Graphical Models - Short Review

Oliver Schulte

Outline

Outline

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:

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

- Product rule:

$$p(X, Y) = p(X)p(Y|X)$$

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

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

$$p(Y|X, Z) = \alpha p(X|Y, Z)p(Y|Z)$$

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

Graphical Models

- Graphical Models provide a visual depiction of probabilistic model
- Conditional indepence assumptions can be seen in graph
- Inference and learning algorithms can be expressed in terms of graph operations
- We will look at 3 types of graph (can be combined)
    - Directed graphs: Bayesian networks
    - Undirected graphs: Markov Random Fields
    - Factor graphs

Outline

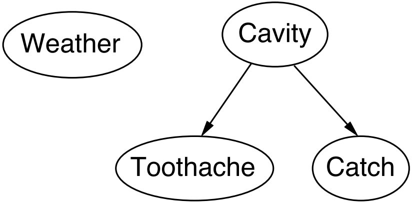
Bayesian Networks

- 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

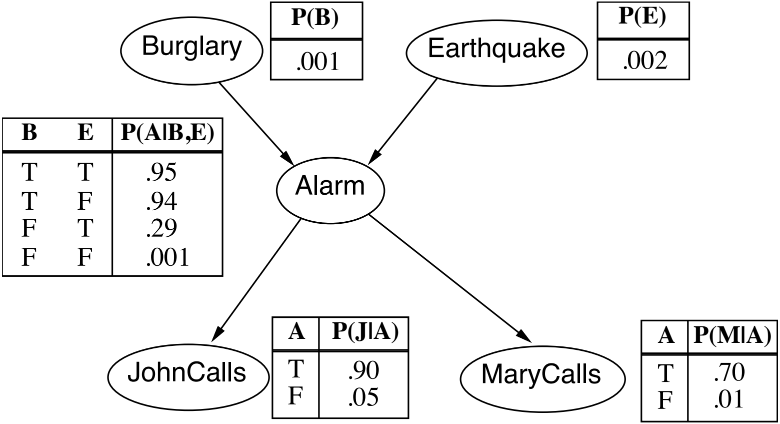


- Topology of network encodes conditional independence assertions:
  - *Weather* is independent of the other variables
  - *Toothache* and *Catch* are conditionally independent given *Cavity*

Example

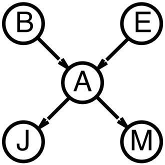
- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?
- Variables: $Burglar$, $Earthquake$, $Alarm$, $JohnCalls$, $MaryCalls$
- Network topology reflects causal knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call
- (Causal models and conditional independence seem hardwired for humans!)

Example contd.



Global Semantics

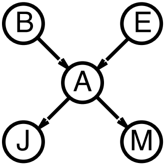
- Global semantics defines the full joint distribution as the product of the local conditional distributions:



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

e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e) =$

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., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e) =$

$$\begin{aligned} & P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \\ = \quad & 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ \approx \quad & 0.00063 \end{aligned}$$

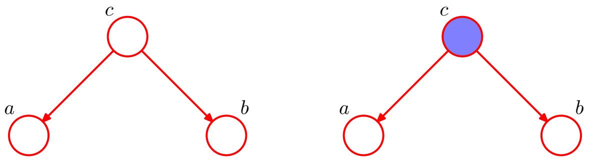
Conditional Independence in Bayesian Networks

- 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 $\approx$ "direct influences"
- We now develop a criterion for what conditional independences (the absence of) links represent.
  - This will be useful for general-purpose inference methods
  - It provides a fast solution to the *relevance problem*: determine whether X is relevant to Y given knowledge of Z.

D-separation

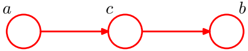
- A general statement of conditional independence
- For sets of nodes A, B, C, check all paths from A to B in graph
- If all paths are blocked, then $A \perp\!\!\!\perp B | C$
- Path is blocked if:
  - Arrows meet head-to-tail or tail-to-tail at a node in C
  - Arrows meet head-to-head at a node—the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

A Tale of Three Graphs - Part 2

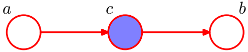


- The graph above means

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

- Again a and b not independent

A Tale of Three Graphs - Part 2



- However, conditioned on c

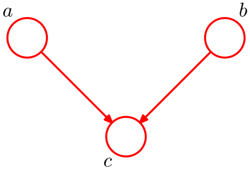
$$
\begin{aligned}
p(a,b|c) &= \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(b|c)}{p(c)}p(c|a) \\
&= \frac{p(a)p(b|c)}{p(c)} \underbrace{\frac{p(a|c)p(c)}{p(a)}}_{\text{Bayes' Rule}} \\
&= p(a|c)p(b|c)
\end{aligned}
$$

