Slides originally taken from
http://research.microsoft.com/en-us/um/people/cmbishop/prml/
and modified by Pushpendre Rastogi for cs475-2017

GRAPHICAL MODELS

Approaches t

But you can still optimize this function !!

The goal in Machine Learning is to minimize True Risk

- True Risk = The **Expected** Loss = $E_{p(x,y)}[l(y, f_\theta(x))]$
  True Risk is a function that you can not observe.

Approach 1) ERM algo: To minimize (TR) you can minimize Empirical Risk (ER) by finding an optimal function from a family of functions.

- ***If*** data is *plentiful* ***Then*** minimization of ER $\Rightarrow$ minimization of TR
- Uses data to fit a function

Approach 2) Probabilistic Approach: Use data to estimate $\hat{p}_\theta(x, y)$ that approximates $p(x, y)$, then choose

$$f_\theta(x) = \arg\min_{\hat{y} \in \mathcal{Y}} E_{\hat{p}_\theta(y|x)}\left[l(y, \hat{y})\right]$$

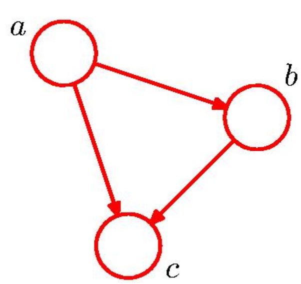
- Search for optimal $\hat{p}_\theta(y; x)$ or $\hat{p}_\theta(y, x)$ from some family of distributions
  - Graphical Models are a language for specifying a family of distributions.
  - Estimating $\hat{p}_\theta(x, y)$ requires Estimation Methods
  - Given $\hat{p}_\theta(x, y)$ we must perform inference to minimize Risk

Summary

- We can estimate the distribution of data from samples by searching in a family of distributions.
- Graphical Models are a high level language for specifying "families of distributions"
- Estimating the optimal parameters from a model family is *Parameter Estimation*.
- Using a distribution to make predictions/decisions is called *Inference*.

Bayesian Networks

- Definition: A BN is a Directed Acyclic Graph (DAG) of Random Variables

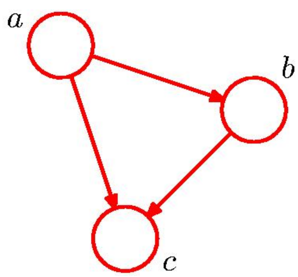


- a has two children
- b is one of the parents of *c*
- *a* has no parent
- c has two *ancestors,* but no descendants

Bayesian Networks

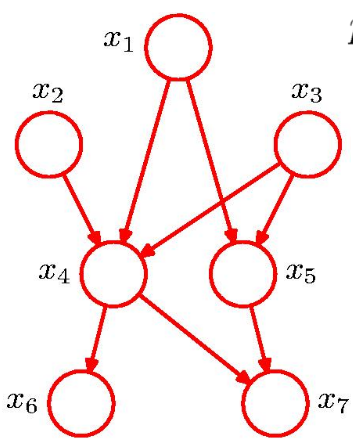
- Definition: A BN is a Directed Acyclic Graph (DAG)
  of Random Variables whose joint probability **factorizes** according to the graph.

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



Bayesian Networks (Example)

$$p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$



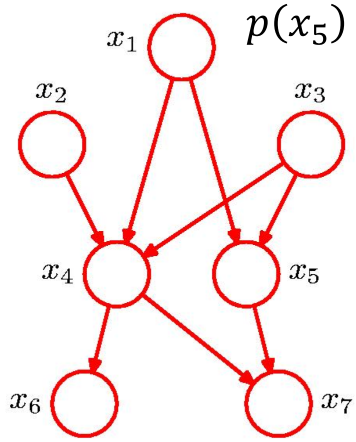
General Factorization

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$

Bayesian Networks (More Concrete Example)

$p(x_1) = \mathcal{N}(0, 1)$ *(Possible Implementation)*

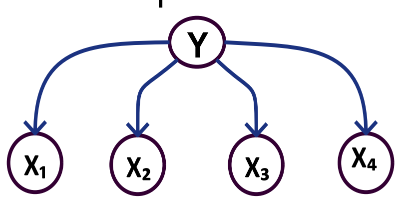
$p(x_5) = Bern(0.3\,\sigma(x_1) + 0.7\,\sigma(x_3))$



