Relational Kernels for Support Vector Machines on Structured Output Spaces

February 25, 2013

IITB-Monash
Research Academy

Examples

- Trees.
- Graphs.
- Lattices.
- Sequences.

Examples

- Trees.
- Graphs.
- Lattices.
- Sequences.

- We consider sequences.
- eg: 1) Natural language text, where words (or its derived characteristics) form a sequence.
  2) Activity Recognition, where activities performed by a person are in a sequential order.

Examples

- Trees.
- Graphs.
- Lattices.
- Sequences.

- We consider sequences.
- eg: 1) Natural language text, where words (or its derived characteristics) form a sequence.
  2) Activity Recognition, where activities performed by a person are in a sequential order.
- Problem: To label each element in a sequence of observations.

Examples

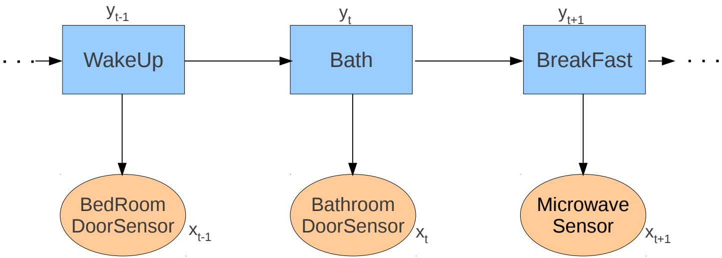
- Trees.
- Graphs.
- Lattices.
- Sequences.

- We consider sequences.
- eg: 1) Natural language text, where words (or its derived characteristics) form a sequence.
  2) Activity Recognition, where activities performed by a person are in a sequential order.
- Problem: To label each element in a sequence of observations.
  Observation: Labels at successive time steps are dependent. Ex: Cooking followed by dinner (Activity Recognition).

Examples

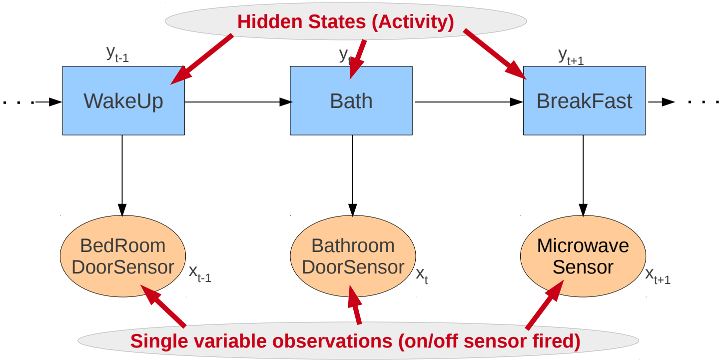
- Trees.
- Graphs.
- Lattices.
- Sequences.

- We consider sequences.
- eg: 1) Natural language text, where words (or its derived characteristics) form a sequence.
  2) Activity Recognition, where activities performed by a person are in a sequential order.
- Problem: To label each element in a sequence of observations.
  Observation: Labels at successive time steps are dependent. Ex: Cooking followed by dinner (Activity Recognition).
  Conventional approaches: Hidden Markov Models [Rabiner,1989], Conditional Random Fields [Lafferty *et al.*,2001].

