Till now - probabilistic models for data — Dirichlet allocat^n

Model in words — Regression

— Clustering

ORIE 4742 - Info Theory and Bayesian ML

Bayesian Networks

April 9, 2020

Sid Banerjee, ORIE, Cornell

From - Ch8, PRML
by Chris Bishop

probabilistic graphical models

graphical representation of complex probability distributions

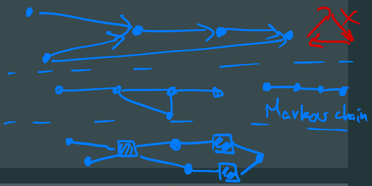
types of graphical models

• ≡ random variable

BayesNets: directed acyclic graphs

Markov random fields: undirected graphs

factor graphs: bipartite graphs

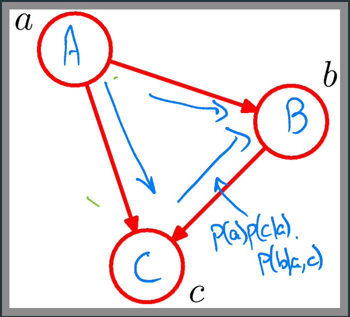


why are they useful?

▨ ≡ functions

- visualizing helps in design of probabilistic models
- complex inference/learning calculations $\rightarrow$ simpler graph operations
- gives insight into properties of model: conditional independence, causal relationships

BayesNets

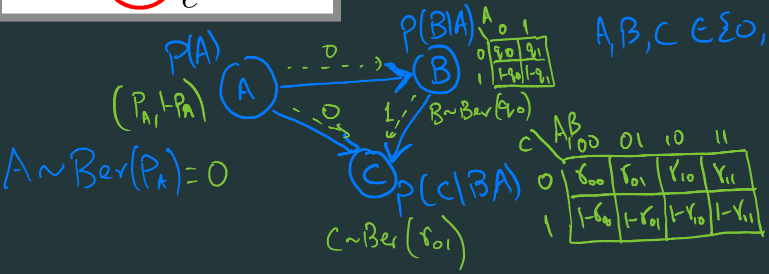


directed acyclic graph (DAG) encoding conditional distributions

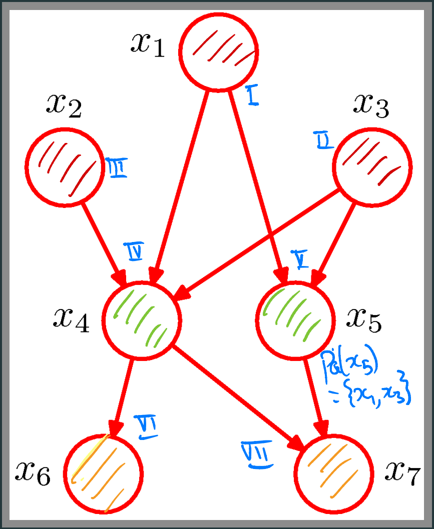
eg. for r.v.s A, B, C, BN on left encodes:

$$\begin{aligned} p(A,B,C) &= p(C|A,B)\,p(\underbrace{A,B}) \\ &= p(c|A,B)\,p(B|A)\,p(A) \end{aligned}$$

$A, B, C \in \{0,1\}$



BayesNets



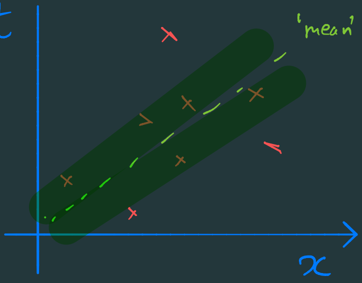
$$P(x_1, x_2, \dots, x_7) = P(x_1) \cdot P(x_3) \cdot P(x_2)\, P(x_4 \mid x_1, x_2, x_3) \cdot P(x_5 \mid x_1, x_3)\, P(x_6 \mid x_4)\, P(x_7 \mid x_5, x_4)$$

- Any DAG has a 'topological ordering' (ie, numbering s.t. no edge from higher to lower number) - use to generate prob expansion / factorization ↓

