# Machine Learning

4771

Instructor: Tony Jebara

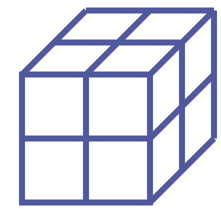
# Topic 14

- Structuring Probability Functions for Storage
- Structuring Probability Functions for Inference
- Basic Graphical Models
- Graphical Models
- Parameters as Nodes

# Structuring PDFs for Storage

- Probability tables quickly grow if p has many variables

$$p(x) = p\left(flu?, headache?, \ldots, temperature?\right)$$



1
- For D true/false "medical" variables $\underline{table\, size = 2^D}$
- Exponential blow-up of storage size for the probability
- Example: 8x8 binary images of digits
- If multinomial with M choices, probabilities are how big?
- As in Naïve Bayes or Multivariate Bernoulli, if words were independent things are much more efficient

$$p(x) = p\left(flu?\right) p\left(headache?\right) \ldots p\left(temperature?\right)$$

independent:

| | |
|---|---|
| 0.73 | 0.27 |

| | |
|---|---|
| 0.2 | 0.8 |

| | |
|---|---|
| 0.54 | 0.46 |

2
- For D true/false "medical" variables $\underline{table\, size = 2 \times D}$
  (really even less than that…)

# Structuring PDFs for Inference

- Inference: goal is to predict some variables given others

  x1: flu
  x2: fever
  x3: sinus infection
  x4: temperature
  x5: sinus swelling
  x6: headache

  Patient claims headache and high temperature. Does he have a flu?

  Given findings variables $X_f$ and unknown variables $X_u$ predict queried variables $X_q$

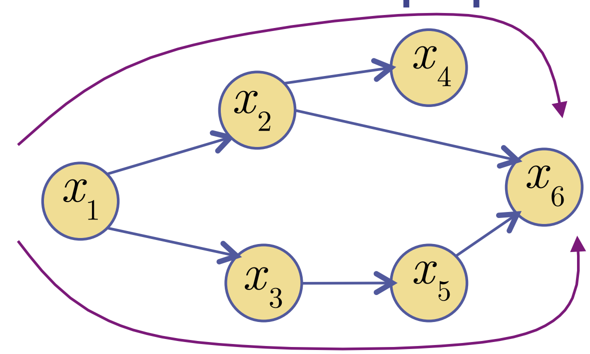
- Classical approach: truth tables (slow) or logic networks

- Modern approach: probability tables (slow) or Bayesian networks (fast belief propagation, junction tree algorithm)

# From Logic Nets to Bayes Nets

- 1980's expert systems & logic networks became popular

| x1 | x2 | x1 v x2 | x1^x2 | x1 -> x2 |
|---|---|---|---|---|
| T | T | T | T | T |
| T | F | T | F | F |
| F | T | T | F | T |
| F | F | F | F | T |