- 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

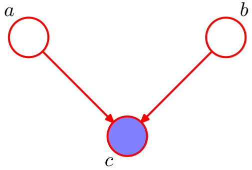


- The graph above means

$$
\begin{aligned}
p(a, b, c) &= p(a)p(b)p(c|a, b) \\
p(a, b) &= \sum_c p(a)p(b)p(c|a, b) \\
&= p(a)p(b)
\end{aligned}
$$

- This time a and b are independent

A Tale of Three Graphs - Part 3

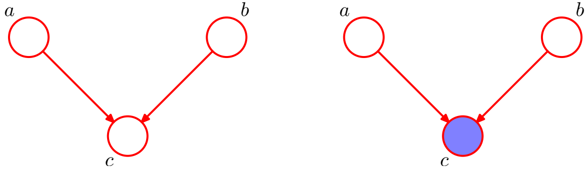


- However, conditioned on c

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

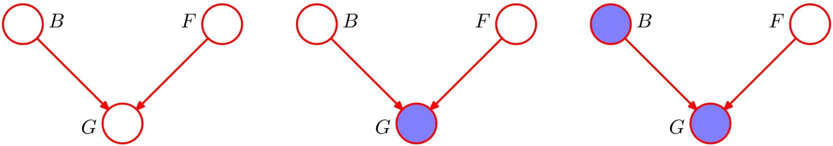
- So a is dependent on b given c

A Tale of Three Graphs - Part 3



- 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

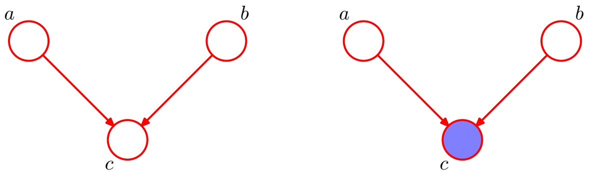
Part 3 - Intuition



- 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 $F \perp\!\!\!\perp B|G$

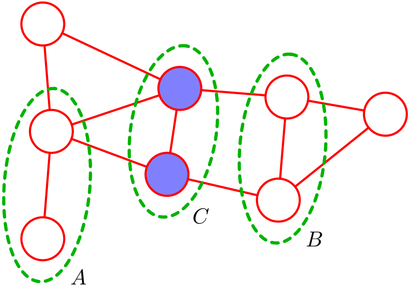
Outline

Conditional Independence in Graphs



- Recall that for Bayesian Networks, conditional independence was a bit complicated
  - d-separation with head-to-head links
- We would like to construct a graphical representation such that conditional independence is straight-forward path checking

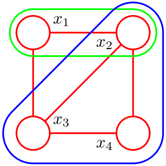
Markov Random Fields



- Markov random fields (MRFs) contain one node per variable
- Undirected graph over these nodes
- Conditional independence will be given by simple separation, blockage by observing a node on a path
    - e.g. in above graph, $A \perp\!\!\!\perp B | C$

Cliques

- A clique in a graph is a subset of nodes such that there is a link between every pair of nodes in the subset
- A maximal clique is a clique for which one cannot add another node and have the set remain a clique



MRF Joint Distribution

- Note that nodes in a clique cannot be made conditionally independent from each other
  - So defining factors $\psi(\cdot)$ on nodes in a clique is "safe"
- The joint distribution for a Markov random field is:

$$p(x_1, \ldots, x_K) = \frac{1}{Z} \prod_C \psi_C(\boldsymbol{x}_C)$$

  where $\boldsymbol{x}_C$ is the set of nodes in clique C, and the product runs over all maximal cliques
- Each $\psi_C(\boldsymbol{x}_C) \geq 0$
- Z is a normalization constant

MRF Joint - Terminology

- The joint distribution for a Markov random field is:

$$p(x_1, \ldots, x_K) = \frac{1}{Z} \prod_C \psi_C(\boldsymbol{x}_C)$$

- Each $\psi_C(\boldsymbol{x}_C) \geq 0$ is called a potential function
- Z, the normalization constant, is called the partition function:

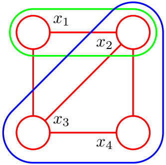
$$Z = \sum_{\boldsymbol{x}} \prod_C \psi_C(\boldsymbol{x}_C)$$

  - Z is very costly to compute, since it is a sum/integral over all possible states for all variables in $\boldsymbol{x}$
  - Don't always need to evaluate it though, will cancel for computing conditional probabilities

MRF Joint Distribution Example

- The joint distribution for a Markov random field is:

$$
\begin{aligned}
p(x_1, \ldots, x_4) &= \frac{1}{Z} \prod_C \psi_C(\boldsymbol{x}_C) \\
&= \frac{1}{Z} \psi_{123}(x_1, x_2, x_3) \psi_{234}(x_2, x_3, x_4)
\end{aligned}
$$