- For any x, $Pa(x) \stackrel{\text{def}}{=}$ 'parents' of x • $P\left(\bigcup_{i=1} x_i\right) = \prod_{i=1} P(x_i \mid Pa(x_i))$

example: (Bayesian) regression

Input – $(x_1,t_1),(x_2,t_2),\dots,(x_n,t_n)$



Tasks – 1) $t_i = \sum_{j=1}^{m} w_j f_j(x_i) + w_0$

basisfns noise

- $w_i \sim \mathcal{N}(\mu_i, 1/\tau_i)$
- $\varepsilon_i \sim \mathcal{N}(0, 1/\tau_\varepsilon)$

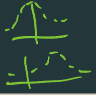
Want to learn $(w_1,w_2,\dots,w_m)$ from data

2) Given new point x_{n+1}, predict/infer t_{n+1}

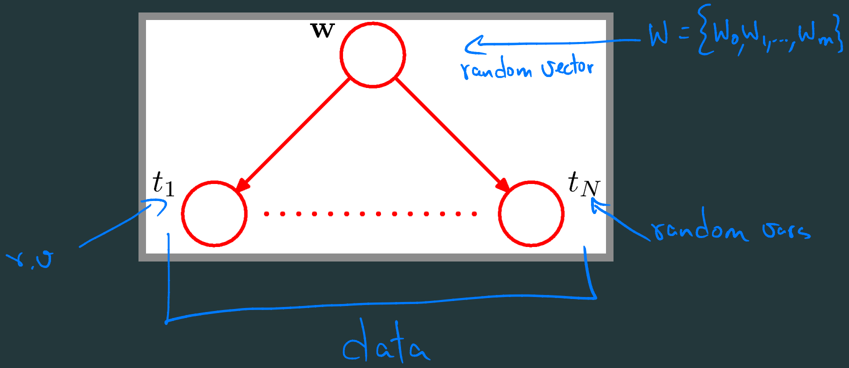
Eg – $f_1(x)=1$ (constant), $f_2(x)=x$ (linear regression)

– $f_k(x) = x^{k-1}$ (polynomial regression)

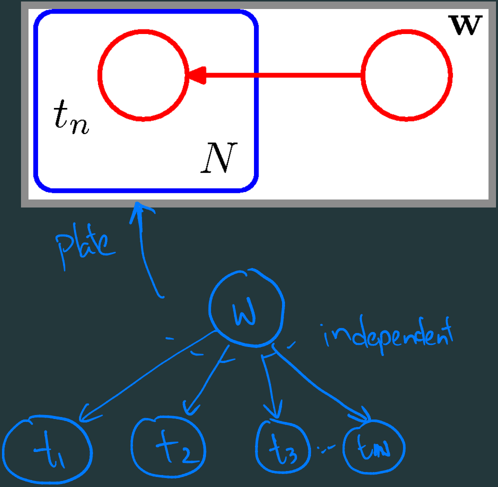
– $f_k(x) = e^{-(x-\mu)^2/2}$ (Gaussian basis fn)



regression: basic BayesNet



regression: plate notation

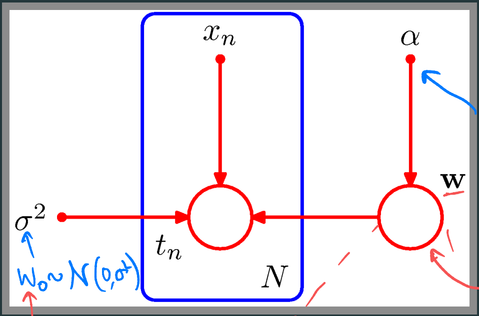


regression: inputs and hyperparameters

hyperparams are represented as solid dots (true for any 'deterministic variable')

Prior $\{w_i \sim \mathcal{N}(\mu_i, 1/\tau_i)\}$

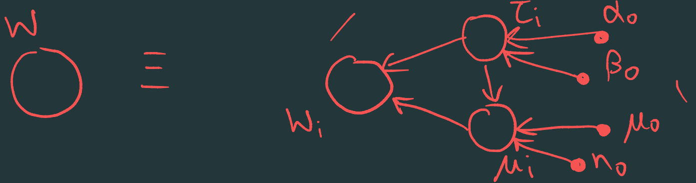
$w \quad \{\mu_1, \tau_1, \ldots, \mu_m, \tau_m\}$