- Problem: inconsistency, 2 paths can give different answers

- Problem: rules are hard, instead use soft probability tables

$x3 = x1 \wedge x2$

x3=0

| | x2=0 | x2=1 |
|---|---|---|
| x1=0 | 1.0 | 1.0 |
| x1=1 | 1.0 | 0.0 |

x3=1

| | x2=0 | x2=1 |
|---|---|---|
| x1=0 | 0.0 | 0.0 |
| x1=1 | 0.0 | 1.0 |

$p(x3|x1,x2)$

$$\sum_{j=0}^{1} p(x_3=j \mid x_1, x_2) = 1$$

x3=0

| | x2=0 | x2=1 |
|---|---|---|
| x1=0 | 0.8 | 0.7 |
| x1=1 | 0.7 | 0.1 |

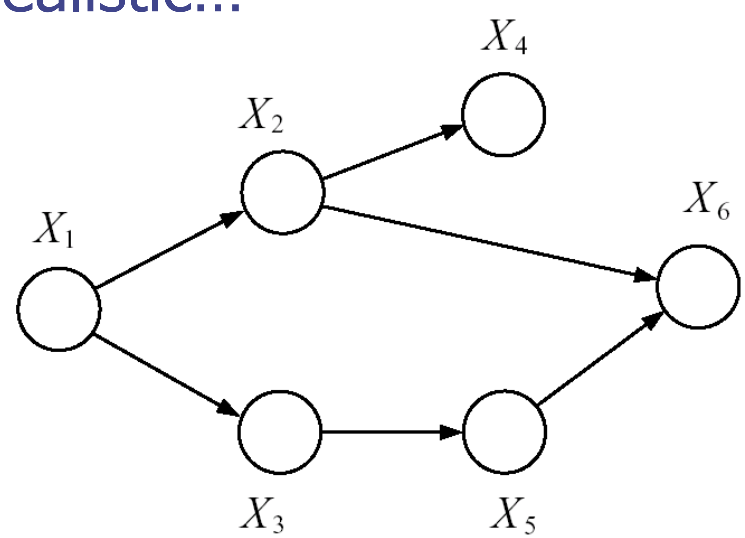
x3=1

| | x2=0 | x2=1 |
|---|---|---|
| x1=0 | 0.2 | 0.3 |
| x1=1 | 0.3 | 0.9 |

- These directed graphs are called Bayesian Networks

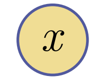
# Graphical Models & Bayes Nets

- Independence assumptions make probability tables smaller
- But real events in the world not completely independent!
- Complete independence is unrealistic…
- Graphical models use a graph to describe more subtle dependencies and independencies:
  …namely: conditional independencies
  (like causality but not exactly…)




- Directed Graphical Model, also called Bayesian Network use a directed acylic graph (DAG).
- Neural Network = Graphical Function Representation
- Bayesian Network = Graphical Probability Representation

# Graphical Models & Bayes Nets

- **Node:** a random variable (discrete or continuous) 

- **Independent:** no link 

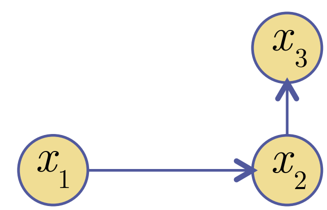
$p(x,y) = p(x)p(y)$
- **Dependent:** link 

$p(x,y) = p(y \mid x)p(x)$
- **Arrow:** from parent to child (like causality, not exactly)
- **Child:** destination of arrow, response
- **Parent:** root of arrow, trigger $parents\,of\,child\,i = pa_i = \pi_i$

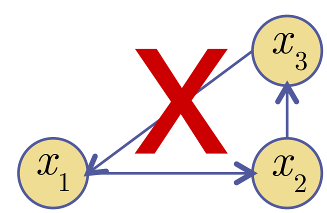
notation

- **Graph:** dependence/independence
- **Graph:** shows factorization of joint
  joint = products of conditionals

$$p\left(x_1,\ldots,x_n\right) = \prod_{i=1}^n p\left(x_i \mid pa_i\right) = \prod_{i=1}^n p\left(x_i \mid \pi_i\right)$$

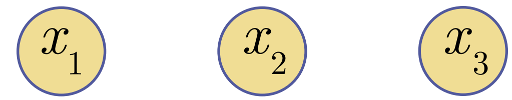



- **DAG:** directed acyclic graph



# Basic Graphical Models

- Independence: all nodes are unlinked

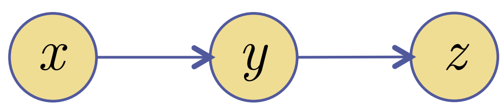


- Shading: variable is 'observed', condition on it moves to the right of the bar in the pdf

- Examples of simplest conditional independence situations…

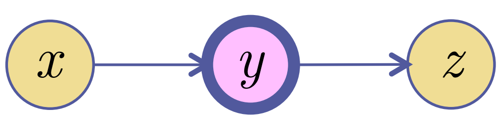
$$p(x_1,\ldots,x_n) = \prod_{i=1}^{n} p(x_i \mid pa_i) = \prod_{i=1}^{n} p(x_i \mid \pi_i)$$

1) Markov chain:



$$p(x,y,z) = p(x)\,p(y \mid x)\,p(z \mid y)$$

**Example binary events:**
**x = president says war**
**y = general orders attack**
**z = soldier shoots gun**



$$x \perp\!\!\!\perp z \mid y$$

$\perp\!\!\!\perp$: independent. observed.

$$p(x \mid y, z) = \frac{p(x,y,z)}{p(y,z)} = p(x \mid y) \times z$$

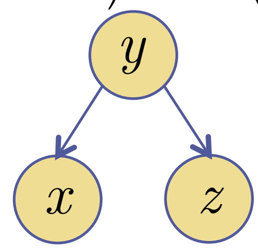
$p(x \mid y, z) = p(x \mid y)$

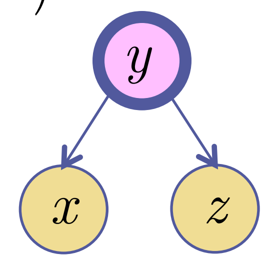
$$= \frac{p(\cancel{z \mid y}) \cdot p(y \mid x) \cdot p(x)}{p(\cancel{z \mid y}) \cdot p(y)} = \frac{p(x,y)}{p(y)}$$

