Graphical Models - Part I

CMPT 419/726

Mo Chen
SFU Computing Science

Feb. 10, 2020

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

Outline

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

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	.108	.012	.072	.008
$\neg cavity$	.016	.064	.144	.576

- Consider model with 3 boolean random variables: $cavity$, $catch$, $toothache$
- Can answer query such as

$$p(\neg cavity | toothache)$$

Joint Distribution

	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	.108	.012	.072	.008
¬cavity	.016	.064	.144	.576

- Consider model with 3 boolean random variables: $cavity,\ catch, toothache$
- Can answer query such as

$$p(\neg cavity|toothache) = \frac{p(\neg cavity, toothache)}{p(toothache)}$$

$$p(\neg cavity|toothache) = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

Joint Distribution

- In general, to answer a query on random variables $\boldsymbol{Q} = Q_1, \dots, Q_N$ given evidence $\boldsymbol{E} = \boldsymbol{e}, \boldsymbol{E} = E_1, \dots, E_M, \boldsymbol{e} = e_1, \dots, e_M$:

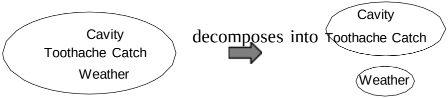
$$p(\boldsymbol{Q}|\boldsymbol{E} = e) = \frac{p(\boldsymbol{Q}, \boldsymbol{E} = \boldsymbol{e})}{P(\boldsymbol{E} = \boldsymbol{e})}$$

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

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

Reminder - Independence



- A and B are independent iff
  $p(A|B) = p(A)$ or $p(B|A) = p(B)$ or $p(A,B) = p(A)p(B)$
- $p(Toothache,\ Catch,\ Cavity,\ Weather) = p(Toothache,\ Catch,\ Cavity)p(Weather)$
  - 32 entries reduced to 12 (*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

- $p(Toothache,\ Cavity,\ Catch)$ 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:
$$P(catch|toothache, cavity) = P(catch|cavity)$$
- The same independence holds if I haven't got a cavity:
$$P(catch|toothache, \neg cavity) = P(catch|\neg cavity)$$
- $Catch$ is conditionally independent of $Toothache$ given $Cavity$: $p(Catch|Toothache, Cavity) = p(Catch|Cavity)$
- Equivalent statements:
  - $p(Toothache|Catch, Cavity) = p(Toothache|Cavity)$
  - $p(Toothache, Catch|Cavity) = p(Toothache|Cavity)p(Catch|Cavity)$
  - $Toothache \perp\!\!\!\perp Catch|Cavity$

Conditional Independence contd.

- Write out full joint distribution using chain rule:

$$p(Toothache, Catch, Cavity)$$
$$= p(Toothache|Catch, Cavity)p(Catch, Cavity)$$
$$= p(Toothache|Catch, Cavity)p(Catch|Cavity)p(Cavity)$$
$$= p(Toothache|Cavity)p(Catch|Cavity)p(Cavity)$$

  $2 + 2 + 1 = 5$ 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 indepence 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

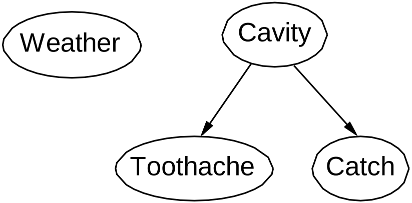
Bayesian Networks

- 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:

$$p\big(X_i | pa(X_i)\big)$$

- 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

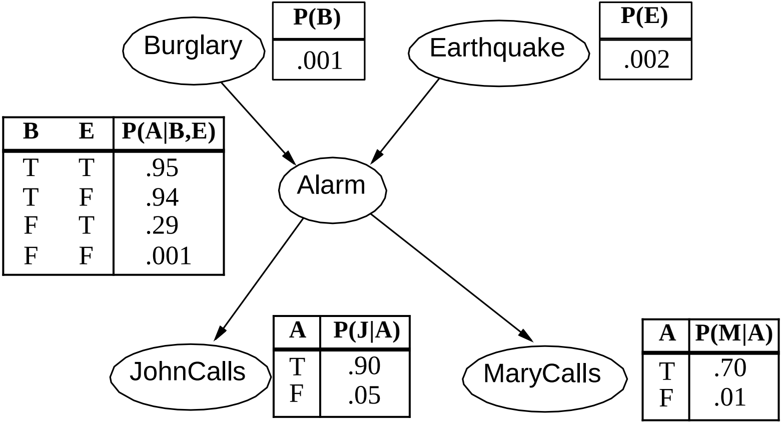


- Topology of network encodes conditional independence assertions:
  - $Weather$ is independent of the other variables
  - $Toothache$ and $Catch$ are conditionally independent given $Cavity$