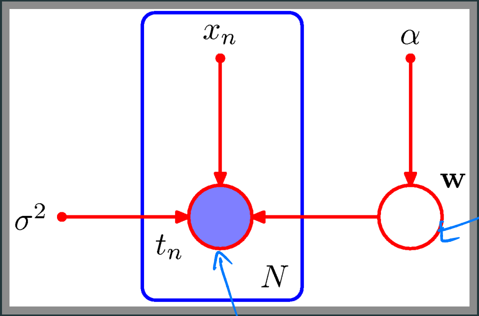
$w_0 \sim \mathcal{N}(0, \sigma^2)$

(want to learn) model params

'nuisance' parameter (do not want to learn)



regression: learning



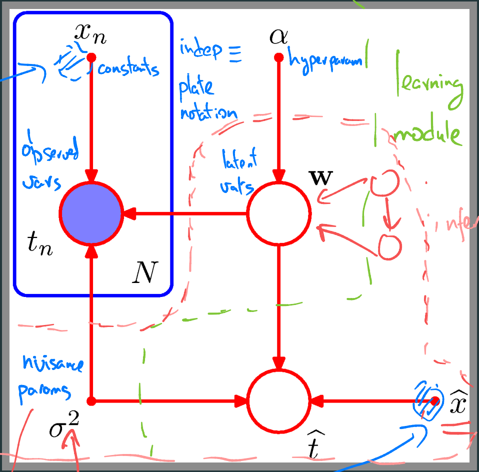
unshaded node
≡ latent variable

Shaded node
≡ 'observed variable'

regression: prediction

$(x_1,t_1),\dots,(x_N,t_N)$

$t_i = \sum_{j=1}^{M} w_j f_j(x_i) + w_0$



If x_i were random

τ_0 w_0

• Note $x_1, x_2, \dots, x_N$, and $\hat{x} \in \mathbb{R}^d$

• However x_i s are not r.v.

$\hat{t} \perp\!\!\!\perp t_i \mid w$

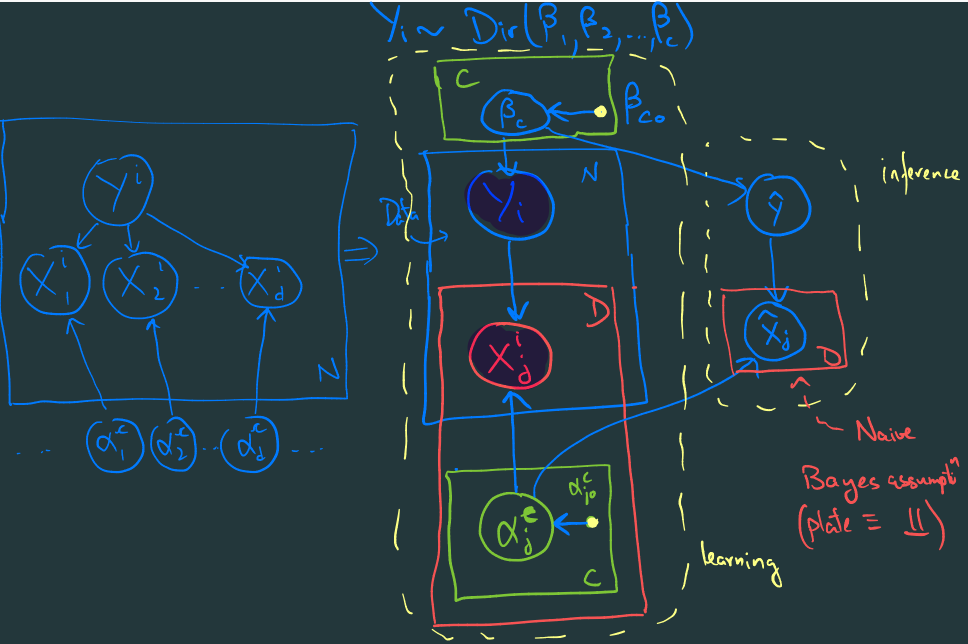
example: naive Bayes

classes - $c \in \{1, 2, \ldots, C\}$, data $(y^i, X_1^i, X_2^i, \ldots, X_d^i)$

dictionary - $d \in \{1, 2, \ldots, D\}$, $i \in \{1, 2, \ldots, N\}$

Assumption - $\underline{X_j^i \perp\!\!\!\perp X_{j'}^i}$ $\forall\ i$, Eg - $(X_1^i, X_2^i, \ldots, X_d^i) \sim \mathrm{Dir}(\{\alpha_j^i\})$

example: naive Bayes

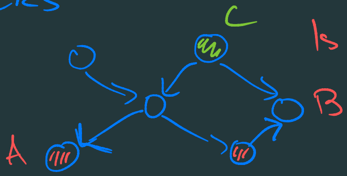


conditional independence

- Use a given Bayes Net to answer is. $A \perp\!\!\!\perp B \mid C$
  - $(A_1, A_2) \perp\!\!\!\perp (B_1, B_2, B_3) \mid (C_1, C_2, C_3)$
- $P(A, B \mid C) \left(\overset{?}{=} \atop \neq \right) P(A \mid C) P(B \mid C)$