# Basic Graphical Models

2) 1 Cause, 2 effects: $p(x,y,z) = p(y)\,p(x \mid y)\,p(z \mid y)$

**y = flu**
**x = sore throat**
**z = temperature**



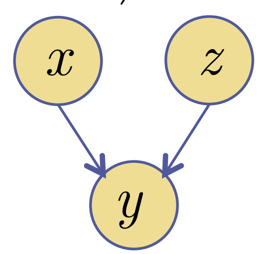


$x \perp\!\!\!\perp z \mid y$

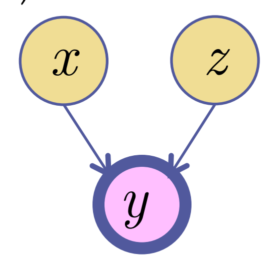
3) 2 Causes, 1 effect: $p(x,y,z) = p(x)\,p(z)\,p(y \mid x,z)$

**x = rain**
**y = wet driveway**
**z = car oil leak**

***Explaining away…***



$x \perp\!\!\!\perp z$



$x \not\perp\!\!\!\perp z \mid y$

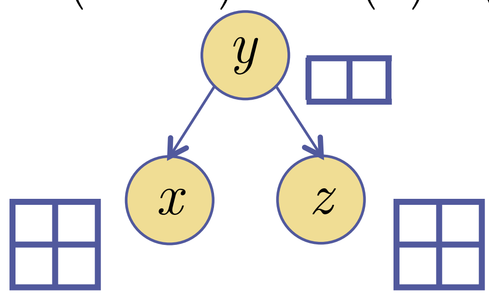
$p(x \mid y,z) = \frac{p(x,y,z)}{p(y,z)} = \frac{p(x)\,p(z)\,p(y \mid x,z)}{p(y \mid x,z)\cdot p(z)} = p(x)$

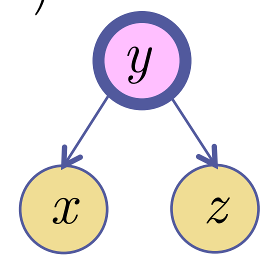
- Each conditional is a mini-table
(Multinomial or Bernoulli conditioned on parents)

# Basic Graphical Models

2) 1 Cause, 2 effects: $p(x,y,z) = p(y)\,p(x \mid y)\,p(z \mid y)$

**y = flu**
**x = sore throat**
**z = temperature**



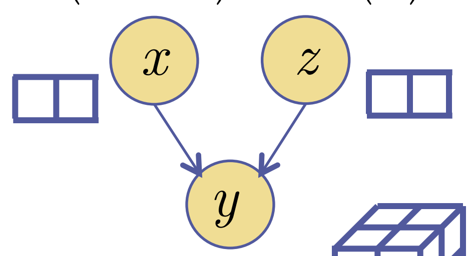


$x \perp\!\!\!\perp z \mid y$

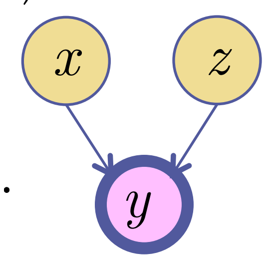
3) 2 Causes, 1 effect: $p(x,y,z) = p(x)\,p(z)\,p(y \mid x,z)$

**x = dad is diabetic**
**y = child is diabetic**
**z = mom is diabetic**

***Explaining away…***



$x \perp\!\!\!\perp z$



$x \cancel{\perp\!\!\!\perp} z \mid y$

following algorithm

•Each conditional is a mini-table
(Multinomial or Bernoulli conditioned on parents)

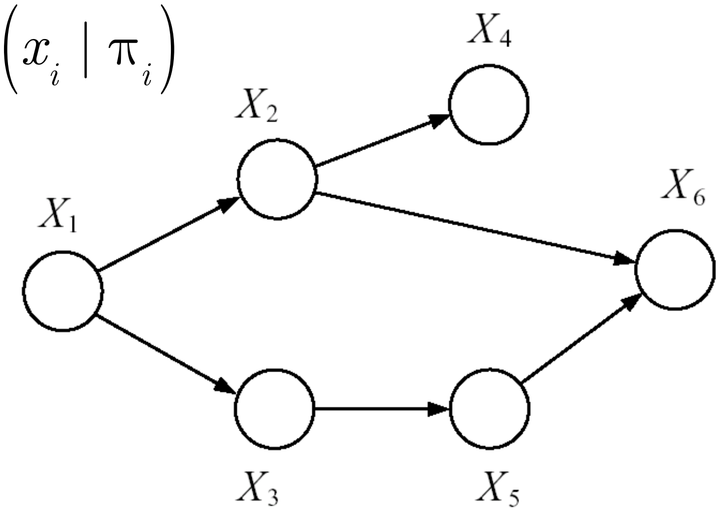
D-Separation. to tell independency

# Graphical Models

- Example: factorization of the following system of variables

$$p\left(x_1,\ldots,x_n\right) = \prod_{i=1}^{n} p\left(x_i \mid pa_i\right) = \prod_{i=1}^{n} p\left(x_i \mid \pi_i\right)$$