General Factorization

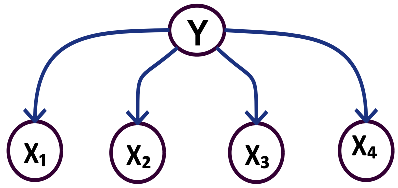
$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k | \mathrm{pa}_k)$$

Example: Naïve Bayes as a BN

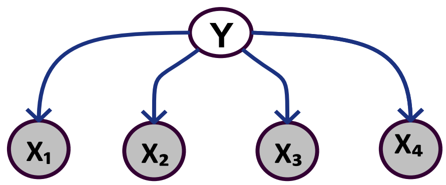


$\mathrm{p}(\mathrm{x}_1, x_2, x_3, x_4, y) = p(x_1 \mid y)\, p(x_2|y)\, p(x_3|y)\, p(x_4|y) p(y)$

Example: Naïve Bayes as a BN

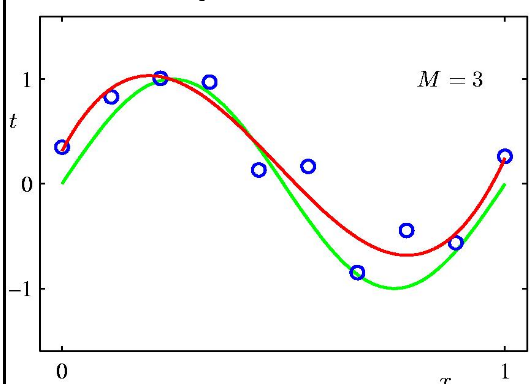


$$p(x_1, x_2, x_3, x_4, y) = p(x_1 \mid y)\, p(x_2|y)\, p(x_3|y)\, p(x_4|y) p(y)$$



$$p(y = c|\, x_1, \dots, x_4) = \frac{p(x_1,x_2,x_3,x_4,c)}{p(x_1,x_2,x_3,x_4)}$$

Bayesian Curve Fitting (1)



Model Summary

x *has no distribution.*

y is predicted value.

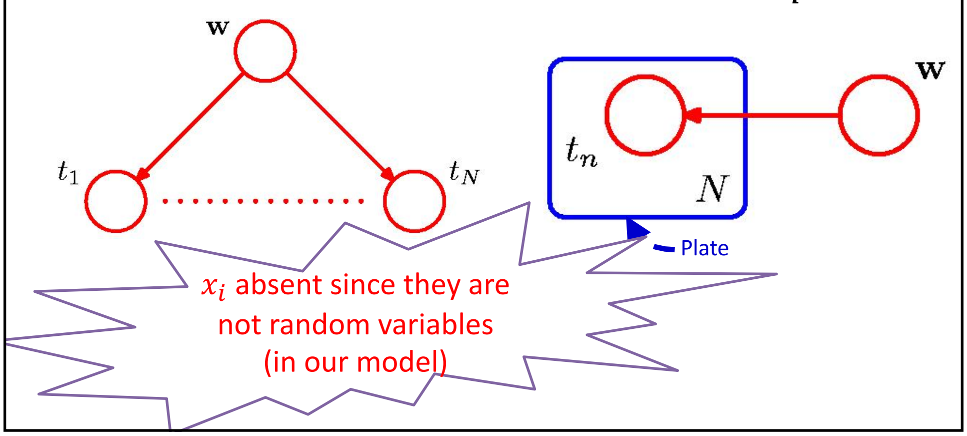
t is the true value.

w has a distribution.

