Machine Learning

10-701/15-781, Spring 2010

Bayesian Networks

Eric Xing

Lecture 13, March 1, 2010

Reading: Chap. 8, C.B book

© Eric Xing @ CMU, 2006-2010
What is a Bayesian Network?  
--- example from a signal transduction pathway

- A possible world for cellular signal transduction:

```
Receptor A  \(X_1\)  Receptor B  \(X_2\)
Kinase C  \(X_3\)  Kinase D  \(X_4\)  Kinase E  \(X_5\)
TF  \(X_6\)
Gene G  \(X_7\)  Gene H  \(X_8\)
```

Recap of Basic Prob. Concepts

- Representation: what is the joint probability dist. on multiple variables?

\[
P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)
\]

- How many state configurations in total? --- \(2^8\)
- Are they all needed to be represented?
- Do we get any scientific/medical insight?

- Learning: where do we get all this probabilities?

  - Maximal-likelihood estimation? but how many data do we need?
  - Where do we put domain knowledge in terms of plausible relationships between variables, and plausible values of the probabilities?

- Inference: If not all variables are observable, how to compute the conditional distribution of latent variables given evidence?

  - Computing \(p(H|A)\) would require summing over all \(2^8\) configurations of the unobserved variables
What is a Bayesian Network?
--- example from a signal transduction pathway

- A possible world for cellular signal transduction:

  - Receptor A $X_1$
  - Receptor B $X_2$
  - Kinase C $X_3$
  - Kinase D $X_4$
  - Kinase E $X_5$
  - TF F $X_6$
  - Gene G $X_7$
  - Gene H $X_8$

BN: Structure Simplifies Representation

- Dependencies among variables
If $X_i$'s are conditionally independent (as described by a BN), the joint can be factored to a product of simpler terms, e.g.,

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1) P(X_2) P(X_3 | X_1) P(X_4 | X_2) P(X_5 | X_2) P(X_6 | X_3, X_4) P(X_7 | X_6) P(X_8 | X_5, X_6)$$

Why we may favor a BN?
- Representation cost: how many probability statements are needed?
  - $2^8 = 256$ to $4^8 = 36$, an 8-fold reduction from $2^8$!
- Algorithms for systematic and efficient inference/learning computation
  - Exploring the graph structure and probabilistic semantics
  - Incorporation of domain knowledge and causal (logical) structures

Specification of a BN
- There are two components to any GM:
  - the qualitative specification
  - the quantitative specification

© Eric Xing @ CMU, 2006-2010
**Bayesian Network: Factorization Theorem**

*Theorem:*

Given a DAG, the most general form of the probability distribution that is consistent with the (probabilistic independence properties encoded in the) graph factors according to “node given its parents”:

\[
P(X) = \prod_{i} P(X_{i} | X_{\pi(i)})
\]

where \(X_{\pi(i)}\) is the set of parents of \(x_{i}\). \(d\) is the number of nodes (variables) in the graph.

**Examples**

\[
\begin{aligned}
P(B, E, J, M) & = P(B) P(E | B) P(A | E, B) P(J | A, B) P(M | A, B) \\
P(L | B)
\end{aligned}
\]
Qualitative Specification

- Where does the qualitative specification come from?
  - Prior knowledge of causal relationships
  - Prior knowledge of modular relationships
  - Assessment from experts
  - Learning from data
  - We simply link a certain architecture (e.g. a layered graph)
  - …

Local Structures & Independencies

- Common parent
  - Fixing B decouples A and C
    "given the level of gene B, the levels of A and C are independent"

- Cascade
  - Knowing B decouples A and C
    "given the level of gene B, the level gene A provides no extra prediction value for the level of gene C"

- V-structure
  - Knowing C couples A and B because A can "explain away" B w.r.t. C
    "If A correlates to C, then chance for B to also correlate to B will decrease"

- The language is compact, the concepts are rich!
A simple justification

Graph separation criterion

- D-separation criterion for Bayesian networks (D for Directed edges):

  **Definition**: variables $x$ and $y$ are *D-separated* (conditionally independent) given $z$ if they are separated in the *moralized* ancestral graph

- Example:

  - Original graph
  - Ancestral graph
  - Moral ancestral graph
Global Markov properties of DAGs

- X is d-separated (directed-separated) from Z given Y if we can't send a ball from any node in X to any node in Z using the "Bayes-ball" algorithm illustrated below (and plus some boundary conditions):

  - Defn: $\mathcal{I}(G) =$ all independence properties that correspond to d-separation:

    $$\mathcal{I}(G) = \{X \perp Z | Y : dsep \subseteq I(X; Y)\}$$

  - D-separation is sound and complete

Example:

- Complete the $\mathcal{I}(G)$ of this graph:

  - $X_4 \perp X_5 \perp X_6 | X_1$
  - $X_4 \perp X_5 \perp X_6 | X_1$

Essentially: A BN is a database of Pr. Independence statements among variables.
Bayesian Network: Conditional Independence Semantics

Structure: DAG

- Meaning: a node is conditionally independent of every other node in the network outside its Markov blanket

- Local conditional distributions (CPD) and the DAG completely determine the joint dist.

