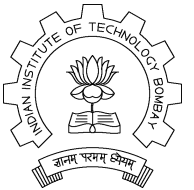
Graphical Models for Data Mining

NLP-AI Seminar



Manoj Kumar Chinnakotla

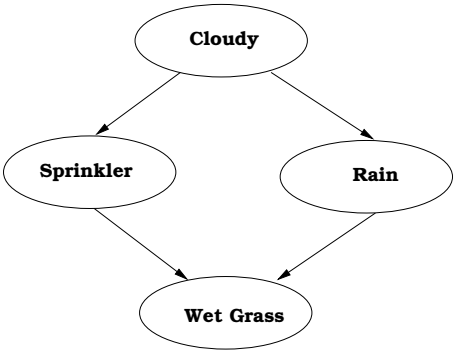
Outline of the Talk

- Graphical Models - Overview
- Motivation
- Bayesian Networks
- Markov Random Fields
- Inferencing and Learning
- Expressive Power
- Example Applications
  - Gene Expression Analysis
  - Web Page Classification
- Summary

Graphical Models - An Introduction

- Graph $G =< V, E >$ representing a family of probability distributions
- Nodes V - Random Variables
- Edges E - Indicate Stochastic Dependence
- G encodes *Conditional Independence* assertions in domain
- Mainly two kinds of Models
  - Directed (a.k.a *Bayesian Networks*)
  - Undirected (a.k.a *Markov Random Fields (MRFs)*)

Graphical Models (Contd. . .)



a

- Direction of edges based on causal knowledge
  - $A \rightarrow B$: A "causes" B
  - $A - B$: Not sure of causality
- Mixed versions also possible - *Chain Graphs*

[a] Figure adapted from [RN95]

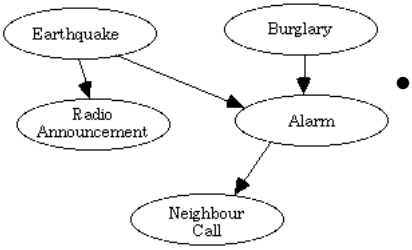
Why Graphical Models?

- Framework for modeling and effeciently reasoning about multiple correlated random variables
- Provides insights into the assumptions of existing models
- Allows qualitative specification of independence assumptions

Why Graphical Models?
Recent Trends in Data Mining

- Traditional learning algorithms assume
  - Data available in record format
  - Instances are *i.i.d* samples
- Recent domains like Web, Biology, Marketing have more *richly* structured data
- Examples : DNA Sequences, Social Networks, Hyperlink structure of Web, Phylogeny Trees
- Relational Data Mining - Data spread across multiple tables
- Relational Structure helps significantly in enhancing accuracy [CDI98, LG03]
- Graphical Models offer a natural formalism to model such data

Directed Models : Bayesian Networks



a

[a]Figure adapted from [RN95]

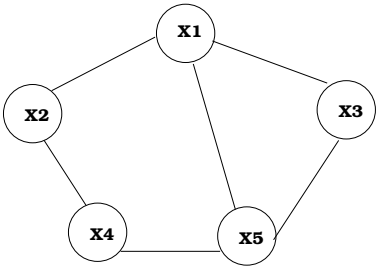
- *Bayes Net* - DAG encoding the conditional independence assumptions among the variables
- Cycles not allowed - Edges usually have causal interpretations
- Specifies a compact representation of joint distribution over the variables given by

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P_i(X_i \mid Pa(X_i))$$

where $Pa(X_i) =$ Parents of Node X_i in the network

- $P_i \rightarrow$ *Conditional Probability Distribution (CPD)* of X_i

Undirected Graphical Models
Markov Random Fields



- Have been well studied and applied in Vision
- No underlying causal structure
- Joint distribution can be factorized into

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(X_c)$$

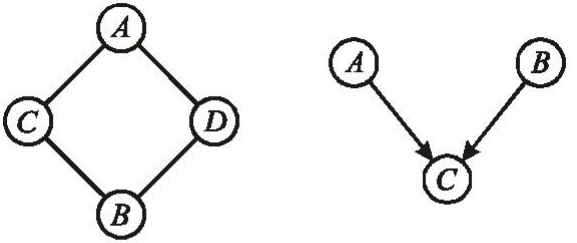
where C - Set of cliques in graph

- ψ_c - Potential function (a positive function) on the clique X_c
- Z - Partition Function given by

$$Z = \sum_{\vec{x}} \prod_{c \in C} \psi_c(X_c)$$

Expressive Power
Directed vs Undirected Models

- Dependencies which can be modeled - Not exactly similar
- Example :