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

Bayesian Curve Fitting (2)

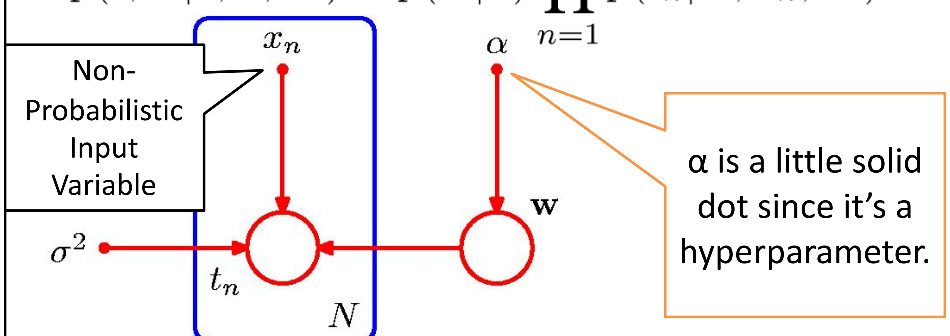
$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n | y(\mathbf{w}, x_n)) \quad y(\boldsymbol{w}, x_i) = \sum_{p=0}^{M} w_j x_i^p$$



Bayesian Curve Fitting (3)

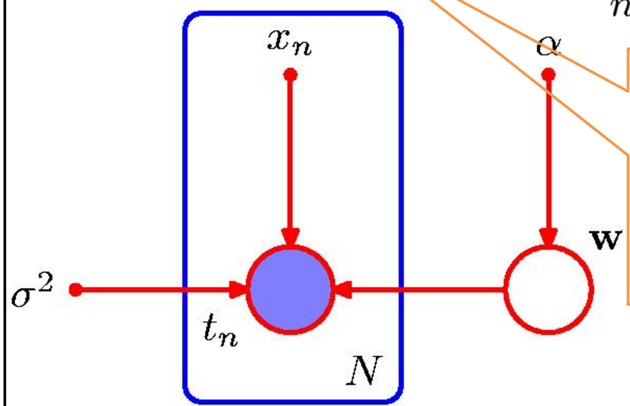
- Input variables and explicit hyperparameters

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



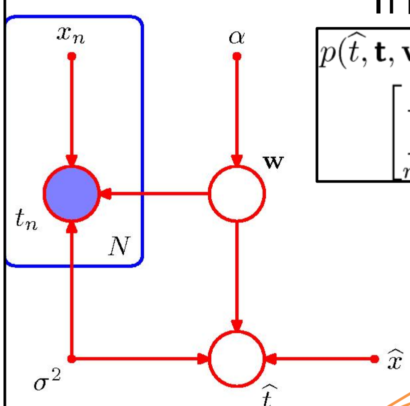
Bayesian Curve Fitting —Learning

- Condition on data $p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w})$



Probability distribution **(Not a point estimate)** of parameters given the data is called Posterior Distribution.

Bayesian Curve Fitting — Inference



$$p(\widehat{t}, \mathbf{t}, \mathbf{w}|\widehat{x}, \mathbf{x}, \alpha, \sigma^2) = \left[\prod_{n=1}^{N} p(t_n|x_n, \mathbf{w}, \sigma^2)\right] p(\mathbf{w}|\alpha) p(\widehat{t}|\widehat{x}, \mathbf{w}, \sigma^2)$$

Distribution of variable to predict given some data is called Predictive Distribution.

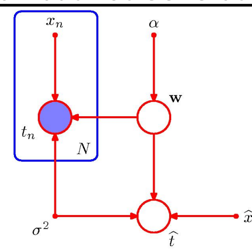
$$p(\widehat{t}|\widehat{x}, \mathbf{x}, \mathbf{t}, \alpha, \sigma^2) \propto \int p(\widehat{t}, \mathbf{t}, \mathbf{w}|\widehat{x}, \mathbf{x}, \alpha, \sigma^2)\, \mathrm{d}\mathbf{w}$$

$$\propto \int p(\hat{t} \mid \hat{x}, \mathbf{w}, \sigma^2)\; p(\mathbf{w} \mid \mathbf{t}, \mathbf{x}, \alpha, \sigma^2) d\mathbf{w}$$

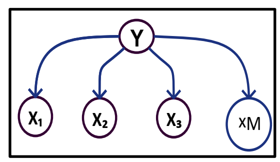
Know Your Jargon
Generative vs Discriminative vs Bayesian

- Bayesian Model – Puts a probability distribution on the parameters of the model. May or may not be generative.
- Generative Model – A probabilistic model that is capable of generating the data that it is modelling.
- Discriminative Model – A model that specifies the distribution of output variables as a function of the input variables. May or may not be Bayesian.