- Give causality relationships, and facilitate a generative process

Towards quantitative specification of probability distribution

- Separation properties in the graph imply independence properties about the associated variables

- For the graph to be useful, any conditional independence properties we can derive from the graph should hold for the probability distribution that the graph represents

- The Equivalence Theorem
  For a graph G,
  Let \( \mathcal{D}_1 \) denote the family of all distributions that satisfy \( I(G) \),
  Let \( \mathcal{D}_2 \) denote the family of all distributions that factor according to G,
  Then \( \mathcal{D}_1 \equiv \mathcal{D}_2 \).
Quantitative Specification

\[ p(A,B,C) = P(A|B,C) \cdot P(B|C) \cdot P(C) \]

Conditional probability tables (CPTs)

\[

table
\]

\[
P(a,b,c,d) = P(a)P(b)P(c|a,b)P(d|c)
\]

\[
\begin{array}{c|c|c|c|c|}
   & a^0 & a^1 & b^0 & b^1 \\
\hline
   c^0 & 0.45 & 1 & 0.9 & 0.7 \\
   c^1 & 0.55 & 0 & 0.1 & 0.3 \\
\end{array}
\]

\[
\begin{array}{c|c|c}
   & c^0 & c^1 \\
\hline
d^0 & 0.3 & 0.5 \\
d^1 & 0.7 & 0.5 \\
\end{array}
\]
Conditional probability density func. (CPDs)

\[ P(a, b, c, d) = P(a)P(b)P(c|a,b)P(d|c) \]

\[ A \sim N(\mu_a, \Sigma_a) \]
\[ B \sim N(\mu_b, \Sigma_b) \]
\[ C \sim N(A+B, \Sigma_c) \]
\[ D \sim N(\mu_a + C, \Sigma_a) \]

Conditional Independencies

What is this model

1. When Y is observed?
2. When Y is unobserved?
Conditionally Independent Observations

Data = \{y_1, \ldots, y_n\}

“Plate” Notation

Plate = rectangle in graphical model

variables within a plate are replicated in a conditionally independent manner
Example: Gaussian Model

Generative model:

\[ p(x_1, \ldots, x_n | \mu, \sigma) = \prod_{i=1}^{n} p(x_i | \mu, \sigma) \]

\[ = p(\text{data} | \text{parameters}) \]

\[ = p(D | \theta) \]

where \( \theta = \{\mu, \sigma\} \)

- Likelihood \( = p(\text{data} | \text{parameters}) \)
  \[ = p(D | \theta) \]
  \[ = L(\theta) \]

- Likelihood tells us how likely the observed data are conditioned on a particular setting of the parameters
  - Often easier to work with \( \log L(\theta) \)

Bayesian models

\[ p(b | \alpha) \]

\[ \theta \]

\[ x_i \]

\[ i=1:n \]
Example: modeling text

A Hierarchical Phrase-Based Model for Statistical Machine Translation

We present a statistical phrase-based Translation model that uses hierarchical phrases—phrases that contain sub-phrases. The model is formally a synchronous context-free grammar but is learned from a bitext without any syntactic information. Thus it can be seen as a shift to the formal machinery of syntax based-translation systems without any linguistic commitment. In our experiments using BLEU as a metric, the Hierarchical Phrase-based model achieves a relative improvement of 7.5% over Pharaoh, a state-of-the-art phrase-based system.

More examples

Density estimation

- Parametric and nonparametric methods

Regression

- Linear, conditional mixture, nonparametric

Classification

- Generative and discriminative approach
Example, con'd

- Evolution

![Evolution Diagram](image1)

- Speech recognition

![Speech Recognition Diagram](image2)
Example, con'd

- Genetic Pedigree

An (incomplete) genealogy of BNs

(Picture by Zoubin Ghahramani and Sam Roweis)
BN and Graphical Models

- A Bayesian network is a special case of **Graphical Models**

- A Graphical Model refers to a family of distributions on a set of random variables that are **compatible** with all the probabilistic independence propositions encoded by a graph that connects these variables.