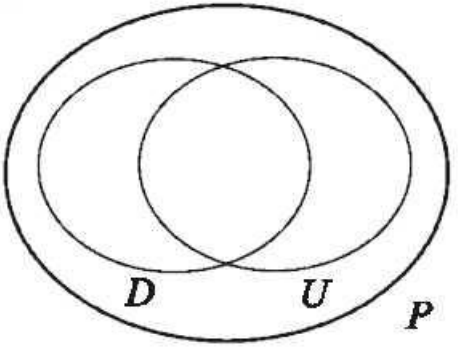


a

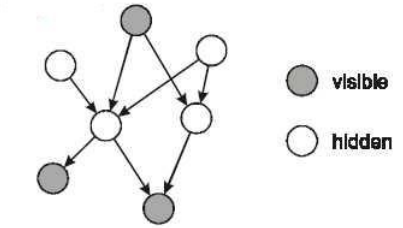
- Decomposable Models - Class of dependencies which both can model

[a] Figure adapted from [JP98]

What Class of Distributions Can be Modeled?



Inference



- Given a subset of variables X_K, compute distribution of $P(X_U|X_K)$ where $\vec{X} = \{X_U\} \cup \{X_K\}$
- Marginals - involve summation over exponential terms
- Complexity handled by exploiting the graphical structure
- Algorithms : *Exact* and *Approximate*
- Some Examples : *Variable Elimination, Sum-Product Algorithm, Sampling Algorithm*

Learning

- Estimating graphical structure G and parameters from data
- Standard ML estimates used when variables in the model are fully *Observable*
- MRFs use Iterative Algorithms for parameter estimation
- Structure Learning relatively hard

Applications

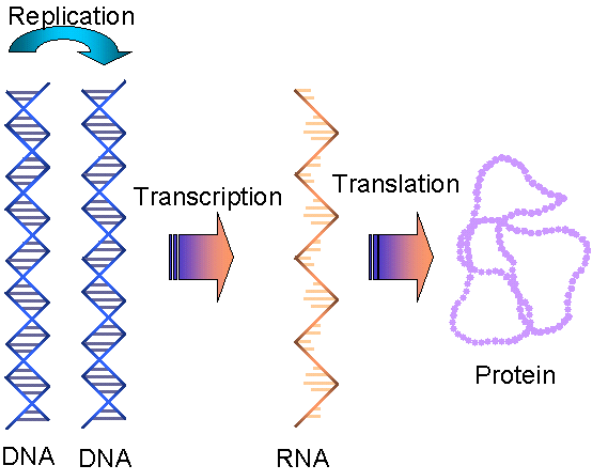
Bio-informatics
Gene Expression Analysis

- Gene Expression Analysis - Introduction
- Standard Techniques - Clustering and Bayesian Networks
- Probabilistic Relational Models (PRMs)
- Integrating Additional Information into PRM
- Learning PRMs from Data

DNA - The Blueprint of Life!

- DNA - *Deoxyribo Nucleic Acid*
- Double Helix Structure
- Each Strand - Sequence of *Nucleotides* $\{$*Adenine (A),Guanine (G),Cytosine (C), Thymine (T)*$\}$
- Complementary Strands - A $\leftrightarrow$ G, C $\leftrightarrow$ T
- *Gene* - Portions of DNA that code for Proteins or large biomolecules

The Central Dogma - Transcription and Translation



[a]Figure Source : www.swbic.org/education/comp-bio/images/

Gene Expression

- Each cell has same copy of DNA still different cells synthesize different Proteins!
  - Example : Cells making the proteins needed for muscles, eye lens etc.
- Gene said to be *expressed* if it produces it's corresponding protein
- Genes expressed vary - Based on time, location, environmental and biological conditions
- Expression regulated by a complex collection of proteins

DNA Micro-array Technology

- *Micro-array or Gene chips* used for experiments
- Allows measurement of *expression levels* of tens of thousands of genes simultaneously
- Many experiments measure *expression* of same set of genes under various environmental/biological conditions
  - Example : Cell is heated up, cooled down, drug added
- Expression Level
  - Estimated based on amount of mRNA for that gene currently present in that cell
  - Ratio of expression level under experiment condition to expression under normal condition taken instead

Gene Expression Data

Features →

↓Examples	Experiment 1	Experiment 2	…	Experiment N
Gene 1	1083	1464	…	1115
Gene 2	1585	398	…	511
…	…	…	…	…
Gene M	170	302	…	751

a

- Enormous amount of expression data for various species publicly available
- Some Examples
  - EBI Micro-array data repository (`http://www.ebi.ac.uk/arrayexpress/`)
  - Stanford Micro-array Database (`http://genome-www5.stanford.edu/`) etc.