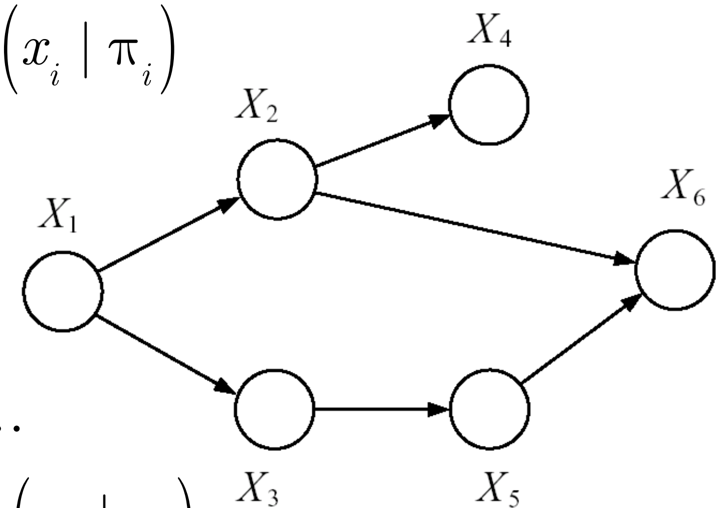
$$p\left(x_1,\ldots,x_6\right) = p\left(x_1\right)\ldots$$



# Graphical Models

- Example: factorization of the following system of variables

$$p(x_1,\ldots,x_n) = \prod_{i=1}^{n} p(x_i \mid pa_i) = \prod_{i=1}^{n} p(x_i \mid \pi_i)$$



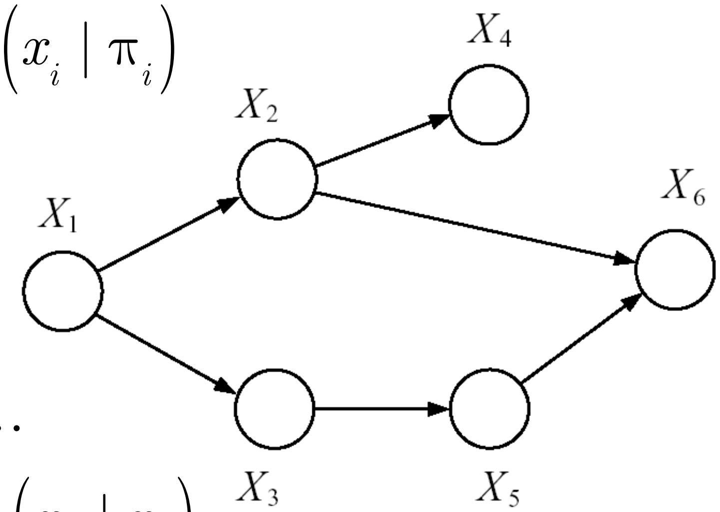

$$\begin{aligned} p(x_1,\ldots,x_6) &= p(x_1)\ldots \\ &= p(x_1)p(x_2 \mid x_1)\ldots \\ &= p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)\ldots \\ &= p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2)\ldots \\ &= p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2)p(x_5 \mid x_3)\ldots \\ &= p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2)p(x_5 \mid x_3)p(x_6 \mid x_2,x_5) \end{aligned}$$

- How big are these tables (if binary variables)?

# Graphical Models

- Example: factorization of the following system of variables

$$p\left(x_1,\ldots,x_n\right) = \prod_{i=1}^{n} p\left(x_i \mid pa_i\right) = \prod_{i=1}^{n} p\left(x_i \mid \pi_i\right)$$




$$\begin{aligned}
p\left(x_1,\ldots,x_6\right) &= p\left(x_1\right)\ldots \\
&= p\left(x_1\right)p\left(x_2 \mid x_1\right)\ldots \\
&= p\left(x_1\right)p\left(x_2 \mid x_1\right)p\left(x_3 \mid x_1\right)\ldots \\
&= p\left(x_1\right)p\left(x_2 \mid x_1\right)p\left(x_3 \mid x_1\right)p\left(x_4 \mid x_2\right)\ldots \\
&= p\left(x_1\right)p\left(x_2 \mid x_1\right)p\left(x_3 \mid x_1\right)p\left(x_4 \mid x_2\right)p\left(x_5 \mid x_3\right)\ldots \\
&= p\left(x_1\right)p\left(x_2 \mid x_1\right)p\left(x_3 \mid x_1\right)p\left(x_4 \mid x_2\right)p\left(x_5 \mid x_3\right)p\left(x_6 \mid x_2,x_5\right)
\end{aligned}$$

