Graphical Models - Part I

Greg Mori - CMPT 419/726

Bishop PRML Ch. 8, some slides from Russell and Norvig
AIMA2e
Outline

Probabilistic Models

Bayesian Networks

Markov Random Fields

Inference
Outline

Probabilistic Models

Bayesian Networks

Markov Random Fields

Inference
Probabilistic Models

- We now turn our focus to probabilistic models for pattern recognition
  - Probabilities express beliefs about uncertain events, useful for decision making, combining sources of information
- Key quantity in probabilistic reasoning is the joint distribution
  \[ p(x_1, x_2, \ldots, x_K) \]
  where \( x_1 \) to \( x_K \) are all variables in model
- Address two problems
  - **Inference**: answering queries given the joint distribution
  - **Learning**: deciding what the joint distribution is (involves inference)
- All inference and learning problems involve manipulations of the joint distribution
Reminder - Three Tricks

- **Bayes’ rule:**

\[
p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} = \alpha p(X|Y)p(Y)
\]

- **Marginalization:**

\[
p(X) = \sum_y p(X, Y = y) \quad \text{or} \quad p(X) = \int p(X, Y = y) dy
\]

- **Product rule:**

\[
p(X, Y) = p(X)p(Y|X)
\]

- **All 3 work with extra conditioning, e.g.:**

\[
p(X|Z) = \sum_y p(X, Y = y|Z)
\]

\[
p(Y|X, Z) = \alpha p(X|Y, Z)p(Y|Z)
\]
Joint Distribution

Consider model with 3 boolean random variables: cavity, catch, toothache

Can answer query such as

\[ p(\neg \text{cavity} | \text{toothache}) \]
Consider model with 3 boolean random variables: \(\text{cavity}, \text{catch}, \text{toothache}\).

Can answer query such as

\[ p(\neg \text{cavity} | \text{toothache}) \]
Consider model with 3 boolean random variables: \( \textit{cavity} \), \( \textit{catch} \), \( \textit{toothache} \)

Can answer query such as \( p(\neg \text{cavity}| \text{toothache}) = p(\neg \text{cavity}, \text{toothache}) / p(\text{toothache}) \)

\[
p(\neg \text{cavity}| \text{toothache}) = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4
\]
Joint Distribution

Consider model with 3 boolean random variables: \textit{cavity}, \textit{catch}, \textit{toothache}

Can answer query such as

\[
p(\neg\text{cavity}|\text{toothache}) = \frac{p(\neg\text{cavity}, \text{toothache})}{p(\text{toothache})}
\]

\[
p(\neg\text{cavity}|\text{toothache}) = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4
\]
In general, to answer a query on random variables $Q = Q_1, \ldots, Q_N$ given evidence $E = e, E = E_1, \ldots, E_M$, $e = e_1, \ldots, e_M$:

$$p(Q|E = e) = \frac{p(Q, E = e)}{p(E = e)}$$

$$= \frac{\sum_h p(Q, E = e, H = h)}{\sum_{q,h} p(Q = q, E = e, H = h)}$$
Problems

• The joint distribution is large
  • e.g. with $K$ boolean random variables, $2^K$ entries
• Inference is slow, previous summations take $O(2^K)$ time
• Learning is difficult, data for $2^K$ parameters
• Analogous problems for continuous random variables
• $A$ and $B$ are independent iff
  
  $$p(A|B) = p(A) \quad \text{or} \quad p(B|A) = p(B) \quad \text{or} \quad p(A, B) = p(A)p(B)$$

• $p(\text{Toothache, Catch, Cavity, Weather}) = p(\text{Toothache, Catch, Cavity})p(\text{Weather})$
  
  • 32 entries reduced to 12 ($\text{Weather}$ takes one of 4 values)

• Absolute independence powerful but rare

• Dentistry is a large field with hundreds of variables, none of which are independent. What to do?
Reminder - Conditional Independence

- $p(\text{Toothache}, \text{Cavity}, \text{Catch})$ has $2^3 - 1 = 7$ independent entries

- If I have a cavity, the probability that the probe catches in it doesn’t depend on whether I have a toothache:
  (1) $P(\text{catch}|\text{toothache}, \text{cavity}) = P(\text{catch}|\text{cavity})$

- The same independence holds if I haven’t got a cavity:
  (2) $P(\text{catch}|\text{toothache}, \neg \text{cavity}) = P(\text{catch}|\neg \text{cavity})$

- Catch is conditionally independent of Toothache given Cavity: $p(\text{Catch}|\text{Toothache}, \text{Cavity}) = p(\text{Catch}|\text{Cavity})$

- Equivalent statements:
  - $p(\text{Toothache}|\text{Catch}, \text{Cavity}) = p(\text{Toothache}|\text{Cavity})$
  - $p(\text{Toothache}, \text{Catch}|\text{Cavity}) = p(\text{Toothache}|\text{Cavity})p(\text{Catch}|\text{Cavity})$
  - Toothache $\perp \perp$ Catch|Cavity
Conditional Independence contd.