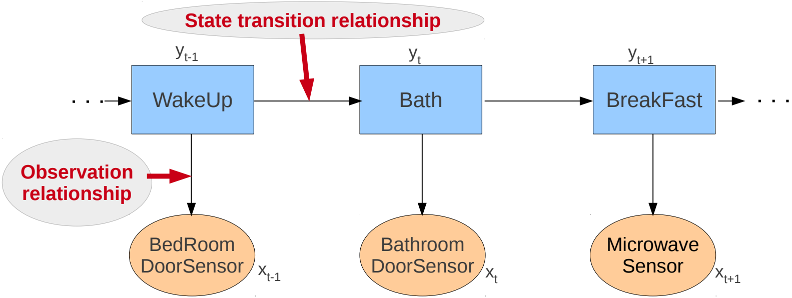


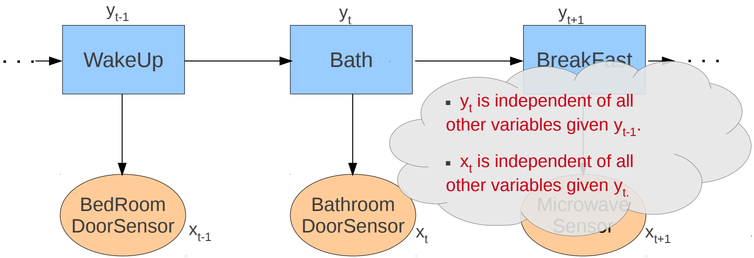
y t-1
y t
y t+1
WakeUp
Bath
BreakFast
BedRoom DoorSensor
x t-1
Bathroom DoorSensor
x t
Microwave Sensor
x t+1

Hidden Markov Model (HMM) [Rabiner,1989]



Hidden Markov Model (HMM) [Rabiner,1989]

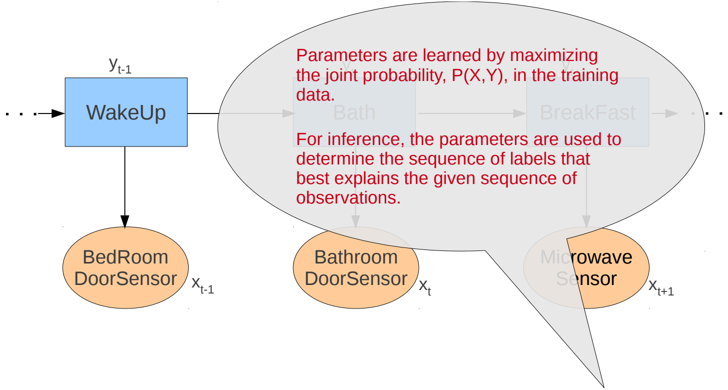




Initial state distribution: $P(y_1)$
Transition distribution: $P(y_t|y_{t-1})$
Emission distribution: $P(x_t|y_t)$.

Joint probability

$$P(X,Y) = \prod_{t=1}^{T} P(y_t|y_{t-1})P(x_t|y_t)$$

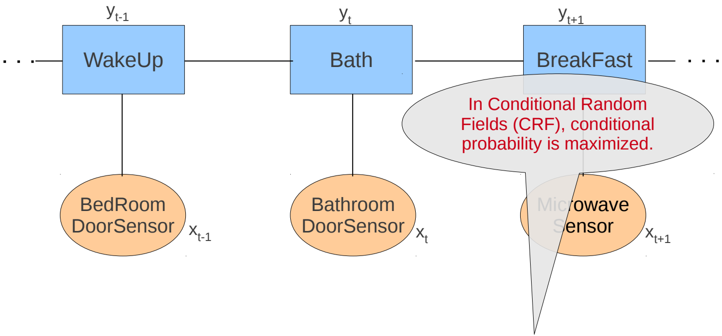


Initial state distribution: $P(y_1)$
Transition distribution: $P(y_t|y_{t-1})$
Emission distribution: $P(x_t|y_t)$.

Joint probability

$$P(X,Y) = \prod_{t=1}^{T} P(y_t|y_{t-1})P(x_t|y_t)$$

Conditional Random Field (CRF) [Lafferty *et al.*,2001]



$\phi_t(y_t, X)$, $\phi_{t-1}(y_{t-1}, y_t, X)$ are potential functions
$Z(X)$ is the partition function.

Conditional probability

$$p(Y|X) = \frac{1}{Z(X)}\, exp \sum_{t=1}^{T} \phi_t(y_t, X) + \phi_{t-1}(y_{t-1}, y_t, X)$$

- Models structured output classification in a large margin framework.
- Generalizes Support Vector Machines to capture relationships in output space.

- Models structured output classification in a large margin framework.
- Generalizes Support Vector Machines to capture relationships in output space.
- We focus on sequence labeling.

