mark A. Hall.*Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Amsterdam, 3rd edition, 2011.

Bibliography

- Wanhong Xu. *Supervising latent topic model for maximum-margin text classification and regression.* In Mohammed Zaki, Jeffrey Yu, B. Ravindran, and Vikram Pudi, editors, Advances in Knowledge Discovery and Data Mining, volume 6118 of Lecture Notes in Computer Science, pages 403–414. Springer Berlin / Heidelberg.
- Jieping Ye and Shuiwang Ji. *Discriminant analysis for dimensionality reduction: An overview of recent developments.*
- Jun Zhu, Amr Ahmed, and Eric Xing. *MedLDA: Maximum Margin Supervised Topic Models.* Journal of Machine Learning Research, 1:1–48, 2010.
- Amrita Saha, Ashwin Srinivasan, and Ganesh Ramakrishnan. *What kind of relational features are useful for Statistical learning?.* Inducive Logic Programming, Dubrovnik, Croatia, 2012.
- Naveen Nair, Ajay Nagesh, and Ganesh Ramakrishnan. *Markov Logic Chains.* Inducive Logic Programming, Dubrovnik, Croatia, 2012.
- Ajay Nagesh, Ganesh Ramakrishnan, Laura Citicariu, Rajasekar Krishnamurthy, Pushpak Bhattacharyya. *Efficient Rule Induction Toward Customizability.* Empirical Methods in Natural Language Processing, Jeju Island, Korea, 2012.
- Anup Kumar Chalamalla, Sumit Negi, L. Venkata Subramaniam, and Ganesh Ramakrishnan. *Identification of Class Specific Discourse Patterns.* ACM 17th Conference on Information and Knowledge Management (CIKM), Napa Valley, California, 2008.

Bibliography

- Wanhong Xu.*Supervising latent topic model for maximum-margin text classification and regression.* In Mohammed Zaki, Jeffrey Yu, B. Ravindran, and Vikram Pudi, editors, Advances in Knowledge Discovery and Data Mining, volume 6118 of Lecture Notes in Computer Science, pages 403–414. Springer Berlin / Heidelberg.
- Jieping Ye and Shuiwang Ji.*Discriminant analysis for dimensionality reduction: An overview of recent developments.*
- Jun Zhu, Amr Ahmed, and Eric Xing.*MedLDA: Maximum Margin Supervised Topic Models*. Journal of Machine Learning Research, 1:1–48, 2010.
- Amrita Saha, Ashwin Srinivasan, and Ganesh Ramakrishnan. *What kind of relational features are useful for Statistical learning?*. Inducive Logic Programming, Dubrovnik, Croatia, 2012.
- Naveen Nair, Ajay Nagesh, and Ganesh Ramakrishnan. *Markov Logic Chains*. Inducive Logic Programming, Dubrovnik, Croatia, 2012.
- Ajay Nagesh, Ganesh Ramakrishnan, Laura Citicariu, Rajasekar Krishnamurthy, Pushpak Bhattacharyya. *Efficient Rule Induction Toward Customizability*.Empirical Methods in Natural Language Processing, Jeju Island, Korea, 2012.
- Anup Kumar Chalamalla, Sumit Negi, L. Venkata Subramaniam, and Ganesh Ramakrishnan. *Identification of Class Specific Discourse Patterns*. ACM 17th Conference on Information and Knowledge Management (CIKM), Napa Valley, California, 2008.