[a] Figure Source : [**?**]

The Problem - Drowning in Data! Where is Information?

- Enormous amount of data
  - EBI data repository has grown 100-fold just in a year!
- Difficult for humans to comprehend, detect patterns
- Biological experiments - Costly and Time consuming
- Machine Learning/Data Mining techniques to the rescue
  - Allow learning of models which provide useful insight into the biological processes
  - Reduce the number of biological experiments needed

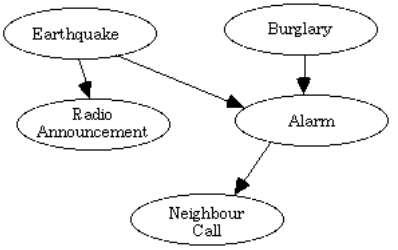
Gene Expression Analysis - Approaches

- Aim
  - To identify co-regulated genes
  - To gain biological insight into gene regulatory mechanisms
- Approaches
  - Clustering
  - Bayesian Networks
  - Probabilistic Relational Models (PRMs)
- Focus of the Presentation
  - Probabilistic Models for Gene Expression using PRMs

Clustering

- Two-Side Clustering
  - Genes and Experiments partitioned into clusters $G_1, \ldots, G_k$ and $E_1, \ldots, E_l$ simultaneously
  - Summarizes data into groups of $k \times l$
  - Assumption - Expression governed by a distribution specific to each combination of Gene/Experiment clusters
- Clustering Techniques - Problems
  - Similarity based on all the measurements. What if similarity exists only over a subset of measurements?
  - Difficult to integrate additional information - Gene Annotation, Cell-Type/Strain used, Gene Promoters

Bayesian Networks



- *Bayes Net* - DAG encoding the conditional independence assumptions among the variables
- Specifies a compact representation of joint distribution over the variables given by

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i))$$

where $Pa(X_i) =$ Parents of Node X_i in the network

- Provides insight into the influence patterns across variables
- Friedman et al have applied it to learn gene regulatory mechanisms

Bayesian Networks (Contd. . .)
Modeling Relational Data

- *Relational Data* - Data spread across multiple tables
- Provides valuable additional information for learning models
  - Example : DNA Sequence Information, Gene Annotations
- Bayes Nets not suitable for modeling
  - Bayes Net Learning Algorithms - Attribute Based
  - Assume all the data to be present in a single table
  - Make sample independence assumption
- Solution : Why not “flatten” the data?
  - Will make the samples dependent
  - Can’t be used to reach conclusions based on relational dependencies

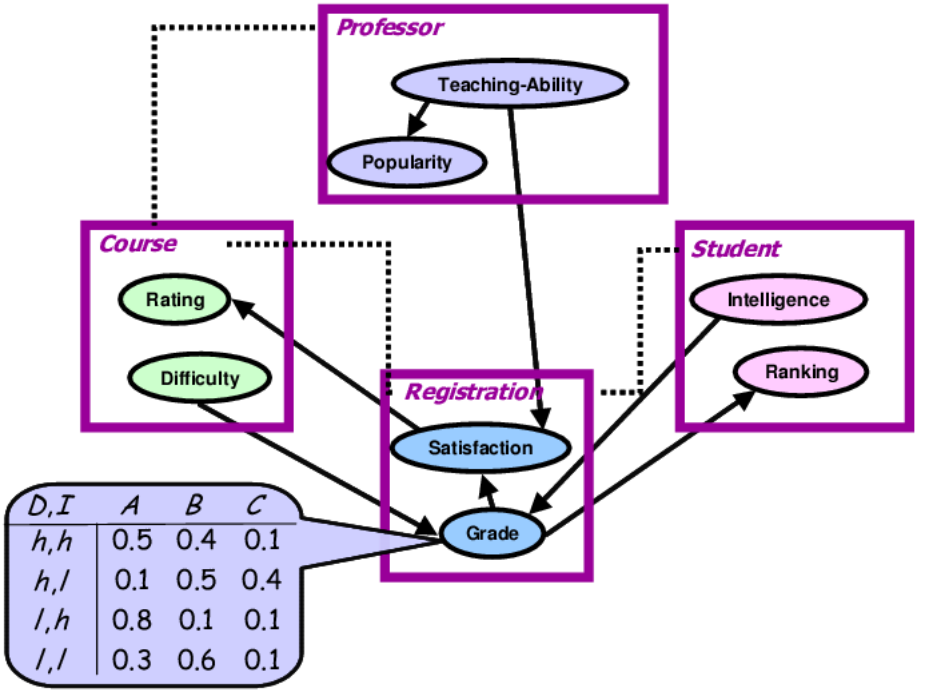
Probabilistic Relational Models (PRMs)