- Write out full joint distribution using chain rule:
  \[ p(Toothache, Catch, Cavity) = p(Toothache|Catch, Cavity)p(Catch, Cavity) = p(Toothache|Catch, Cavity)p(Catch|Cavity)p(Cavity) = p(Toothache|Cavity)p(Catch|Cavity)p(Cavity) \]
  \[ 2 + 2 + 1 = 5 \text{ independent numbers} \]

- In many cases, the use of conditional independence greatly reduces the size of the representation of the joint distribution
Graphical Models

- Graphical Models provide a visual depiction of probabilistic model
- Conditional independence assumptions can be seen in graph
- Inference and learning algorithms can be expressed in terms of graph operations
- We will look at 2 types of graph (can be combined)
  - Directed graphs: Bayesian networks
  - Undirected graphs: Markov Random Fields
  - Factor graphs (won’t cover)
Outline

Probabilistic Models

Bayesian Networks

Markov Random Fields

Inference
Bayesian Networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

- Syntax:
  - a set of nodes, one per variable
  - a directed, acyclic graph (link $\approx$ “directly influences”)
  - a conditional distribution for each node given its parents:
    \[
    p(X_i | pa(X_i))
    \]

- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over $X_i$ for each combination of parent values
Example

- Topology of network encodes conditional independence assertions:
  - Weather is independent of the other variables
  - Toothache and Catch are conditionally independent given Cavity
Example

• I’m at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn’t call. Sometimes it’s set off by minor earthquakes. Is there a burglar?

• Variables: Burglar, Earthquake, Alarm, JohnCalls, MaryCalls

• Network topology reflects “causal” knowledge:
  • A burglar can set the alarm off
  • An earthquake can set the alarm off
  • The alarm can cause Mary to call
  • The alarm can cause John to call
Example contd.

P(B) = 0.001
P(E) = 0.002

| B | E | P(A|B,E) |
|---|---|---------|
| T | T | 0.95    |
| T | F | 0.94    |
| F | T | 0.29    |
| F | F | 0.001   |

P(J|A) =
| A | P(J|A) |
|---|-------|
| T | 0.90  |
| F | 0.05  |

P(M|A) =
| A | P(M|A) |
|---|-------|
| T | 0.70  |
| F | 0.01  |
Compactness

- A CPT for Boolean $X_i$ with $k$ Boolean parents has $2^k$ rows for the combinations of parent values.
- Each row requires one number $p$ for $X_i = true$ (the number for $X_i = false$ is just $1 - p$).
- If each variable has no more than $k$ parents, the complete network requires $O(n \cdot 2^k)$ numbers.
- i.e., grows linearly with $n$, vs. $O(2^n)$ for the full joint distribution.
- For burglary net, ?? numbers.
Compactness

- A CPT for Boolean $X_i$ with $k$ Boolean parents has $2^k$ rows for the combinations of parent values.
- Each row requires one number $p$ for $X_i = true$ (the number for $X_i = false$ is just $1 - p$).
- If each variable has no more than $k$ parents, the complete network requires $O(n \cdot 2^k)$ numbers.
- i.e., grows linearly with $n$, vs. $O(2^n)$ for the full joint distribution.
- For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$).
Global Semantics

- **Global semantics** defines the full joint distribution as the product of the local conditional distributions:

\[
P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | pa(X_i))
\]

E.g., \(P(j \land m \land a \land \neg b \land \neg e) = \\)
Global Semantics

- Global semantics defines the full joint distribution as the product of the local conditional distributions:

\[
P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|pa(X_i))
\]

E.g., \( P(j \land m \land a \land \neg b \land \neg e) = \)

\[
P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) = 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\
\approx 0.00063
\]
Constructing Bayesian Networks