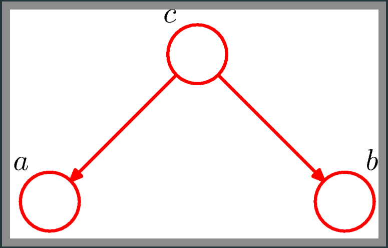
conditional independence

- TLDR – You can answer this given a Bayes Net
  - $\boxed{\text{d-separation}}$ (Pearl '88)
  - 3 building blocks
- Question



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

conditional independence: splits



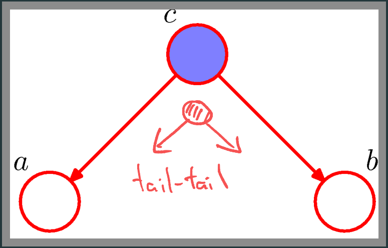
$$p(a,b,c) = p(c) \cdot p(a|c) \cdot p(b|c)$$

without conditioning

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

$$\Rightarrow \boxed{a \not\perp\!\!\!\perp b}$$

conditional independence: splits

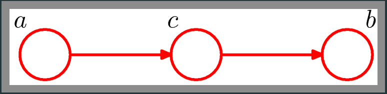


$$p(a,b|c) = \frac{p(a,b,c)}{p(c)}$$

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

$$= p(a|c).p(b|c) \Rightarrow \boxed{a \perp\!\!\!\perp b \mid c}$$

conditional independence: chains

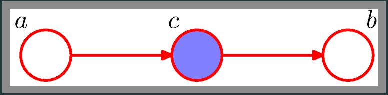


$$p(a,b,c) = p(a) \cdot p(c|a) \cdot p(b|c)$$

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

$$\Rightarrow \boxed{a \not\perp\!\!\!\perp b} \quad (\text{Eg. } b=c,\ c=a)$$

conditional independence: chains



head-tail

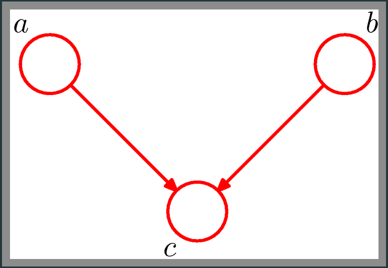
$$p(a,b|c) = \frac{p(a)\,p(c|a)\,p(b|c)}{p(c)}$$

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

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

(Markov chain)

conditional independence: joins

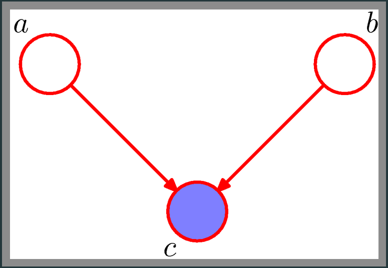


$$p(a,b,c) = p(a) \cdot p(b) \cdot p(c|a,b)$$

$$\Rightarrow p(a,b) = p(a) \cdot p(b) \cdot \underbrace{\sum_c p(c|a,b)}_{=1}$$

$$\Rightarrow \boxed{a \perp\!\!\!\perp b}$$

conditional independence: joins ('explaining away')

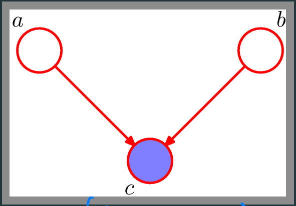


$$p(a,b|c) = \frac{p(a)\,p(b)\,p(c|a,b)}{p(c)}$$

$$\neq p(a|c)\cdot p(b|c)$$

$$\Rightarrow \quad a \not\perp\!\!\!\perp b \,|\, c$$

'explaining away'



(from Bishop)

C \ AB	00	01	10	11
0	0.9	0.8	0.8	0.2
1	0.1	0.2	0.2	0.8

Eg - $C = \mathbb{1}\{$burglar alarm rang$\}$ (fever + cough)