- It is a smart way to write/specify/compose/design exponentially-large probability distributions without paying an exponential cost, and at the same time endow the distributions with structured semantics.

Two types of GMs

- **Directed edges** give causality relationships (Bayesian Network or Directed Graphical Model):

  \[
P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1) P(X_2) P(X_3 | X_1) P(X_4 | X_2) P(X_5 | X_2) P(X_6 | X_3, X_4) P(X_7 | X_6) P(X_8 | X_5, X_6)
  \]

- **Undirected edges** simply give correlations between variables (Markov Random Field or Undirected Graphical model):

  \[
P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = \frac{1}{Z} \exp\{E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) + E(X_6) + E(X_7) + E(X_8)\}
  \]
Probabilistic Inference

- Computing statistical queries regarding the network, e.g.:
  - Is node X independent on node Y given nodes Z,W?
  - What is the probability of X=true if (Y=false and Z=true)?
  - What is the joint distribution of (X,Y) if Z=false?
  - What is the likelihood of some full assignment?
  - What is the most likely assignment of values to all or a subset the nodes of the network?

- General purpose algorithms exist to fully automate such computation
  - Computational cost depends on the topology of the network
  - Exact inference:
    - The junction tree algorithm
  - Approximate inference:
    - Loopy belief propagation, variational inference, Monte Carlo sampling

Learning BNs (or GMs)

The goal:

Given set of independent samples (assignments of random variables), find the best (the most likely?) Bayesian Network (both DAG and CPDs)

<table>
<thead>
<tr>
<th>(B,E,A,C,R)</th>
<th>E</th>
<th>B</th>
<th>R</th>
<th>A</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>(T,F,F,T,F)</td>
<td>0.9</td>
<td>0.1</td>
<td>0.2</td>
<td>0.8</td>
<td>0.01</td>
</tr>
</tbody>
</table>
MLE for general BN parameters

- If we assume the parameters for each CPD are globally independent, and all nodes are fully observed, then the log-likelihood function decomposes into a sum of local terms, one per node:

\[
\ell(\theta; D) = \log p(D \mid \theta) = \log \prod_{i} \prod_{x_i} p(x_{i,j} \mid x_{n.i}, \theta_j) = \sum_{i} \sum_{x_i} \log p(x_{i,j} \mid x_{n.i}, \theta_j)
\]

Example: decomposable likelihood of a directed model

- Consider the distribution defined by the directed acyclic GM:

\[
p(x \mid \theta) = p(x_1 \mid \theta_1) p(x_2 \mid x_1, \theta_1) p(x_3 \mid x_1, \theta_2) p(x_4 \mid x_2, x_3, \theta_1)
\]

- This is exactly like learning four separate small BNs, each of which consists of a node and its parents.
E.g.: MLE for BNs with tabular CPDs

- Assume each CPD is represented as a table (multinomial) where
  \[ \theta_{jk}^{\text{def}} = p(X_j = j \mid X_{-j} = k) \]
  - Note that in case of multiple parents, \( X_{-j} \) will have a composite state, and the CPD will be a high-dimensional table
  - The sufficient statistics are counts of family configurations
    \[ n_{jk}^{\text{def}} = \sum_i x_i^j x_i^{X_{-j}} \]
  - The log-likelihood is
    \[ \ell(\theta, D) = \log \prod_{i,j,k} \theta_{ijk}^{n_{ijk}} = \sum_{i,j,k} n_{ijk} \log \theta_{ijk} \]
  - Using a Lagrange multiplier to enforce \( \sum_j \theta_{jk} = 1 \), we get:
    \[ \theta_{jk}^{\text{ML}} = \frac{n_{ijk}}{\sum_{i,j,k'} n_{ij'k}} \]

What if some nodes are not observed?

- Consider the distribution defined by the directed acyclic GM:
  \[ p(x \mid \theta) = p(x_1 \mid \theta_1) p(x_2 \mid x_1, \theta_1) p(x_3 \mid x_1, \theta_2) p(x_4 \mid x_2, x_3, \theta_3) \]
- Need to compute \( p(x_4 \mid X_3) \) → inference
Summary

- Represent dependency structure with a directed acyclic graph
  - Node ↔ random variable
  - Edges encode dependencies
    - Absence of edge → conditional independence
  - Plate representation
  - A BN is a database of prob. Independence statement on variables

- The factorization theorem of the joint probability
  - Local specification → globally consistent distribution
  - Local representation for exponentially complex state-space

- Support efficient inference and learning – next lecture