Bibliography

- Wanhong Xu. *Supervising latent topic model for maximum-margin text classification and regression.* In Mohammed Zaki, Jeffrey Yu, B. Ravindran, and Vikram Pudi, editors, Advances in Knowledge Discovery and Data Mining, volume 6118 of Lecture Notes in Computer Science, pages 403–414. Springer Berlin / Heidelberg.
- Jieping Ye and Shuiwang Ji. *Discriminant analysis for dimensionality reduction: An overview of recent developments.*
- Jun Zhu, Amr Ahmed, and Eric Xing. *MedLDA: Maximum Margin Supervised Topic Models*. Journal of Machine Learning Research, 1:1–48, 2010.
- Amrita Saha, Ashwin Srinivasan, and Ganesh Ramakrishnan. *What kind of relational features are useful for Statistical learning?*. Inducive Logic Programming, Dubrovnik, Croatia, 2012.
- Naveen Nair, Ajay Nagesh, and Ganesh Ramakrishnan. *Markov Logic Chains*. Inducive Logic Programming, Dubrovnik, Croatia, 2012.
- Ajay Nagesh, Ganesh Ramakrishnan, Laura Citicariu, Rajasekar Krishnamurthy, Pushpak Bhattacharyya. *Efficient Rule Induction Toward Customizability.* Empirical Methods in Natural Language Processing, Jeju Island, Korea, 2012.
- Anup Kumar Chalamalla, Sumit Negi, L. Venkata Subramaniam, and Ganesh Ramakrishnan. *Identification of Class Specific Discourse Patterns.* ACM 17th Conference on Information and Knowledge Management (CIKM), Napa Valley, California, 2008.

Bibliography

- Wanhong Xu.*Supervising latent topic model for maximum-margin text classification and regression.* In Mohammed Zaki, Jeffrey Yu, B. Ravindran, and Vikram Pudi, editors, Advances in Knowledge Discovery and Data Mining, volume 6118 of Lecture Notes in Computer Science, pages 403–414. Springer Berlin / Heidelberg.
- Jieping Ye and Shuiwang Ji.*Discriminant analysis for dimensionality reduction: An overview of recent developments.*
- Jun Zhu, Amr Ahmed, and Eric Xing.*MedLDA: Maximum Margin Supervised Topic Models*. Journal of Machine Learning Research, 1:1–48, 2010.
- Amrita Saha, Ashwin Srinivasan, and Ganesh Ramakrishnan. *What kind of relational features are useful for Statistical learning?*. Inducive Logic Programming, Dubrovnik, Croatia, 2012.
- Naveen Nair, Ajay Nagesh, and Ganesh Ramakrishnan. *Markov Logic Chains*. Inducive Logic Programming, Dubrovnik, Croatia, 2012.
- Ajay Nagesh, Ganesh Ramakrishnan, Laura Citicariu, Rajasekar Krishnamurthy, Pushpak Bhattacharyya. *Efficient Rule Induction Toward Customizability.*Empirical Methods in Natural Language Processing, Jeju Island, Korea, 2012.
- Anup Kumar Chalamalla, Sumit Negi, L. Venkata Subramaniam, and Ganesh Ramakrishnan. *Identification of Class Specific Discourse Patterns.* ACM 17th Conference on Information and Knowledge Management (CIKM), Napa Valley, California, 2008.

Bibliography

- Wanhong Xu. *Supervising latent topic model for maximum-margin text classification and regression.* In Mohammed Zaki, Jeffrey Yu, B. Ravindran, and Vikram Pudi, editors, Advances in Knowledge Discovery and Data Mining, volume 6118 of Lecture Notes in Computer Science, pages 403–414. Springer Berlin / Heidelberg.
- Jieping Ye and Shuiwang Ji. *Discriminant analysis for dimensionality reduction: An overview of recent developments.*
- Jun Zhu, Amr Ahmed, and Eric Xing. *MedLDA: Maximum Margin Supervised Topic Models*. Journal of Machine Learning Research, 1:1–48, 2010.
- Amrita Saha, Ashwin Srinivasan, and Ganesh Ramakrishnan. *What kind of relational features are useful for Statistical learning?*. Inductive Logic Programming, Dubrovnik, Croatia, 2012.
- Naveen Nair, Ajay Nagesh, and Ganesh Ramakrishnan. *Markov Logic Chains*. Inductive Logic Programming, Dubrovnik, Croatia, 2012.
- Ajay Nagesh, Ganesh Ramakrishnan, Laura Citicariu, Rajasekar Krishnamurthy, Pushpak Bhattacharyya. *Efficient Rule Induction Toward Customizability.* Empirical Methods in Natural Language Processing, Jeju Island, Korea, 2012.
- Anup Kumar Chalamalla, Sumit Negi, L. Venkata Subramaniam, and Ganesh Ramakrishnan. *Identification of Class Specific Discourse Patterns.* ACM 17th Conference on Information and Knowledge Management (CIKM), Napa Valley, California, 2008.