$$2^6 \qquad 2^1 \qquad 2^2 \qquad 2^2 \qquad 2^2 \qquad 2^2 \qquad 2^3$$

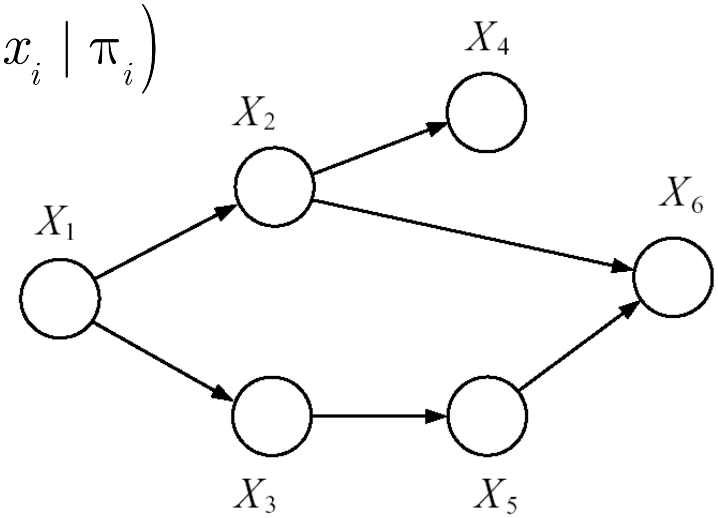
- How big are these tables (if binary variables)?

# Graphical Models

- Example: factorization of the following system of variables

$$p(x_1,\ldots,x_n) = \prod_{i=1}^{n} p(x_i \mid pa_i) = \prod_{i=1}^{n} p(x_i \mid \pi_i)$$

- Interpretation???



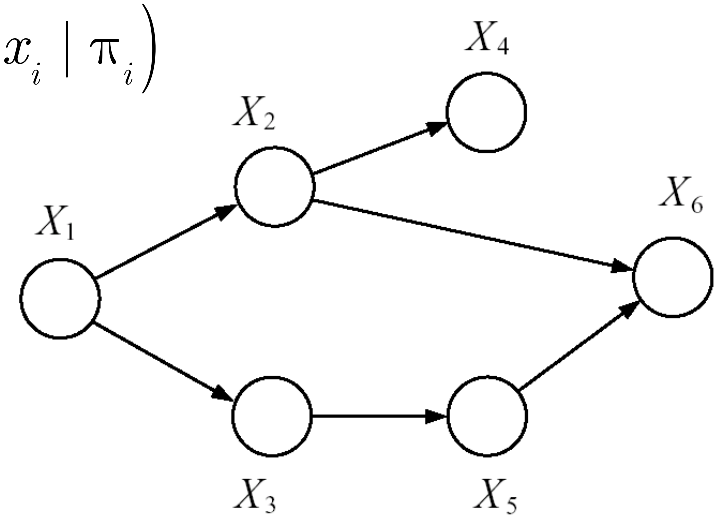

$$p(x_1,\ldots,x_6) = p(x_1)\,p(x_2 \mid x_1)\,p(x_3 \mid x_1)\,p(x_4 \mid x_2)\,p(x_5 \mid x_3)\,p(x_6 \mid x_2,x_5)$$

$$2^6 \qquad 2^1 \qquad 2^2 \qquad 2^2 \qquad 2^2 \qquad 2^2 \qquad 2^3$$

# Graphical Models

- Example: factorization of the following system of variables

$$p(x_1,\ldots,x_n) = \prod_{i=1}^{n} p(x_i \mid pa_i) = \prod_{i=1}^{n} p(x_i \mid \pi_i)$$




- Interpretation:
  1: flu
  2: fever
  3: sinus infection
  4: temperature
  5: sinus swelling
  6: headache

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

$$2^6 \Rightarrow 2^1 + 2^2 + 2^2 + 2^2 + 2^2 + 2^3$$

# Graphical Models

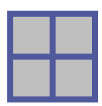
- Normalizing probability tables. Joint distributions sum to 1.
- BUT, conditionals sum to 1 for *each* setting of parents.