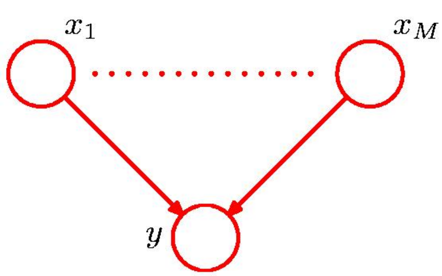
Bayesian but not Generative



Naïve Bayes – Generative (but not necessarily Bayesian)



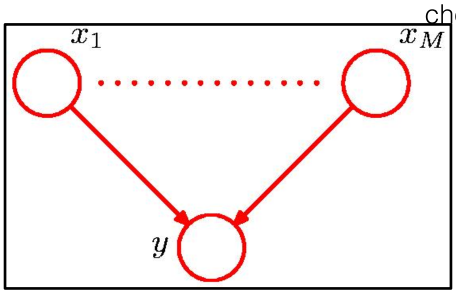
Parameterized Conditional Distributions
(Discriminative – May or May not be Bayesian)

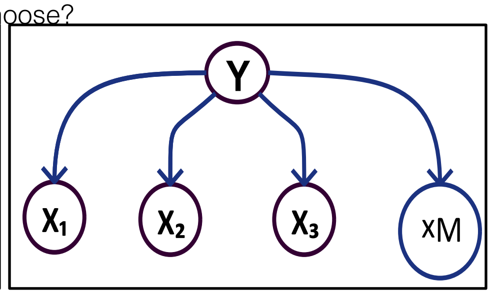


If $x_1, \ldots, x_M$ are discrete K-state variables, $p(y = 1 \mid x_1, \ldots, x_M)$ in general has $O(K^M)$ parameters. OTOH, The linear parameterized form requires only $M + 1$ parameters

$$p(y = 1|x_1, \ldots, x_M) = \sigma\left(w_0 + \sum_{i=1}^{M} w_i x_i\right) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x})$$

Regression) vs Generative (Naïve Bayes): How to choose?





On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, A. Y. Ng and M. I. Jordan, NIPS (2002)

tl;dr

1. Big data – Choose logistic regression
2. Small data – NB can outperform logistic regression.

Semantics of Graphical Models – Algorithms for Graphical Models

Bayesian Networks – DAGs of RVs – are a language for specifying sets of probability distributions.

- What properties do these sets of distributions have? Conditional Independence, D-Separation, Markov Blanket

- What other specifications exist for specifying sets of probability distributions? MRF, Factor Graphs
  - Can BN specify all possible joint distributions? No
  - Is there some formalism that is more expressive? Yes
    - How to convert a BN to this general form?

We saw how to do inference in a specific BN.

- ∃ general algorithm for inference in arbitrary BN and Factor Graphs? Yes, It's called Belief Propagation

Summary

- There are 3 dominant languages for designing probability distributions over interdependent RVs.
  - Bayesian Networks
  - Markov Random Fields
  - Factor Graph (General, Contains the above.)
- Together these methods of specifying probability distributions are called *Probabilistic Graphical Models*
- *Belief Propagation (BP)* is a general algorithm for doing inference in instances of *a useful subset* of PGMs.
  - Inference means finding the probability of a event.
- To understand BP and PGMs we need to know about
  - Conditional Independence
  - D-Separation
  - Markov Blankets

Conditional Independence

- If a is independent of b given c, then

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

- Equivalently

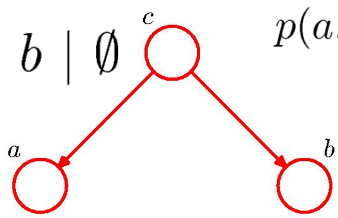
$$\boxed{\begin{aligned} p(a,b|c) &= p(a|b,c)p(b|c) \\ &= p(a|c)p(b|c) \end{aligned}}$$

- Notation

$$\boxed{a \perp\!\!\!\perp b \mid c}$$

Conditional Independence: Example 1

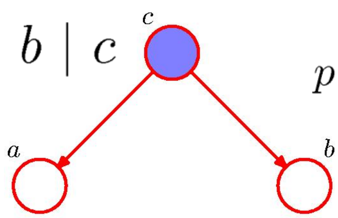
$a \not\perp\!\!\!\perp b \mid \emptyset$