Bibliography

- Wanhong Xu.*Supervising latent topic model for maximum-margin text classification and regression.* In Mohammed Zaki, Jeffrey Yu, B. Ravindran, and Vikram Pudi, editors, Advances in Knowledge Discovery and Data Mining, volume 6118 of Lecture Notes in Computer Science, pages 403–414. Springer Berlin / Heidelberg.
- Jieping Ye and Shuiwang Ji.*Discriminant analysis for dimensionality reduction: An overview of recent developments.*
- Jun Zhu, Amr Ahmed, and Eric Xing.*MedLDA: Maximum Margin Supervised Topic Models*. Journal of Machine Learning Research, 1:1–48, 2010.
- Amrita Saha, Ashwin Srinivasan, and Ganesh Ramakrishnan. *What kind of relational features are useful for Statistical learning?*. Inducive Logic Programming, Dubrovnik, Croatia, 2012.
- Naveen Nair, Ajay Nagesh, and Ganesh Ramakrishnan. *Markov Logic Chains*. Inducive Logic Programming, Dubrovnik, Croatia, 2012.
- Ajay Nagesh, Ganesh Ramakrishnan, Laura Citicariu, Rajasekar Krishnamurthy, Pushpak Bhattacharyya. *Efficient Rule Induction Toward Customizability*.Empirical Methods in Natural Language Processing, Jeju Island, Korea, 2012.
- Anup Kumar Chalamalla, Sumit Negi, L. Venkata Subramaniam, and Ganesh Ramakrishnan. *Identification of Class Specific Discourse Patterns*. ACM 17th Conference on Information and Knowledge Management (CIKM), Napa Valley, California, 2008.

Bibliography

- Wanhong Xu.*Supervising latent topic model for maximum-margin text classification and regression.* In Mohammed Zaki, Jeffrey Yu, B. Ravindran, and Vikram Pudi, editors, Advances in Knowledge Discovery and Data Mining, volume 6118 of Lecture Notes in Computer Science, pages 403–414. Springer Berlin / Heidelberg.
- Jieping Ye and Shuiwang Ji.*Discriminant analysis for dimensionality reduction: An overview of recent developments.*
- Jun Zhu, Amr Ahmed, and Eric Xing.*MedLDA: Maximum Margin Supervised Topic Models*. Journal of Machine Learning Research, 1:1–48, 2010.
- Amrita Saha, Ashwin Srinivasan, and Ganesh Ramakrishnan. *What kind of relational features are useful for Statistical learning?*. Inducive Logic Programming, Dubrovnik, Croatia, 2012.
- Naveen Nair, Ajay Nagesh, and Ganesh Ramakrishnan. *Markov Logic Chains*. Inducive Logic Programming, Dubrovnik, Croatia, 2012.
- Ajay Nagesh, Ganesh Ramakrishnan, Laura Citicariu, Rajasekar Krishnamurthy, Pushpak Bhattacharyya. *Efficient Rule Induction Toward Customizability*.Empirical Methods in Natural Language Processing, Jeju Island, Korea, 2012.
- Anup Kumar Chalamalla, Sumit Negi, L. Venkata Subramaniam, and Ganesh Ramakrishnan. *Identification of Class Specific Discourse Patterns*. ACM 17th Conference on Information and Knowledge Management (CIKM), Napa Valley, California, 2008.