- Need a method such that a series of locally testable assertions of conditional independence guarantees the required global semantics

  1. Choose an ordering of variables $X_1, \ldots, X_n$
  2. For $i = 1$ to $n$
      - add $X_i$ to the network
      - select parents from $X_1, \ldots, X_{i-1}$ such that
      $p(X_i|pa(X_i)) = p(X_i|X_1, \ldots, X_{i-1})$

- This choice of parents guarantees the global semantics:

  $p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i|X_1, \ldots, X_{i-1})$ (chain rule)

  $= \prod_{i=1}^{n} p(X_i|pa(X_i))$ (by construction)
Constructing Bayesian Networks

- Need a method such that a series of locally testable assertions of conditional independence guarantees the required global semantics

  1. Choose an ordering of variables $X_1, \ldots, X_n$
  2. For $i = 1$ to $n$
     - Add $X_i$ to the network
     - Select parents from $X_1, \ldots, X_{i-1}$ such that $p(X_i|\text{pa}(X_i)) = p(X_i|X_1, \ldots, X_{i-1})$

- This choice of parents guarantees the global semantics:

\[
p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i|X_1, \ldots, X_{i-1}) \quad \text{(chain rule)}
\]

\[
= \prod_{i=1}^{n} p(X_i|\text{pa}(X_i)) \quad \text{(by construction)}
\]
Example

Suppose we choose the ordering $M, J, A, B, E$

$P(J|M) = P(J)$?
Example

Suppose we choose the ordering $M, J, A, B, E$

$P(J|M) = P(J)? \quad \text{No}$

$P(A|J, M) = P(A|J)? \quad P(A|J, M) = P(A)?$
Example

Suppose we choose the ordering $M, J, A, B, E$

\[
P(J|M) = P(J)? \quad \text{No}
\]
\[
P(A|J, M) = P(A|J)? \quad P(A|J, M) = P(A)? \quad \text{No}
\]
\[
P(B|A, J, M) = P(B|A)?
\]
\[
P(B|A, J, M) = P(B)?
\]
Example

Suppose we choose the ordering $M, J, A, B, E$

\[
P(J|M) = P(J) \quad \text{No}
\]
\[
P(A|J, M) = P(A|J) \quad P(A|J, M) = P(A) \quad \text{No}
\]
\[
P(B|A, J, M) = P(B|A) \quad \text{Yes}
\]
\[
P(B|A, J, M) = P(B) \quad \text{No}
\]
\[
P(E|B, A, J, M) = P(E|A) \quad ?
\]
\[
P(E|B, A, J, M) = P(E|A, B) \quad ?
\]
Example

Suppose we choose the ordering $M, J, A, B, E$

\[
P(J|M) = P(J) \quad \text{No}
\]
\[
P(A|J, M) = P(A|J) \quad P(A|J, M) = P(A) \quad \text{No}
\]
\[
P(B|A, J, M) = P(B|A) \quad \text{Yes}
\]
\[
P(B|A, J, M) = P(B) \quad \text{No}
\]
\[
P(E|B, A, J, M) = P(E|A) \quad \text{No}
\]
\[
P(E|B, A, J, M) = P(E|A, B) \quad \text{Yes}
\]
Example contd.

- Deciding conditional independence is hard in noncausal directions
  - (Causal models and conditional independence seem hardwired for humans!)
- Assessing conditional probabilities is hard in noncausal directions
- Network is less compact: $1 + 2 + 4 + 2 + 4 = 13$ numbers needed
Example - Car Insurance
Example - Polynomial Regression

- Bayesian polynomial regression model
- Observations $t = (t_1, \ldots, t_N)$
- Vector of coefficients $w$
- Inputs $x$ and noise variance $\sigma^2$ were assumed fixed, not stochastic and hence not in model
- Joint distribution:

$$p(t, w) = p(w) \prod_{n=1}^{N} p(t_n | w)$$
Plates

- A shorthand for writing repeated nodes such as the $t_n$ uses plates
Deterministic Model Parameters

Can also include deterministic parameters (not stochastic) as small nodes

Bayesian polynomial regression model:

\[ p(t, w | x, \alpha, \sigma^2) = p(w | \alpha) \prod_{n=1}^{N} p(t_n | w, x_n, \sigma^2) \]
In polynomial regression, we assumed we had a training set of $N$ pairs $(x_n, t_n)$.