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

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

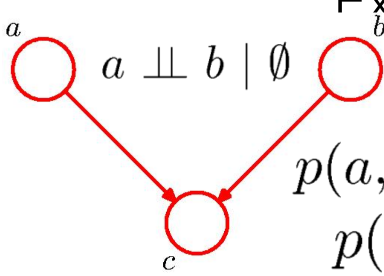
$a \perp\!\!\!\perp b \mid c$



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

Conditional Independence: Example 2

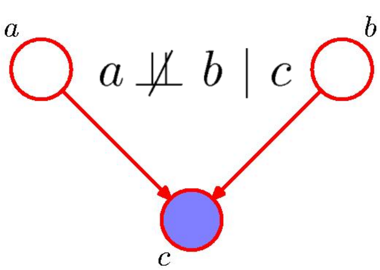
Note: this is the opposite of Example 1, with c observed.



$a \perp\!\!\!\perp b \mid \emptyset$

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

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

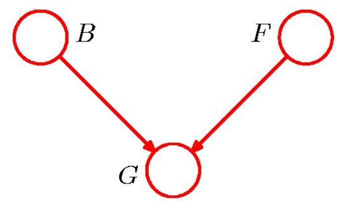


$a \not\perp\!\!\!\perp b \mid c$

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

Conditional Independence: Example 2
Inferring whether a car is out of fuel

$$
\begin{aligned}
p(G=1|B=1,F=1) &= 0.8 \\
p(G=1|B=1,F=0) &= 0.2 \\
p(G=1|B=0,F=1) &= 0.2 \\
p(G=1|B=0,F=0) &= 0.1
\end{aligned}
$$



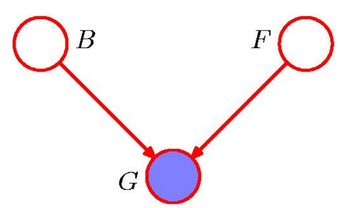
$$
\begin{aligned}
p(B=1) &= 0.9 \\
p(F=1) &= 0.9
\end{aligned}
$$

and hence

$$p(F=0) = 0.1$$

B = Battery (0=flat, 1=fully charged)
F = Fuel Tank (0=empty, 1=full)
G = Fuel Gauge Reading (0=empty, 1=full)

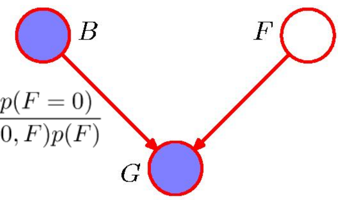
"Am I out of fuel?"



$$
\begin{aligned}
p(F=0|G=0) &= \frac{p(G=0|F=0)p(F=0)}{p(G=0)} \\
&\simeq 0.257
\end{aligned}
$$

Probability of an empty tank increased by observing G = 0

"Am I out of fuel?"



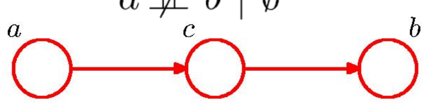
$$p(F=0|G=0,B=0) = \frac{p(G=0|B=0,F=0)p(F=0)}{\sum_{F\in\{0,1\}} p(G=0|B=0,F)p(F)} \simeq 0.111$$

Probability of an empty tank reduced by observing $\mathrm{B} = 0$. This referred to as "explaining away".

More generally, in a directed PGM, a child or any other ancestor of a child can influence the computation of the probability of a random variable.