- Learns a probabilistic model over a *relational schema* involving multiple entities
- Entities in the current problem *Gene*, *Array* and *Expression*
- Each entity X can have attributes of the form
  - X.B - Simple Attribute
  - X.R.C - Attribute of another relation where R is a *Reference Slot*
- *Reference Slots* - Similar to foreign keys in the database world

PRMs (Contd. . .)

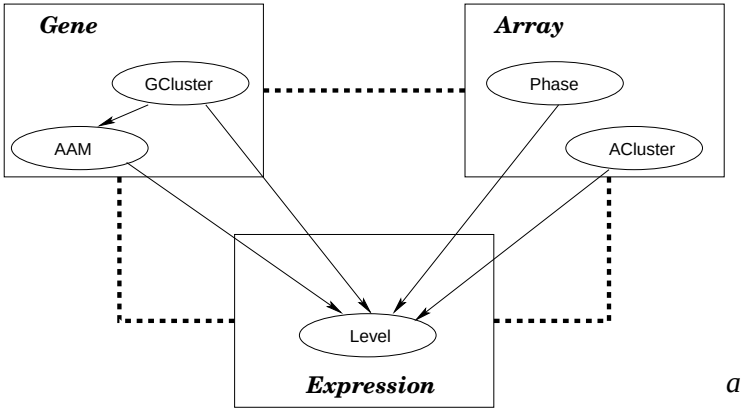
- Attributes of objects - Random Variables
- Given the above, a PRM Π is defined by
  - A class-level dependency structure S
  - The parameter set θ_S for the resultant *Conditional Probability Distribution (CPD)*
- The PRM Π is only a class-level "template" - Gets instantiated for each object

A Sample PRM



[a]Figure Source : [FGKP99]

PRM for Gene Expression



[a]

[a]Figure Source : [STG+01]

Inferencing in PRMs

- A *Relational Skeleton* σ is an instantiation of this schema
- For Example : 1000 gene objects, 100 array objects and 100,000 objects expression objects
- Relational skeleton σ completely specifies the values for the reference slots
- *Objective*
  Given σ, with observed evidence regarding some variables, update the probabilistic distribution over the rest of the variables

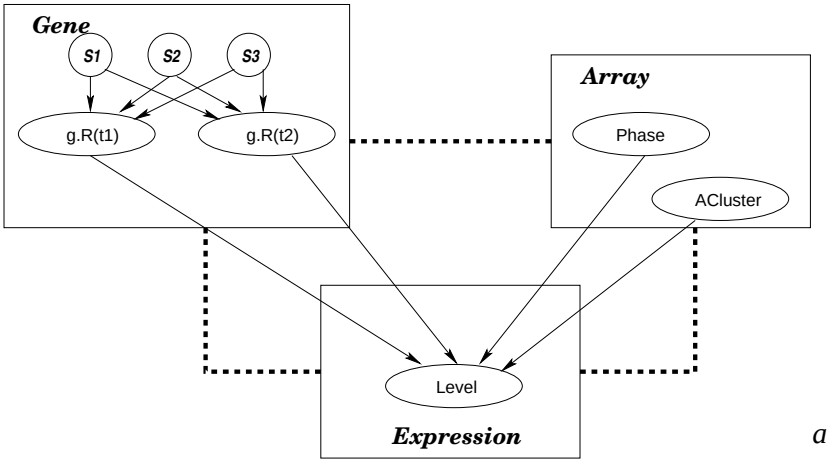
Inferencing in PRMs (Contd. . .)