$A = \mathbb{1}\{$there is a burglar$\}$ (allergy)

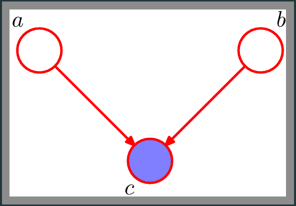
$B = \mathbb{1}\{$there is a raccoon$\}$ (COVID-19)

$\mathbb{P}[A=1] = 0.9, \ \mathbb{P}[B=1] = 0.9$

- $\mathbb{P}[B=0 \mid C=0] = \dfrac{\mathbb{P}[C=0 \mid B=0]\,\mathbb{P}[B=0]}{\mathbb{P}[C=0]} \approx 0.25$

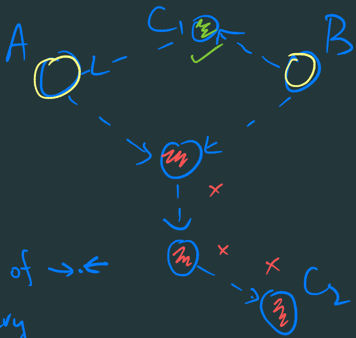
$\mathbb{P}[B=0 \mid C=0, A=0] = \dfrac{\mathbb{P}[C=0 \mid B=0, A=0]\,\mathbb{P}[B=0]}{\sum_{0,1} \mathbb{P}[A=i]\,\mathbb{P}[C=0 \mid A=i, B=0]} \approx 0.11$

‘explaining away’



d-separation

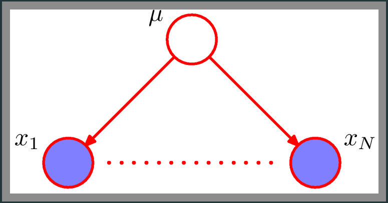
- $A \perp\!\!\!\perp B \mid C$ if C is not a join or a descendant of a join



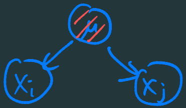
- A path from $A \to B$ is blocked by C if
  i) $\to \overset{c}{\bullet} \to$ or $\leftarrow \overset{c}{\bullet} \to$
  ii) C is not $\to \cdot \leftarrow$ or a descendant of $\to \cdot \leftarrow$

A, B are d-separated by C if every path $A \to B$ is blocked by C

d-separation: i.i.d. data

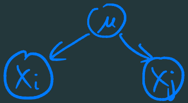


- $X_i \perp\!\!\!\perp X_j \mid \mu$



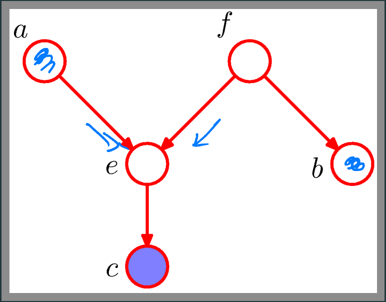
μ blocks path

- $X_i \not\perp\!\!\!\perp X_j$



μ does not block path

d-separation: example



Q: Is

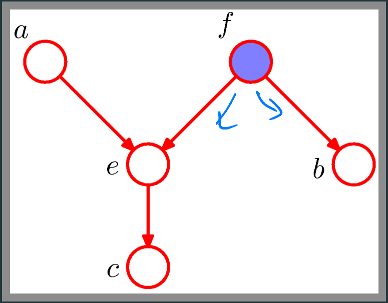
i) $A \perp\!\!\!\perp B$?

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

Ans

- $A \not\perp\!\!\!\perp B$
- $A \not\perp\!\!\!\perp B \mid C$

d-separation: example

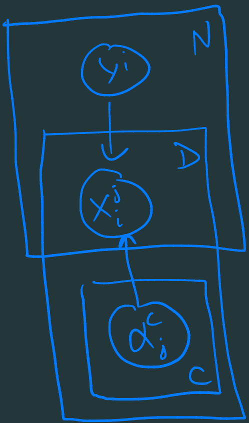


Q: Is

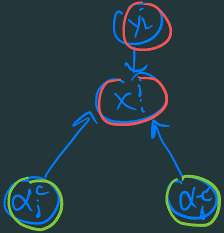
$A \perp\!\!\!\perp B \mid F$

Ans $A \perp\!\!\!\perp B \mid F$

d-separation: model parameters



$$(\alpha_1^c, \alpha_2^c \dots \alpha_d^c) \perp\!\!\!\perp (\alpha_1^{c'}, \alpha_2^{c'} \dots \alpha_d^{c'}) \mid \{y^i, x_1^j, \dots, x_n^j\}$$