p(x) **2-1**

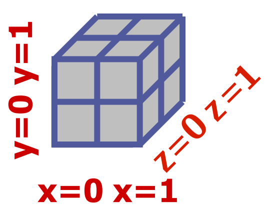
$$\sum_{x=0}^{1} p(x) = 1$$

p(x,y) **4-1**

$$\sum_{x,y} p(x,y) = 1$$



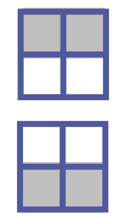
p(x,y,z) **8-1**




$$\sum_{x,y,z} p(x,y,z) = 1$$

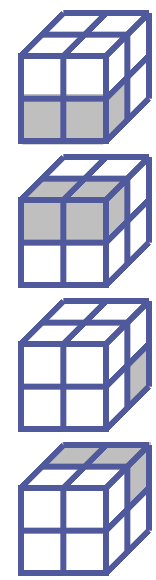
p(x|y) **4-2**

$$\sum_{x} p(x \mid y=0) = 1$$
$$\sum_{x} p(x \mid y=1) = 1$$



p(x|y,z) **8-4**

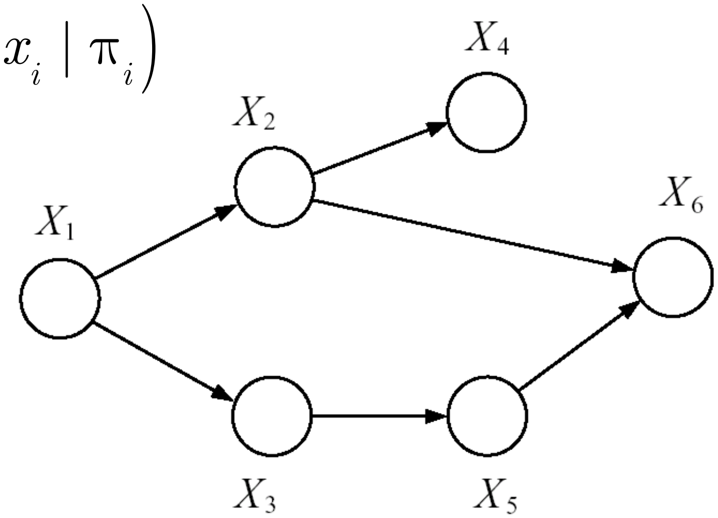
$$\sum_{x} p(x \mid y=0, z=0) = 1$$
$$\sum_{x} p(x \mid y=1, z=0) = 1$$
$$\sum_{x} p(x \mid y=0, z=1) = 1$$
$$\sum_{x} p(x \mid y=1, z=1) = 1$$



# Graphical Models

- Example: factorization of the following system of variables

$$p\left(x_1,\ldots,x_n\right)=\prod_{i=1}^{n}p\left(x_i\mid pa_i\right)=\prod_{i=1}^{n}p\left(x_i\mid \pi_i\right)$$




- Interpretation
  1: flu
  2: fever
  3: sinus infection
  4: temperature
  5: sinus swelling
  6: headache

$$p\left(x_1,\ldots,x_6\right)=p\left(x_1\right)p\left(x_2\mid x_1\right)p\left(x_3\mid x_1\right)p\left(x_4\mid x_2\right)p\left(x_5\mid x_3\right)p\left(x_6\mid x_2,x_5\right)$$

$$2^6-1 \rightarrow \quad 2^1-1 \quad 2^2-2 \quad 2^2-2 \quad 2^2-2 \quad 2^2-2 \quad 2^3-4$$

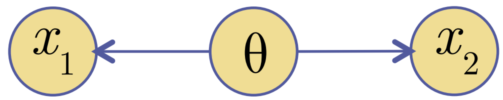
$$1 + 2 + 2 + 2 + 2 + 4$$

63 vs. 13 degrees of freedom

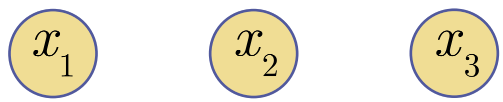
store them. Sum up for storage space.

# Parameters as Nodes

- Consider the model variable $\theta$ ALSO as a random variable

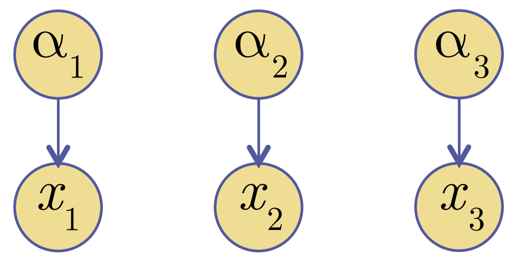



- But would need a prior distribution $P(\theta)$… ignore for now
- Recall: Naïve Bayes, word probabilities are independent



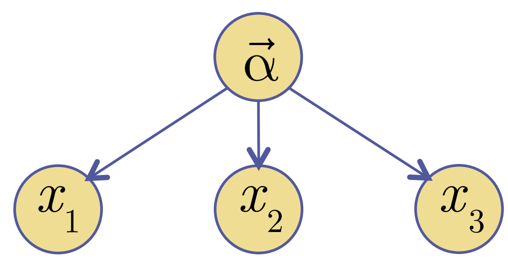

- Text: Multivariate Bernoulli

$$p\left(x \mid \vec{\alpha}\right) = \prod_{d=1}^{50000} \alpha_d^{x_d} \left(1 - \alpha_d\right)^{\left(1 - x_d\right)}$$