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)
.001

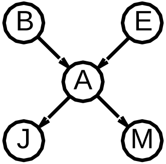
P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

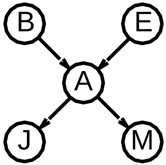
Compactness



- 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
  - $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$)

Global Semantics

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



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

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

$$P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e)$$
$$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998$$
$$\approx 0.00063$$

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 $X_1, \dots, X_n$
  2. For i= 1 to n
     add X_i to the network
     select parents from $X_1, \dots, X_{i-1}$ such that
     $$p\big(X_i | pa(X_i)\big) = p(X_i | X_1, \dots, X_{i-1})$$

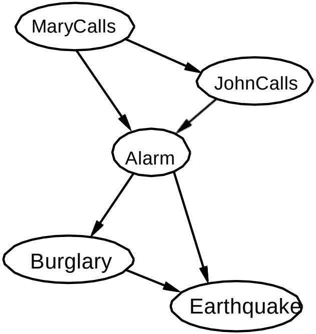
- This choice of parents guarantees the global semantics:

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

$$= \prod_{i=1}^{n} p\big(X_i | pa(X_i)\big) \qquad \text{(by construction)}$$

Example

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



$P(J|M) = P(J)$? No

$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$? No

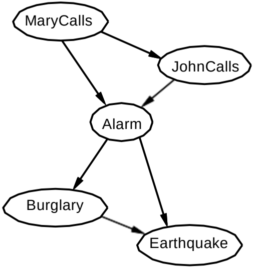
$P(B|A, J, M) = P(B|A)$? Yes

$P(B|A, J, M) = P(B)$? No

$P(E|B, A, J, M) = P(E|A)$? No

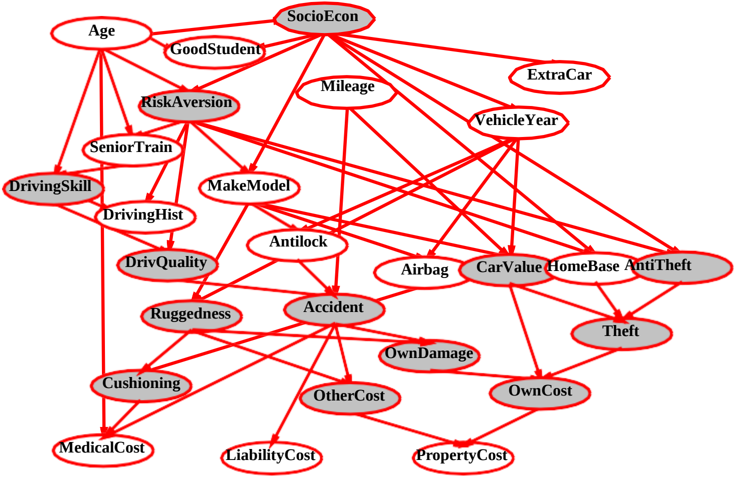
$P(E|B, A, J, M) = P(E|A, B)$? Yes

Example contd.

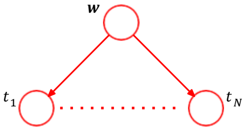


- Deciding conditional independence is hard in noncausal directions
  - (Causal models and conditional independence seem hardwired for humans!)
- Assessing conditional probabilities is hard in noncausal directions
- Network is less compact: $1 + 2 + 4 + 2 + 4 = 13$ numbers needed

Example - Car Insurance



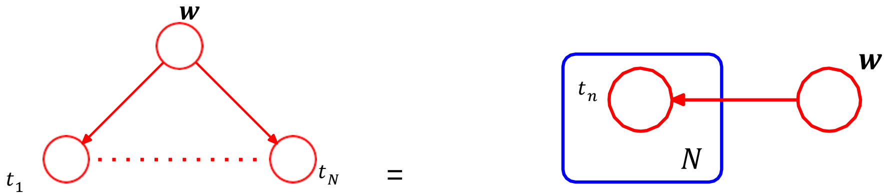
Example - Polynomial Regression



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

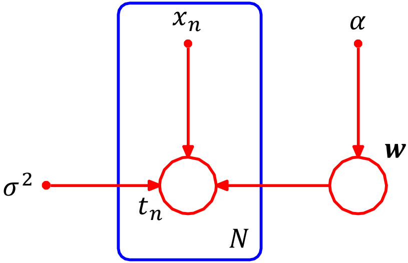
$$p(\boldsymbol{t}, \boldsymbol{w}) = p(\boldsymbol{w}) \prod_{n=1}^{N} p(t_n | \boldsymbol{w})$$