- Given a relational skeleton σ, a PRM induces a *Bayesian Network* over all the random variables
- Parents and *CPDs* of Bayes Net - Obtained from class-level PRM
- Bayesian Network Inferencing Algorithms are then used for *inference* in the resultant network

Integrating Additional Sources of Data
DNA Sequence Information

- *Transcription Factors (TFs)* - Proteins that bind to specific DNA sequence in the promoter region known as *binding sites*
- TFs encourage or repress the start of transcription
- Why is sequence information important?
  - Help in identifying TF *binding sites*
  - Two genes with similar expression profiles - mostly likely to be controlled by same TFs
- New features added
  - Base pairs of Promoter Sequence
  - *Regulates* variable $g.R(t)$ for each TF t

PRM with Promoter Sequence Information



a

[a]Figure Source : [SBS+02]

Learning the Models

- CPD Parameter Estimation
  - Expression.Level modeled using a Gaussian
  - CPD divides the expression values into $k \times l$ groups
  - Parameter set constitutes the mean and variance of each group
- CPD Structure Learning
  - Scoring Function - measure of "goodness" of a structure relative to data
  - Search Algorithm - finding the structure with highest score
  - Bayesian Score as scoring function- Posterior of structure given data $P(S \mid D)$
  - Greedy local structure search used for search algorithm

PRMs for Gene Expression : Conclusion

- Templates for directed graphical models over relational data
- PRMs can be applied to relational data spread across multiple tables
- Capable of learning *unified models* integrating sequence information, expression data and annotation data
- Can easily accommodate additional information related to domain

Web Mining
Collective Web Page Classification [CDI98]

- Class of neighbouring pages (in Web Graph) usually correlated.
- Construct a directed graphical model based on the web graph.
  - Nodes - Random Variables for the category of each page
- Given an assignment of categories for some nodes :
  - Run *inferencing* on the above graphical model
  - Find the *Most Probable Explanation* for the rest

Summary

- Graphical Models - A natural formalism for modeling multiple correlated random variables
- Allows integration of domain knowledge in the form of dependency structures
- Techniques especially useful when data spread across multiple tables
- Allows easy integration of new additional information

Thanks!

References

[NLD99] Nir Friedman, Lise Getoor, Daphne Koller and Avi Pfeffer, Learning Probabilistic Relational Models, In Proceedings of IJCAI 1999, pages 1300-1309, 1999.

[CDI98] Soumen Chakrabarti, Byron E. Dom and Piotr Indyk , Enhanced hypertext categorization using hyperlinks , In Proceedings of SIGMOD-98, ACM International Conference on Management of Data , pages 307–318, 1998.

[Chi02] David Maxwell Chickering, The WinMine Toolkit, Microsoft, MSR-TR-2002-103, 2002, Redmond, WA.

[Col02] Michael Collins, Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms, In the proceedings of EMNLP 2002, pages 1–8, 2002.

[Fri00] Friedman N., Linial, Nachman I. and Pe'er D., Using Bayesian Networks to Analyze Expression Data, Journal of Computational Biology, vol 7, pages 601-620, 2000.

[GS04] Shantanu Godbole and Sunita Sarawagi, Discriminative Methods for Multi-Labeled Classification, In Proceedings of PAKDD 2004, 2004.

[LG03] Qing Lu and Lise Getoor, Link-based Classification, In Proceedings of ICML 2003, page 496, August 2003.

[Mur01] Kevin P. Murphy, The Bayes Net Toolbox for MATLAB, Journal of Computing Science and Statistics, vol. 33, 2001.

[FGKP99] Nir Friedman, Lise Getoor, Daphne Koller and Avi Pfeffer , Learning Probabilistic Relational Models , IJCAI , 1300-1309 , 1999

[STG+01] E. Segal, B. Taskar, A. Gasch, N. Friedman and D. Koller , Rich probabilistic models for gene expression , Bioinformatics , 17 , s243-52 , 2001

[SBS+02] E. Segal, Y. Barash, I. Simon, N. Friechnan and D. Koller , From promoter sequence to expression: A probabilistic framework , RECOMB , 2002

[RN95] S. Russel and P. Norvig, Artificial Intelligence: A Modern Approach, Prentice-Hall, 1995.

[MWJ99] Kevin P. Murphy, Yair Weiss and Michael I. Jordan, Loopy belief propagation for approximate inference : An emperical Study. In Proceedings of UAI 99, Pages 467-475, 1999.

[JP98] Pearl, J., Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann Publishers, 1988.

27 6, 35 6 27 28 32 4, 7 9