- Note that maximal cliques subsume smaller ones: $\psi_{123}(x_1, x_2, x_3)$ could include $\psi_{12}(x_1, x_2)$, though sometimes smaller cliques are explicitly used for clarity

Hammersley-Clifford

- The definition of the joint:

$$p(x_1, \ldots, x_K) = \frac{1}{Z} \prod_C \psi_C(\boldsymbol{x}_C)$$

- Note that we started with particular conditional independences
- We then formulated the factorization based on clique potentials
  - This formulation resulted in the right conditional independences
- The converse is true as well, any strictly positive distribution with the conditional independences given by the undirected graph can be represented using a product of clique potentials
- This is the Hammersley-Clifford theorem

Energy Functions

- Often use exponential, which is non-negative, to define potential functions:

$$\psi_C(\boldsymbol{x}_C) = \exp\{-E_C(\boldsymbol{x}_C)\}$$

- Minus sign $-$ by convention
- $E_C(\boldsymbol{x}_C)$ is called an energy function
  - From physics, low energy = high probability
- This exponential representation is known as the Boltzmann distribution

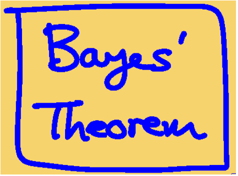
Energy Functions - Intuition

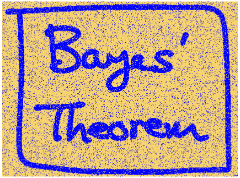
- Joint distribution nicely rearranges as

$$\begin{aligned} p(x_1, \ldots, x_K) &= \frac{1}{Z} \prod_C \psi_C(\boldsymbol{x}_C) \\ &= \frac{1}{Z} \exp\{-\sum_C E_C(\boldsymbol{x}_C)\} \end{aligned}$$

- Intuition about potential functions: ψ_C are describing good (low energy) sets of states for adjacent nodes
- An example of this is next

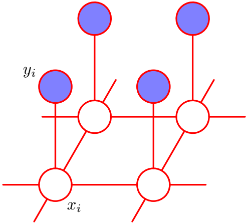
Image Denoising





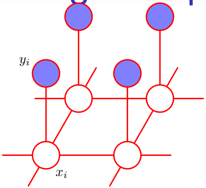
- Consider the problem of trying to correct (denoise) an image that has been corrupted
- Assume image is binary
- Observed (noisy) pixel values $y_i \in \{-1, +1\}$
- Unobserved true pixel values $x_i \in \{-1, +1\}$
- Another application: face sketch synthesis from photos `http://people.csail.mit.edu/celiu/FaceHallucination/fh.html`.

Image Denoising - Graphical Model



- Cliques containing each true pixel value $x_i \in \{-1, +1\}$ and observed value $y_i \in \{-1, +1\}$
  - Observed pixel value is usually same as true pixel value
  - Energy function $-\eta x_i y_i$, $\eta > 0$, lower energy (better) if $x_i = y_i$
- Cliques containing adjacent true pixel values x_i, x_j
  - Nearby pixel values are usually the same
  - Energy function $-\beta x_i x_j$, $\beta > 0$, lower energy (better) if $x_i = x_j$

Image Denoising - Graphical Model



- Complete energy function:

$$E(\boldsymbol{x}, \boldsymbol{y}) = -\beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i$$

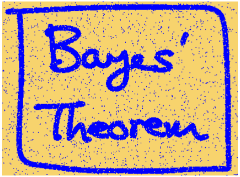
- Joint distribution:

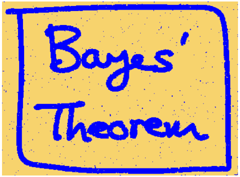
$$p(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z} \exp\{-E(\boldsymbol{x}, \boldsymbol{y})\}$$

- Or, as potential functions $\psi_n(x_i, x_j) = \exp(\beta x_i x_j)$, $\psi_p(x_i, y_i) = \exp(\eta x_i y_i)$:

$$p(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z} \prod_{i,j} \psi_n(x_i, x_j) \prod_i \psi_p(x_i, y_i)$$

Image Denoising - Inference





- The denoising query is $\arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y})$
- Two approaches:
  - Iterated conditional modes (ICM): hill climbing in $\boldsymbol{x}$, one variable x_i at a time
    - Simple to compute, conditional probability depends only on observation plus neighbouring pixels.
    - Demo `http://cs.stanford.edu/people/karpathy/visml/ising_example.html`
  - Graph cuts: formulate as max-flow/min-cut problem, exact inference.

Conclusion

- Graphical models depict conditional independence assumptions
- Directed graphs (Bayesian networks)
  - Factorization of joint distribution as conditional on node given parents
- Undirected graphs (Markov random fields)
  - Factorization of joint distribution as clique potential functions