Conditional Independence: Example 3

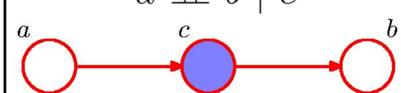
$$a \not\!\perp\!\!\!\perp b \mid \emptyset$$



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

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

$$a \perp\!\!\!\perp b \mid c$$

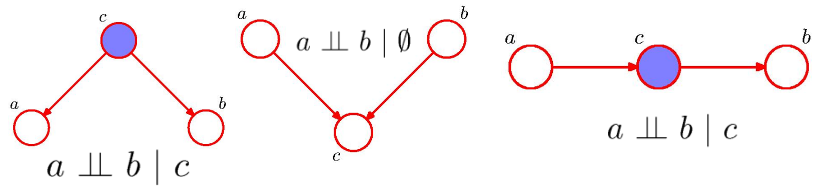


$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(c|a)p(b|c)}{p(c)} = p(a|c)p(b|c)$$

D-separation (In Directed PGMs – aka BayesNets)

- Let A, B, and C be disjoint subsets of nodes in a directed graph.
- A path from A to B is blocked by C, if it passes through a (vertex, edge pair) combination blocked by C.
- A vertex, V, edge pair (e_1 , e_2) is blocked by C, if, either
  a) $V \in C$ and (e_1 , e_2) meet either head-to-tail, tail-to-head, or tail-to-tail at V, OR
  b) (e_1 , e_2) meet head-to-head at V and $\left(V \notin C \text{ and any descendant}(V) \notin C\right)$

❑ Homework: Prove that if C blocks $A \to B$ then C blocks $B \to A$



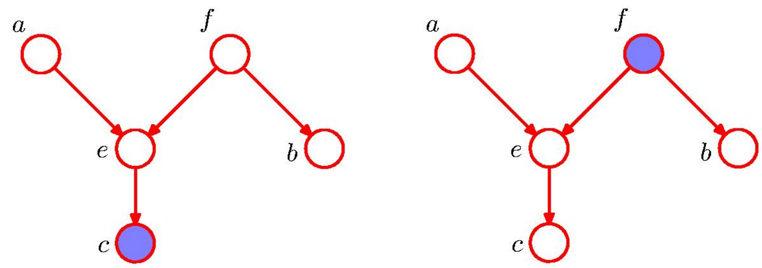
D-separation (In Directed PGMs – aka BayesNets)

- Let A, B, and C be disjoint subsets of nodes in a directed graph.
- A path from A to B is blocked by C, if it passes through a (vertex, edge pair) combination blocked by C.
- A vertex, V, edge pair (e_1 , e_2) is blocked by C, if, either
  a) $V \in C$ and (e_1 , e_2) meet either head-to-tail, tail-to-head, or tail-to-tail at V, OR
  b) (e_1 , e_2) meet head-to-head at V and $\left(V \notin C \text{ and any descendant}(V) \notin C\right)$

❑ Homework: Prove that if C blocks $A \to B$ then C blocks $B \to A$

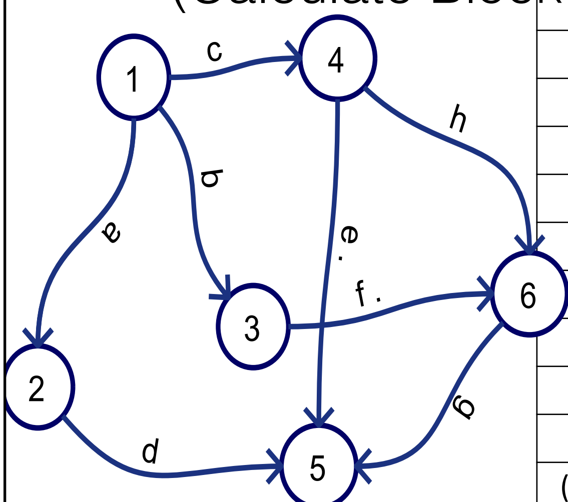
- If all paths from A to B are blocked, A is said to be d-separated from B by C. If A is d-separated from B by C, the joint distribution over all variables in the graph satisfies $A \perp\!\!\!\perp B \mid C$

D-separation: Example



$a \not\perp\!\!\!\perp b \mid c$ $\qquad$ $a \perp\!\!\!\perp b \mid f$

D-Separation: Exercise
(Homework)
(Calculate Block and Sep)



V	(e_1 , e_2)	C	Block
4	(c, e)	2	No
6	(f, h)	5	No
5	(d, e)	ϕ	Yes
6	(d, g)	2	Invalid
5	(d, g)	(2,5)	No
3	(b,f)	3	Yes
A	**B**	**C**	**D-Sep**
(1)	(5,6)	(2,34)	Yes
(1,2)	6	(4,3,5)	No

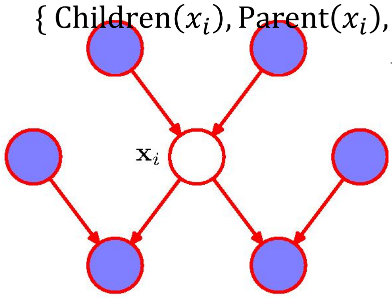
The Markov Blanket of a random variable

Let $\mathcal{X}$ be the set of all random variables.