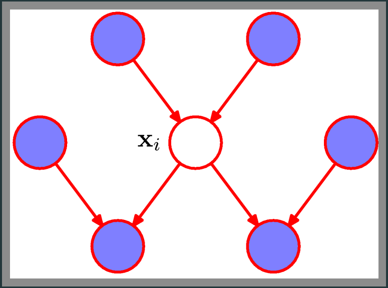


example: naive Bayes

example: naive Bayes

d-separation: model parameters

the Markov blanket



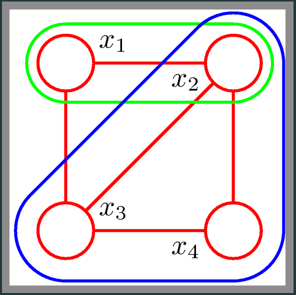
Markov random fields

- BayesNets encode 'local conditioning'

  $\prod_i P(x_i \mid Pa(x_i))$

- Don't directly capture global conditional indep/dep

cliques and potentials

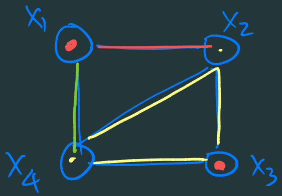


maximal cliques - set of nodes which form a clique, and are not subsets of a larger 'selected' clique

- Clique cover - collection of (maximal) cliques s.t. every edge is in a clique

- $$P(x_1, x_2, x_3, x_4) = \underbrace{\frac{1}{Z}}_{\text{normalization}} \cdot \prod_{\text{cliques } c} \underbrace{\psi_c(x_i \mid i \in c)}_{\text{clique potential}}$$

conditional independence and Markov blanket in MRF



$$C = \{ \{x_1, x_2\} \ \{x_1, x_4\} \ \{x_2 x_4 x_3\} \}$$

$$P(x_1, x_2, x_3, x_4) = \frac{1}{Z} \psi_{12}(x_1, x_2) \cdot \psi_{14}(x_1, x_4) \cdot \psi_{234}(x_2, x_3, x_4)$$

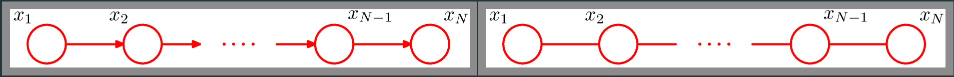
$\Longleftrightarrow$

- Conditional indep $\Longleftrightarrow$ separation

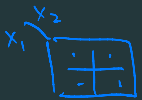
Q: $X_3 \perp\!\!\!\perp X_1 \mid X_4, X_2$?

Yes as (X_2, X_4) $\underbrace{\text{separate}}_{\text{disconnect}}$ X_1 and X_3

BayesNet vs MRF - Markov Chain



$$p(x) = p(x_1)\,p(x_2|x_1)\cdots p(x_N|x_{N-1})$$



$$p(x) = \frac{1}{Z}\,\psi_{12}(x_1,x_2)\,\psi_{23}(x_2,x_3)\cdots\psi_{N-1,N}(x_{N-1},x_N)$$

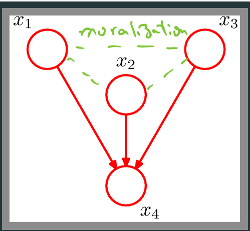
- Choose $\psi_{12}(x_1,x_2) = p(x_1)\,p(x_2|x_1)$