Plates



- A shorthand for writing repeated nodes such as the t_n uses plates

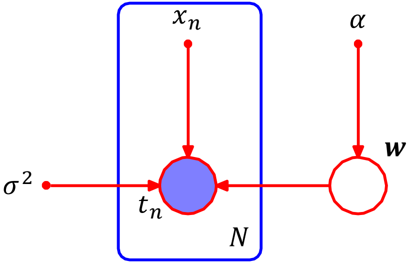
Deterministic Model Parameters



- Can also include deterministic parameters (not stochastic) as small nodes
- Bayesian polynomial regression model:

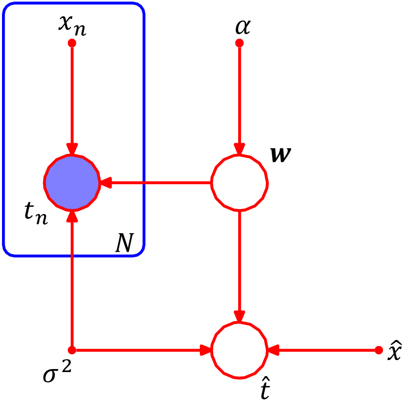
$$p(\boldsymbol{t}, \boldsymbol{w} | \boldsymbol{x}, \alpha, \sigma^2) = p(\boldsymbol{w}|\alpha) \prod_{n=1}^{N} p(t_n | \boldsymbol{w}, x_n, \sigma^2)$$

Observations



- In polynomial regression, we assumed we had a training set of N pairs $(\boldsymbol{x}_n, t_n)$
- Convention is to use shaded nodes for observed random variables

Predictions



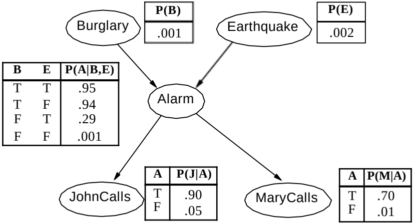
- Suppose we wished to predict the value $\hat{t}$ for a new input $\hat{x}$
- The Bayesian network used for this inference task would be this one

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:

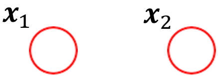
$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

  - e.g. $p(B) = 0.001^{B_1} 0.999^{B_2}$, 1-of-$K$ representation



Specifying Distributions - Discrete Variables cont.

- With two variables $\boldsymbol{x}_1, \boldsymbol{x}_2$ can have two cases



- Dependent

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2 | \boldsymbol{\mu}) = p(\boldsymbol{x}_1 | \boldsymbol{\mu}) p(\boldsymbol{x}_2 | \boldsymbol{x}_1, \mu)$$

$$= \left(\prod_{k=1}^{K} \mu_{k1}^{x_{1k}} \right) \left(\prod_{k=1}^{K} \prod_{j=1}^{K} \mu_{kj2}^{x_{1k} x_{2j}} \right)$$

- Independent

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2 | \boldsymbol{\mu}) = p(\boldsymbol{x}_1 | \boldsymbol{\mu}) p(\boldsymbol{x}_2 | \mu)$$

$$= \left(\prod_{k=1}^{K} \mu_{k1}^{x_{1k}} \right) \left(\prod_{k=1}^{K} \mu_{k2}^{x_{2k}} \right)$$

- $K^2 - 1$ free parameters in $\boldsymbol{\mu}$
- $2(K-1)$ free parameters in $\boldsymbol{\mu}$

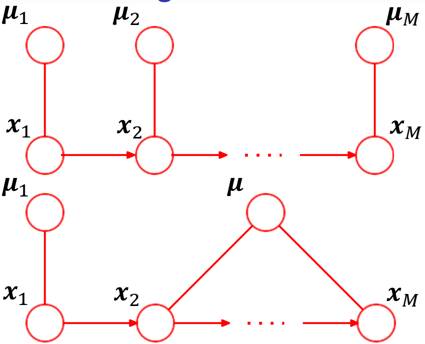
Chains of Nodes

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

$$\underbrace{(K-1)}_{\boldsymbol{x}_1} + (M-1)\underbrace{K(K-1)}_{\text{others}}$$

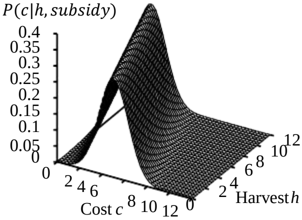
- Compare to:
  - $K^M - 1$ for fully connected graph
  - $M(K-1)$ for graph with no edges (all independent)

Sharing Parameters



- Another way to reduce number of parameters is sharing parameters (a. k. a. tying of parameters)
- Lower graph reuses same $\boldsymbol{\mu}$ for nodes $2 - M$
  - $\boldsymbol{\mu}$ 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