- Text: Multinomial

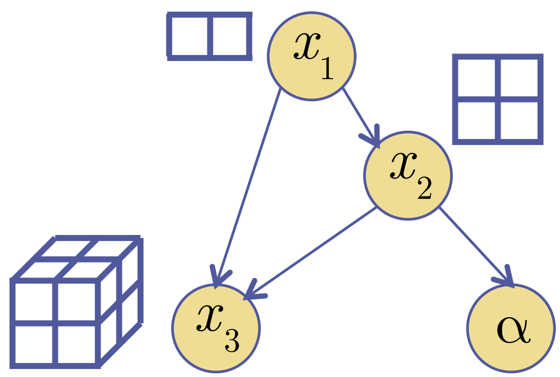
$$p\left(X \mid \vec{\alpha}\right) = \frac{\left(\sum_{m=1}^{M} X_m\right)!}{\prod_{m=1}^{M} X_m!} \prod_{m=1}^{M} \alpha_m^{X_m}$$

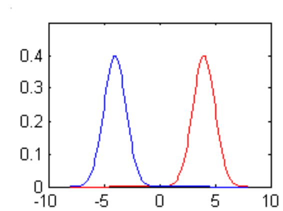


# Continuous Conditional Models

- In previous slide, $\theta$ and $\alpha$ were a random variable in graph
- But, $\theta$ and $\alpha$ are continuous
- Network can have both discrete & continuous nodes

- Joint factorizes into conditionals that are either:
  1) discrete conditional probability tables
  2) continuous conditional probability distributions




- Most popular continuous distribution = Gaussian

# Graphical Models

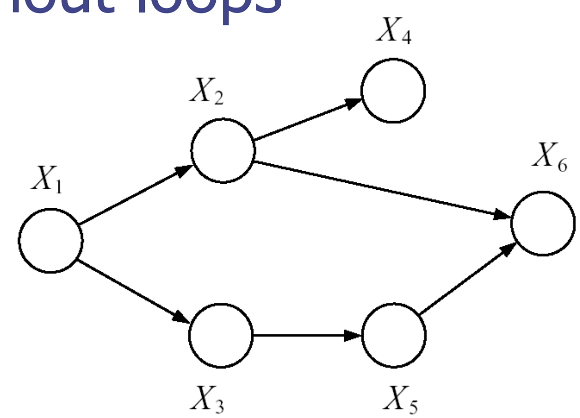
- In EM, we saw how to handle nodes that are: observed (shaded), hidden variables (E), parameters (M)
- But, only considered simple iid, single parent, structures
- More generally, have arbitrary DAG without loops
- Notation:

$$G = \{X, E\} = \{\texttt{nodes / randomvars,edges}\}$$

$$X = \{x_1, \ldots, x_M\}$$

$$E = \{(x_i, x_j) : i \neq j\}$$

$$X_c = \{x_1, x_3, x_4\} = subset$$

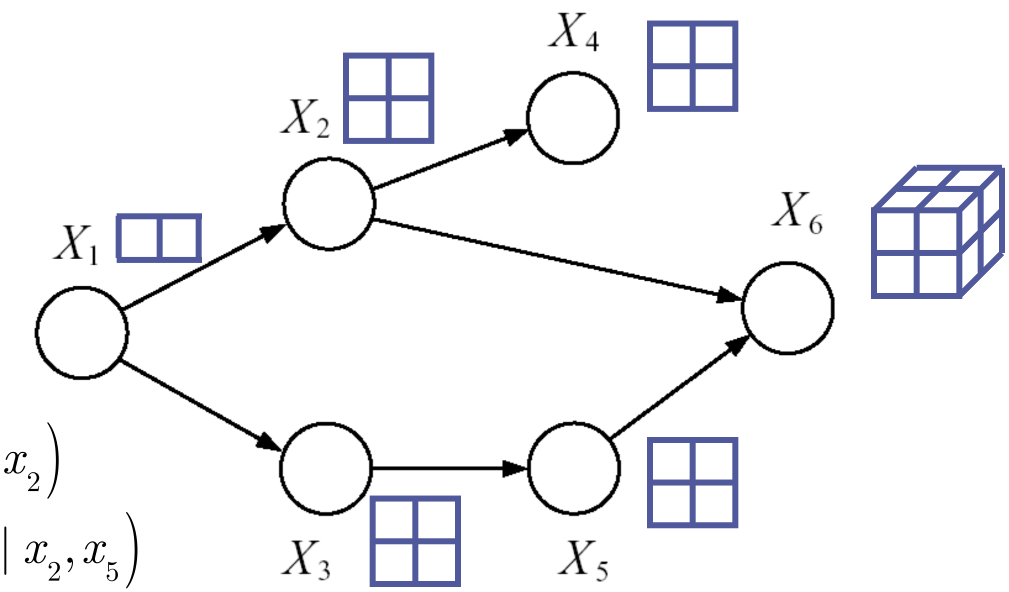



- Want to do 4 things with these graphical models:
  1) Learn Parameters (to fit to data)
  2) Query independence/dependence
  3) Perform Inference (get marginals/max a posteriori)
  4) Compute Likelihood (e.g. for classification)