Let scoring function, $F(X,Y;\mathbf{f}) = \langle \mathbf{f}, \psi(X,Y) \rangle$

X: input sequence; Y: output sequence; $\psi(X,Y)$: feature vector (observation and transition); $\mathbf{f}$: parameter vector

Objective: Learn features that maximize $F(X,Y;\mathbf{f}) - \max_{\hat{Y} \neq Y} F(X,\hat{Y};\mathbf{f})$

Inference:

$$\hat{Y} = \mathscr{F}(X;\mathbf{f}) = \arg\max_{Y \in \mathscr{Y}} F(X,Y;\mathbf{f})$$

Loss function:

- $\Delta(Y,\hat{Y})$ for true output Y and prediction $\hat{Y}$.
- Predicted sequences that deviate more from the actual should be penalized more.

$$\min_{\mathbf{f},\xi} \frac{1}{2} \parallel \mathbf{f} \parallel^2 + \frac{C}{m}\sum_{i=1}^{m} \xi_i, \qquad s.t.\ \forall i:\ \xi_i \geq 0,$$

$$\forall i,\ \forall\, Y \neq Y_i: \quad \langle \mathbf{f}, \psi_i^\delta(Y)\rangle \ \geq\ 1 - \frac{\xi_i}{\Delta(Y_i, Y)}$$

- $\parallel . \parallel$: 2-norm regularizer.
- C: regularization parameter.
- m: no of example sequences in training set
- $\langle \mathbf{f}, \psi_i^\delta(Y)\rangle = \langle \mathbf{f}, \psi(X_i, Y_i)\rangle - \langle \mathbf{f}, \psi(X_i, Y)\rangle$
- X_i and Y_i: i^{th} input and output sequences in the training set.
- ξ: slack variables to allow errors in the training set in a soft margin SVM.

$$\min_{\mathbf{f},\xi} \frac{1}{2} \| \mathbf{f} \|^2 + \frac{C}{m} \sum_{i=1}^{m} \xi_i, \qquad s.t.\ \forall i:\ \xi_i \geq 0,$$

$$\forall i,\ \forall\, Y \neq Y_i: \quad \langle \mathbf{f}, \psi_i^{\delta}(Y) \rangle \ \geq\ 1 - \frac{\xi_i}{\Delta(Y_i, Y)}$$

Difference with regular SVMs ?

- Margin is defined as the difference in scores of true and wrong output sequences.
- The loss function also scales the slackness in margin. If the loss is large, less tolerance is allowed; and vice-versa

SVM for structured output spaces (StructSVM)[Tsochantaridis *et al.*, 2004,2006]

$$\min_{\mathbf{f},\xi} \frac{1}{2} \| \mathbf{f} \|^2 + \frac{C}{m} \sum_{i=1}^{m} \xi_i, \qquad s.t.\ \forall i:\ \xi_i \geq 0,$$

$$\forall i,\ \forall\, Y \neq Y_i :\quad \langle \mathbf{f}, \psi_i^{\delta}(Y) \rangle \ \geq\ 1 - \frac{\xi_i}{\Delta(Y_i, Y)}$$

- The number of constraints can be extremely large.
- Cutting plane method for finding polynomially sized subset of constraints [Tsochantaridis *et al.*, 2004,2006].
  - Start with no constraints.
  - Incrementally add constraints that violates the margin more than a threshold ε.
  - Repeat until no constraint violates the margin more than ε.

Input: kernels, C, ε_{margin}

1. $S_i \leftarrow \phi \quad \forall i = 1, \ldots, m$
2. **repeat**
3. **for** $i = 1, \ldots, m$ **do** //for each example
4. Define $H(Y) \equiv \left[1 - \langle \mathbf{f}, \psi_i^{\delta}(Y)\rangle\right] \Delta(Y_i, Y)$ //Margin Violation
5. Compute $\hat{Y} = \arg\max_{Y \in \mathscr{Y}} H(Y)$. //Max Margin Violation
6. Compute $\xi_i = \max\{0, \max_{Y \in S_i} H(Y)\}$. //Current Max Margin Violation
7. **if** $H(\hat{Y}) > \xi_i + \varepsilon_{margin}$, **then**
8. $S_i \leftarrow S_i \bigcup \{\hat{Y}\}$. //adding constraints
9. $\alpha \leftarrow$ optimize dual over S, $S = \bigcup_i S_i$. //$\mathbf{f}$ can be derived from α
10. **end if**
11. **end for**