Convention is to use shaded nodes for observed random variables.
Suppose we wished to predict the value $\hat{t}$ for a new input $\hat{x}$.

The Bayesian network used for this inference task would be this one.
Specifying Distributions - Discrete Variables

- Earlier we saw the use of conditional probability tables (CPT) for specifying a distribution over discrete random variables with discrete-valued parents.
- For a variable with no parents, with \( K \) possible states:

\[
p(x|\mu) = \prod_{k=1}^{K} \mu_k^{x_k}
\]

- e.g. \( p(B) = 0.001^B_1 0.999^B_2 \), 1-of-\( K \) representation.
Specifying Distributions - Discrete Variables cont.

- With two variables $x_1, x_2$ can have two cases

  - Dependent
    \[
    p(x_1, x_2 | \mu) = p(x_1 | \mu)p(x_2 | x_1, \mu) = \left( \prod_{k=1}^{K} \mu_{k1}^{x_{1k}} \right) \left( \prod_{k=1}^{K} \prod_{j=1}^{K} \mu_{kj2}^{x_{1k}x_{2j}} \right)
    \]
    - $K^2 - 1$ free parameters in $\mu$

  - Independent
    \[
    p(x_1, x_2 | \mu) = p(x_1 | \mu)p(x_2 | \mu) = \left( \prod_{k=1}^{K} \mu_{k1}^{x_{1k}} \right) \left( \prod_{k=1}^{K} \mu_{k2}^{x_{2k}} \right)
    \]
    - $2(K - 1)$ free parameters in $\mu$
Chains of Nodes

- With $M$ nodes, could form a chain as shown above.
- Number of parameters is:

$$\underbrace{(K - 1)}_{x_1} + \underbrace{(M - 1)K(K - 1)}_{\text{others}}$$

- Compare to:
  - $K^M - 1$ for fully connected graph
  - $M(K - 1)$ for graph with no edges (all independent)
Another way to reduce number of parameters is sharing parameters (a. k. a. tying of parameters)

Lower graph reuses same $\mu$ for nodes 2-$M$
  - $\mu$ is a random variable in this network, could also be deterministic

$(K - 1) + K(K - 1)$ parameters
Specifying Distributions - Continuous Variables

- One common type of conditional distribution for continuous variables is the **linear-Gaussian**

\[
p(x_i | p_{a_i}) = \mathcal{N}(x_i; \sum_{j \in p_{a_i}} w_{ij} x_j + b_i, v_i)
\]
Specifying Distributions - Continuous Variables

- One common type of conditional distribution for continuous variables is the **linear-Gaussian**

\[
p(x_i | \text{pa}_i) = \mathcal{N} \left( x_i; \sum_{j \in \text{pa}_i} w_{ij} x_j + b_i, v_i \right)
\]

- e.g. With one parent \textit{Harvest}:

\[
p(c|h) = \mathcal{N} (c; -0.5h + 5, 1)
\]

- For harvest \(h\), mean cost is 
  \(-0.5h + 5\), variance is \(1\)
Linear Gaussian

- Interesting fact: if all nodes in a Bayesian Network are linear Gaussian, joint distribution is a multivariate Gaussian

\[
p(x_i | p_{ai}) = \mathcal{N} \left( x_i; \sum_{j \in p_{ai}} w_{ij} x_j + b_i, v_i \right)
\]

\[
p(x_1, \ldots, x_N) = \prod_{i=1}^{N} \mathcal{N} \left( x_i; \sum_{j \in p_{ai}} w_{ij} x_j + b_i, v_i \right)
\]

- Each factor looks like \(\exp((x_i - (w^T_i x_{pa_i})^2)\), this product will be another quadratic form
- With no links in graph, end up with diagonal covariance matrix
- With fully connected graph, end up with full covariance matrix
Conditional Independence in Bayesian Networks

- Recall again that \( a \) and \( b \) are conditionally independent given \( c \) (\( a \perp\!\!\!\!\perp b|c \)) if
  - \( p(a|b, c) = p(a|c) \) or equivalently
  - \( p(a, b|c) = p(a|c)p(b|c) \)

- Before we stated that links in a graph are \( \approx \) “directly influences”

- We now develop a correct notion of links, in terms of the conditional independences they represent
  - This will be useful for general-purpose inference methods
A Tale of Three Graphs - Part 1