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

- e.g. With one parent *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

$$p(x_i|pa_i) = \mathcal{N}\left(x_i; \sum_{j \in pa_i} x_{ij}x_j + b_i, v_i\right)$$

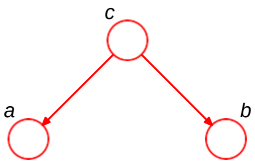
$$p(x_1, \ldots, x_N) = \prod_{i=1}^{N} \mathcal{N}\left(x_i; \sum_{j \in pa_i} x_{ij}x_j + b_i, v_i\right)$$

- Each factor looks like $\exp\left(\left(x_i - \boldsymbol{w}_i^\top \boldsymbol{x}_{pa_i}\right)^2\right)$, 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

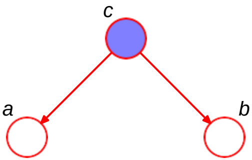


- The graph above means

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

- So a and b not independent

A Tale of Three Graphs - Part 1

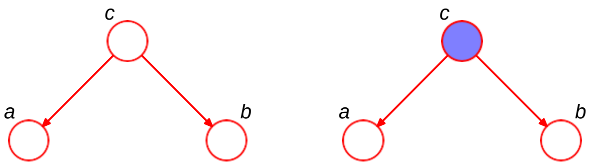


- However, conditioned on c,

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

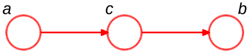
- So $a \perp\!\!\!\perp b|c$

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

A Tale of Three Graphs - Part 2

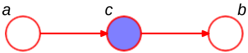


- The graph above means

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

- Again a and b not independent

A Tale of Three Graphs - Part 2



- However, conditioned on c

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

$$= \frac{p(a)p(b|c)}{p(c)}\underbrace{\frac{p(a|c)p(c)}{p(a)}}_{\text{Bayes' rule}}$$

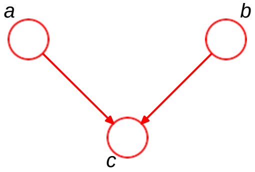
$$= p(a|c)p(b|c)$$

- 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

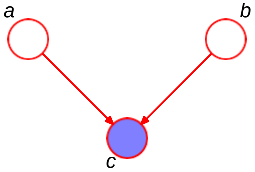


- The graph above means

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

- This time a and b are independent

A Tale of Three Graphs - Part 3



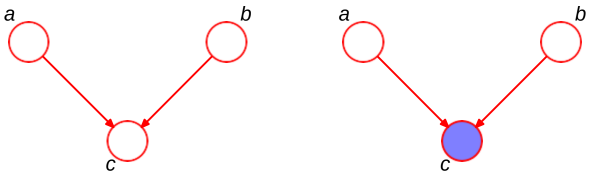
- However, conditioned on c

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

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

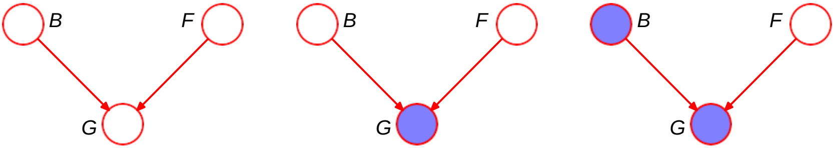
- So $a \perp\!\!\!\perp b|c$

A Tale of Three Graphs - Part 3



- 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

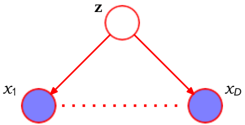


- Binary random variables B (battery charged), F (fuel tank full), G (fuel gauge reads full)
- B and F independent
- But if we observe $G = 0$ (false) things change
  - e.g. $p(F = 0|G = 0, B = 0)$ could be less than $p(F = 0|G = 0)$, as $B = 0$ explains away the fact that the gauge reads empty
  - Recall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

D-separation

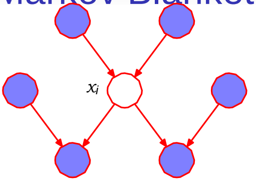
- A general statement of conditional independence
- For sets of nodes A, B, C, check all paths from A to B in graph
- If all paths are blocked, then $A \perp\!\!\!\perp B | C$
- Path is blocked if:
  - Arrows meet head-to-tail or tail-to-tail at a node in C
  - Arrows meet head-to-head at a node, and 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

Markov Blanket



- 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:

$$p\left(x_i | x_{\{j \neq i\}}\right) = \frac{p(x_1, \dots, x_D)}{\int p(x_1, \dots, x_D) dx_i}$$

$$= \frac{\prod_k p(x_k | pa_k)}{\int \prod_k p(x_k | pa_k)\, dx_i}$$

- 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

Outline

Outline