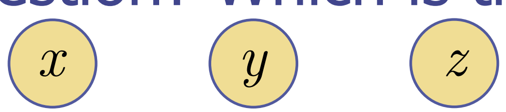
# Graphical Models

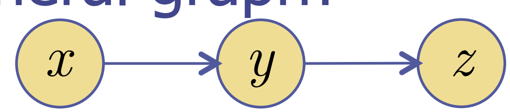
- Graph factorizes probability: $p\left(x_1,\ldots,x_n\right) = \prod_{i=1}^{n} p\left(x_i \mid \pi_i\right)$
- Topological graph: nodes are in order so that parents $\pi$ come before children

$$p\left(x_1,\ldots,x_6\right) = p\left(x_1\right)p\left(x_2 \mid x_1\right) \times p\left(x_3 \mid x_1\right)p\left(x_4 \mid x_2\right) \times p\left(x_5 \mid x_3\right)p\left(x_6 \mid x_2,x_5\right)$$



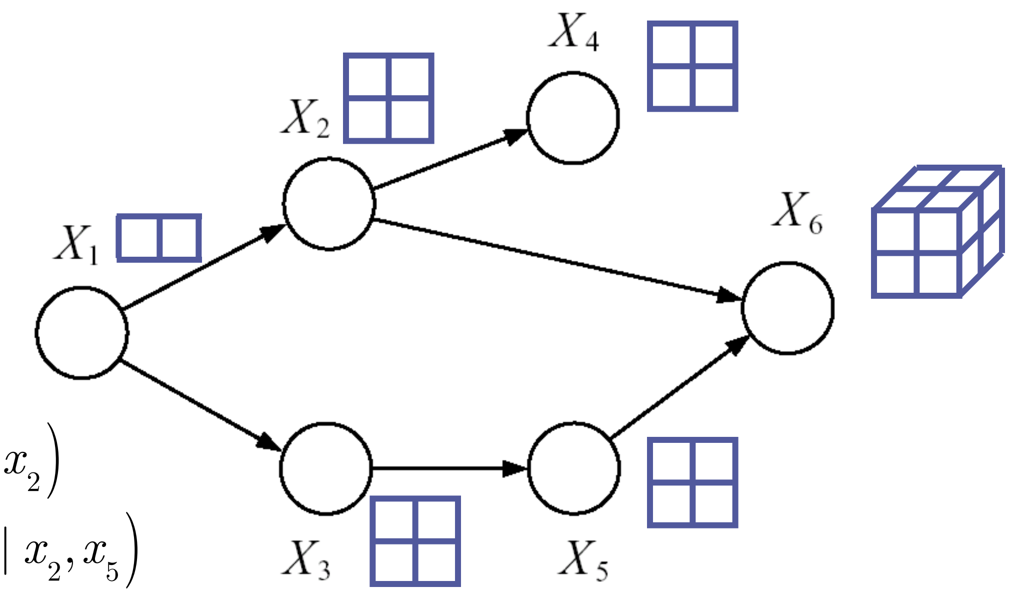

- Question? Which is the more general graph?



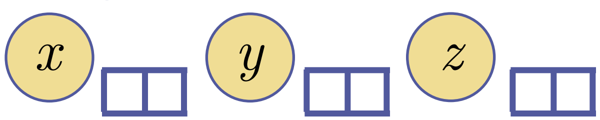
# Graphical Models

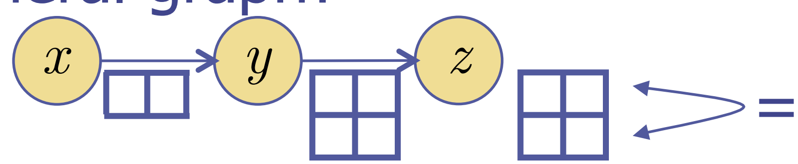
- Graph factorizes probability: $p\left(x_1,\ldots,x_n\right)=\prod_{i=1}^{n} p\left(x_i \mid \pi_i\right)$
- Topological graph:
  nodes are in order so that parents $\pi$ come before children




$$p\left(x_1,\ldots,x_6\right)=p\left(x_1\right)p\left(x_2 \mid x_1\right)\times p\left(x_3 \mid x_1\right)p\left(x_4 \mid x_2\right)\times p\left(x_5 \mid x_3\right)p\left(x_6 \mid x_2,x_5\right)$$

- Question? Which is the more general graph?




- Conditional probability tables can be chosen to make ‘busier’ graph look like simpler graph