- The graph above means

\[ p(a, b, c) = p(a|c)p(b|c)p(c) \]
\[ p(a, b) = \sum_c p(a|c)p(b|c)p(c) \]


- So \( a \) and \( b \) not independent
However, conditioned on $c$

$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a|c)p(b|c)p(c)}{p(c)} = p(a|c)p(b|c)$$

So $a \indep b|c$
A Tale of Three Graphs - Part 1

- Note the path from $a$ to $b$ in the graph
  - When $c$ is not observed, path is open, $a$ and $b$ not independent
  - When $c$ is observed, path is blocked, $a$ and $b$ independent
- In this case $c$ is tail-to-tail with respect to this path
The graph above means

\[ p(a, b, c) = p(a)p(b|c)p(c|a) \]

Again \( a \) and \( b \) not independent
However, conditioned on $c$

$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a)p(b|c)}{p(c)}p(c|a)$$

$$= \frac{p(a)p(b|c)}{p(c)} \frac{p(a|c)p(c)}{p(a)}$$

Bayes' Rule

$$= p(a|c)p(b|c)$$

So $a \perp\!\!\!\!\!\!\perp b|c$
A Tale of Three Graphs - Part 2

- As before, the path from $a$ to $b$ in the graph
  - When $c$ is not observed, path is open, $a$ and $b$ not independent
  - When $c$ is observed, path is blocked, $a$ and $b$ independent
- In this case $c$ is head-to-tail with respect to this path
A Tale of Three Graphs - Part 3

- The graph above means

\[ p(a, b, c) = p(a)p(b)p(c|a, b) \]

\[ p(a, b) = \sum_c p(a)p(b)p(c|a, b) \]

\[ = p(a)p(b) \]

- This time \(a\) and \(b\) are independent
However, conditioned on $c$

$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a)p(b)p(c|a, b)}{p(c)} \neq p(a|c)p(b|c) \text{ in general}$$

So $a \perp b | c$
Frustratingly, the behaviour here is different

- When $c$ is not observed, path is blocked, $a$ and $b$ independent
- When $c$ is observed, path is unblocked, $a$ and $b$ not independent

- In this case $c$ is head-to-head with respect to this path
- Situation is in fact more complex, path is unblocked if any descendent of $c$ is observed
Part 3 - Intuition

- Binary random variables $B$ (battery charged), $F$ (fuel tank full), $G$ (fuel gauge reads full)
- $B$ and $F$ independent
- But if we observe $G = 0$ (false) things change
  - e.g. $p(F = 0|G = 0, B = 0)$ could be less than $p(F = 0|G = 0)$, as $B = 0$ explains away the fact that the gauge reads empty
  - Recall that $p(F|G, B) = p(F|G)$ is another $F \perp \perp B|G$
D-separation

- A general statement of conditional independence
- For sets of nodes $A$, $B$, $C$, check all paths from $A$ to $B$ in graph
- If all paths are blocked, then $A \perp\!
\!
\!\perp B|C$
- Path is blocked if:
  - Arrows meet head-to-tail or tail-to-tail at a node in $C$
  - Arrows meet head-to-head at a node, and neither node nor any descendent is in $C$
Naive Bayes

- Commonly used **naive Bayes** classification model
- Class label $z$, features $x_1, \ldots, x_D$
- Model assumes features independent given class label
  - **Tail-to-tail** at $z$, blocks path between features
• What is the minimal set of nodes which makes a node $x_i$ conditionally independent from the rest of the graph?
  • $x_i$’s parents, children, and children’s parents (co-parents)

• Define this set $MB$, and consider:

$$p(x_i|x_{\{j \neq i\}}) = \frac{p(x_1, \ldots, x_D)}{\int p(x_1, \ldots, x_D)dx_i} = \frac{\prod_k p(x_k|pa_k)}{\int \prod_k p(x_k|pa_k)dx_i}$$

• All factors other than those for which $x_i$ is $x_k$ or in $pa_k$ cancel
Learning Parameters

- When all random variables are observed in training data, relatively straight-forward
  - Distribution factors, all factors observed
  - e.g. Maximum likelihood used to set parameters of each distribution $p(x_i|pa_i)$ separately

- When some random variables not observed, it’s tricky
  - This is a common case
  - Expectation-maximization (later) is a method for this
Probabilistic Models

Bayesian Networks

Markov Random Fields

Inference
Outline

Probabilistic Models
Bayesian Networks
Markov Random Fields
Inference