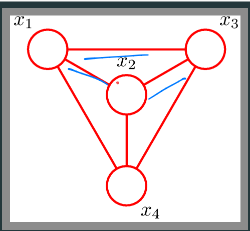
  $\psi_{23}(x_2,x_3) = p(x_3|x_2)$

  $\vdots$

  $\psi_{N-1,N}(x_{N-1},x_N) = p(x_N|x_{N-1})$

BayesNet vs MRF





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

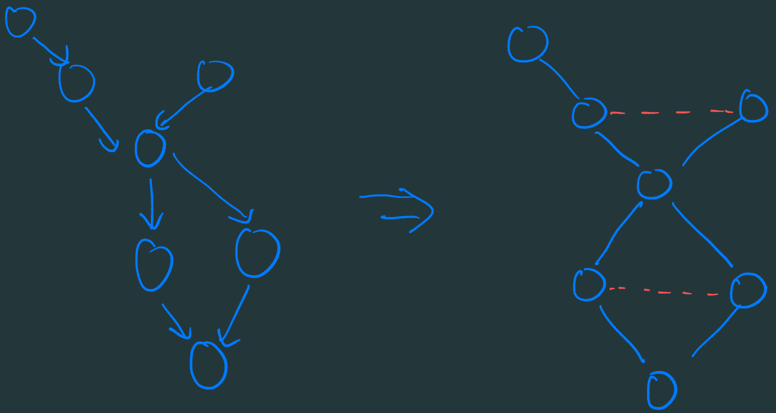
$$C = \{\{x_1\},\{x_2\},\{x_3\},\{x_1x_2x_3x_4\}\}$$

$$\psi_1(x_1) = p(x_1), \; \dots \; \psi_{1234}(x_1x_2x_3x_4) = p(x_4|x_1,x_2,x_3)$$

$$C = \{x_1x_2x_3\}\{x_1x_2x_4\}\{x_2x_3x_4\}$$

$$\Rightarrow \frac{1}{Z}\psi_{123}(x_1x_2x_3)\,\psi(x_1x_2x_4)\,\psi_{234}(x_2x_3x_4)$$

converting BayesNets to MRFs – 'moralization'

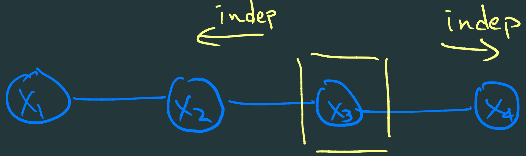


- Add edges between unconnected parents of each child node

d-separation and moralization

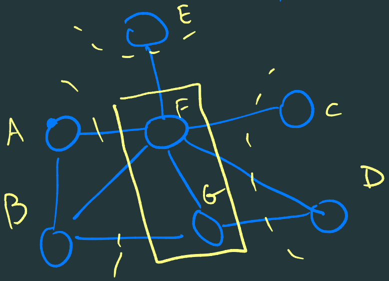
- Is $A \perp\!\!\!\perp B \mid C$?
  - Convert Bayes Net of 'ancestors of C' into MRF (via moralization)
  - Check for conditional independence

Eg. Markov chain



- $X_i \perp\!\!\!\perp X_j \mid X_k$ if $i < k < j$ or $i > k > j$

Eg

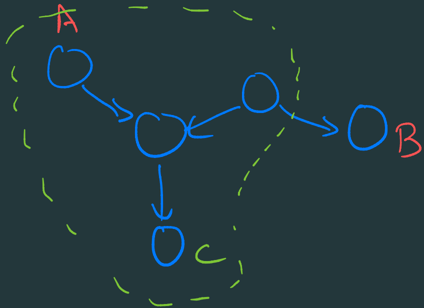


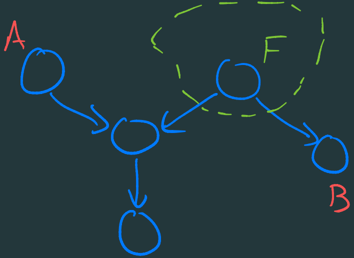
$A, B \perp\!\!\!\perp C, D \mid F, G$

$C \perp\!\!\!\perp D \mid F, G$

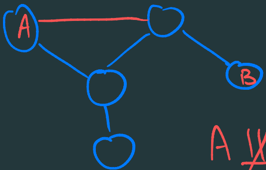
$A \not\!\perp\!\!\!\perp B \mid F, G$

$E \perp\!\!\!\perp C, D \mid F, G$



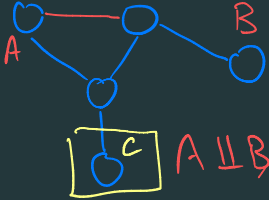


- Unconditional



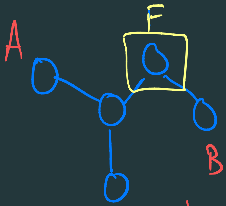
$A \not\!\perp\!\!\!\perp B$

- Condn on C



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

- Condn on F



$A \perp\!\!\!\perp B \mid F$