Markov Blanket of variable x_i is the smallest subset $\mathcal{S} \subseteq \mathcal{X}$ such that $x_i \perp\!\!\!\perp (\mathcal{X} \setminus \mathcal{S}) \mid \mathcal{S}$

The Markov Blanket of a random variable in a directed graphical model

Let $\mathcal{X}$ be the set of all random variables in a directed PGM $\mathcal{G}$.
Markov Blanket of x_i =
$\{\,\mathrm{Children}(x_i), \mathrm{Parent}(x_i), \mathrm{Co{-}Parent}(x_i)\}$



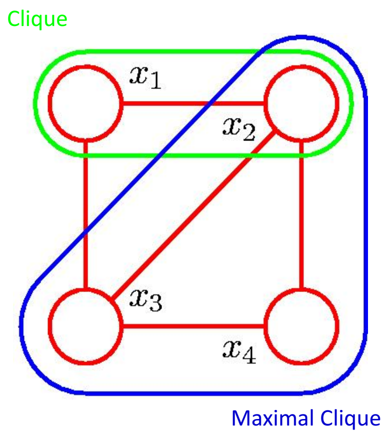
$$p(\mathbf{x}_i|\mathbf{x}_{\{j\neq i\}}) = \frac{p(\mathbf{x}_1,\ldots,\mathbf{x}_M)}{\int p(\mathbf{x}_1,\ldots,\mathbf{x}_M)\,\mathrm{d}\mathbf{x}_i} = \frac{\prod_k p(\mathbf{x}_k|\mathrm{pa}_k)}{\int \prod_k p(\mathbf{x}_k|\mathrm{pa}_k)\,\mathrm{d}\mathbf{x}_i}$$

Brute Force Proof: Factors independent of x_i cancel between numerator and denominator.

Simpler Proof: x_i is D-separated from every other variable given its children, Parents, Co-Parents.

Graph Theory Termino[...] Cliques and Maximal C[...]

Needed for Next Topic: Undirected Probabilistic Models, Markov Random Fields

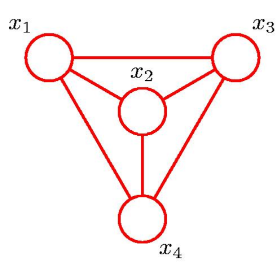


Markov Random Field

Definition: A MRF is an undirected graph of random variables whose joint probability factorizes according to the ***maximal cliques*** in the graph.

$$p(\boldsymbol{x}) = \frac{1}{Z}\prod_{C}\psi_c(x_c)$$

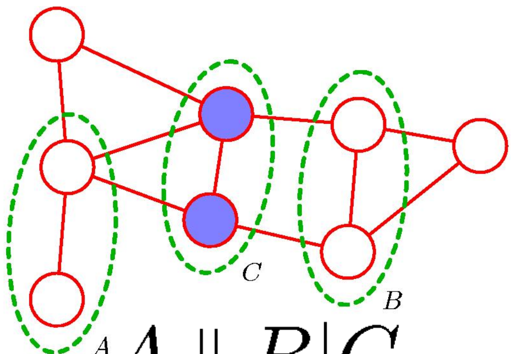
where $\psi_c(x_c)$ is the potential over clique C and Z is the normalization constant



$$Z = \sum_{x}\prod_{C}\psi_C(x_C)$$

D-Separation and Markov Blanket in Markov Random Fields

- A path from A to B is blocked by C, if it passes through a vertex that lies in C
- A is D-Separated from B if all paths between A and B are blocked by C
- The Markov Blanket of variable x_i is simply its set of neighbors.



$$A \perp\!\!\!\perp B \mid C$$