12. **until** no S_i has changed during the iteration.

$$\max_{\alpha} \sum_{i,Y \neq Y_i} \alpha_{iY} - \frac{1}{2} \sum_{i,Y \neq Y_i} \sum_{j,Y' \neq Y_j} \alpha_{iY} \alpha_{j,Y'} \langle \psi_i^{\delta}(Y), \psi_j^{\delta}(Y^{'}) \rangle$$

such that,

$$\forall i, \forall Y \neq Y_i : \quad \alpha_{iY} \geq 0$$
$$\forall i : \quad n \sum_{Y \neq Y_i} \frac{\alpha_{iY}}{\Delta(Y_i, Y)} \leq C$$

$$\max_{\alpha} \sum_{i,Y \neq Y_i} \alpha_{iY} - \frac{1}{2} \sum_{i,Y \neq Y_i} \sum_{j,Y' \neq Y_j} \alpha_{iY} \alpha_{j,Y'} \kappa^{\delta} \Big((X_i, Y_i, Y), (X_j, Y_j, Y') \Big)$$

such that,

$$\forall i, \forall Y \neq Y_i : \quad \alpha_{iY} \geq 0$$
$$\forall i : \quad n \sum_{Y \neq Y_i} \frac{\alpha_{iY}}{\Delta(Y_i, Y)} \leq C$$

Kernel can be split into emission and transition parts.

$$\kappa^{\delta}\Big((X_i, Y_i, Y), (X_j, Y_j, Y^{'})\Big) = \kappa_T^{\delta}(Y_i, Y, Y_j, Y^{'}) + \kappa_E^{\delta}\Big((X_i, Y_i, Y), (X_j, Y_j, Y^{'})\Big)$$

From the definition of $\psi_i^{\delta}(Y)$,

$$\kappa_T^{\delta}\left(Y_i, Y, Y_j, Y^{'}\right) = \kappa_T(Y_i, Y_j) + \kappa_T(Y, Y^{'}) - \kappa_T(Y_i, Y^{'}) - \kappa_T(Y_j, Y)$$

where,

$$\begin{aligned}\kappa_T(Y_i, Y_j) &= \sum_{p=1}^{l_i - 1} \sum_{q=1}^{l_j - 1} \Lambda(y_i^p, y_j^q)\Lambda(y_i^{p+1}, y_j^{q+1}) \\ &= \sum_{p=2}^{l_i} \sum_{q=2}^{l_j} \Lambda(y_i^{p-1}, y_j^{q-1})\Lambda(y_i^p, y_j^q),\end{aligned}$$

where $\Lambda(y_i^p, y_j^q) = 1$ if $y_i^p = y_j^q$; 0 otherwise. y_i^p is the p^{th} label of i^{th} sequence.

Kernel

Similarly,

$$\kappa_E^{\delta}\left((X_i, Y_i, Y), (X_j, Y_j, Y^{'})\right) = \sum_{p=1}^{l_i} \sum_{q=1}^{l_j} \kappa_E(x_i^p, x_j^q) \Big(\Lambda(y_i^p, y_j^q) + \Lambda(y^p, y^{'q}) - \Lambda(y_i^p, y^{'q}) - \Lambda(y^p, y_j^q) \Big)$$

The kernel $\kappa_E(x_i^p, x_j^q)$ can be defined as a Set-Sequence (String) kernel (or any fancy kernel), where we may be considering some window time steps before and after p and q, with p and q as pivots.

Conclusion

- Structured Output Spaces
- Sequence labeling problems.
- Hidden Markov Models.
- Conditional Random Fields.
- StructSVM.
- Cutting Plane Algorithm.
- Dual and Kernel.