Machine Learning

4771

Instructors:
Adrian Weller and Ilia Vovsha

Lectures 17-18

- Decompose Maximum Likelihood with hidden variables
- Expectation Maximization as Bound Maximization
- EM for Maximum A Posteriori (MAP)
- Intro to Graphical Models

Expectation Maximization

- Recall the problem…
- We have observed variables X
- Hidden variables Z (e.g. the class or Gaussian distribution from which we draw)
- Joint distribution $p(X,Z\,|\,\theta)$ depends on parameters θ (e.g. for Gaussian mixture have μ_k, Σ_k, π_k)

  complete likelihood
- Goal is to find $\hat{\theta}$ to maximize likelihood

$$p(X\,|\,\theta) = \sum_Z p(X,Z\,|\,\theta)$$

We'd like the true probability $p(Z\,|\,X,\theta)$

Instead we use an approximation $q_t(Z) = p(Z\,|\,X,\theta_t)$

Decompose Log Likelihood

- Let $q(Z)$ be any probability distribution over the latent variables Z
- Define $\mathcal{L}(q,\theta) = \sum_Z q(Z) \log\left(\frac{p(X,Z\mid\theta)}{q(Z)}\right)$

$p(X,Z\mid\theta) = p(X\mid\theta).p(Z\mid X,\theta)$

$$= \sum_Z q(Z)[\log p(X\mid\theta) + \log p(Z\mid X,\theta) - \log q(Z)]$$

$$= \log p(X\mid\theta) - \sum_Z q(Z) \log\left(\frac{q(Z)}{p(Z\mid X,\theta)}\right)$$

Does this look familiar?

Decompose Log Likelihood

- Let $q(Z)$ be any probability distribution over the latent variables Z
- Define $\mathcal{L}(q,\theta) = \sum_Z q(Z)\log\left(\frac{p(X,Z\mid\theta)}{q(Z)}\right)$

$p(X,Z\mid\theta) = p(X\mid\theta).p(Z\mid X,\theta)$

$$= \sum_Z q(Z)[\log p(X\mid\theta) + \log p(Z\mid X,\theta) - \log q(Z)]$$

$$= \log p(X\mid\theta) - \sum_Z q(Z)\log\left(\frac{q(Z)}{p(Z\mid X,\theta)}\right)$$

Does this look familiar?

Our log likelihood $- KL(q \,\|\, p(Z\mid X,\theta))$!

Decompose Log Likelihood

- Hence, the log likelihood

$$l(\theta) := \log p(X \mid \theta) = \mathcal{L}(q,\theta) + KL(q \,\|\, p(Z \mid X,\theta))$$

independent of q

lower bound

- E step: Lock $\theta = \theta_t$, maximize lower bound $\mathcal{L}$ wrt q
  - Recall $KL(q \,\|\, p) \geq 0$, best can do is $q_t = p(Z \mid X, \theta_t)$
- M step: Lock $q = q_t$, maximize lower bound $\mathcal{L}$ wrt θ
  - Observe can write $\mathcal{L}(q_t,\theta) = \sum_Z q_t(Z) \log\left(\frac{p(X,Z \mid \theta)}{q_t(Z)}\right)$

$$= \mathbb{E}_{q_t} \log p(X,Z \mid \theta) + H(q_t)$$

Sometimes called
Auxiliary Function $Q(\theta \mid \theta_t)$
or $Q_t(\theta)$

E step selects $Q_t(\theta)$ function
M steps sets $\theta_{t+1} = \arg\max_\theta Q_t(\theta)$

Expected complete likelihood using q_t

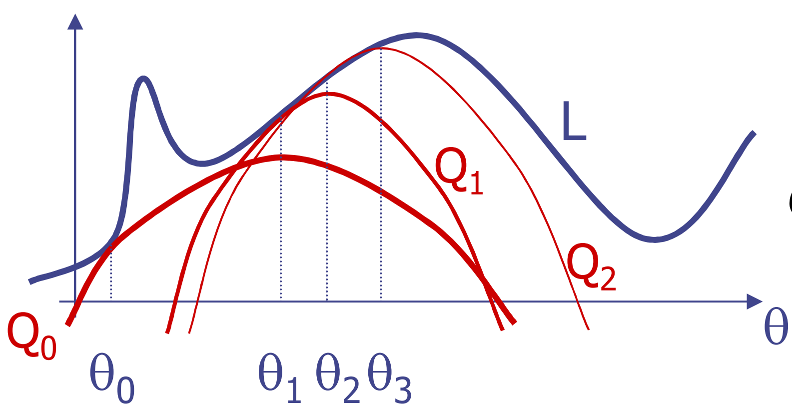
Entropy of q_t
indep of θ
treat as const

EM as Bound Maximization

- Bound Maximization: optimize a lower bound on $l(\theta)$

- Since log-likelihood $l(\theta)$ not concave, can't max it directly
- Consider an auxiliary function $Q(\theta)$ which is concave
- $Q(\theta)$ kisses $l(\theta)$ at a point and is less than it elsewhere

  matches gradient there – why?



$$l(\theta) \geq Q_t(\theta) \quad \forall\theta \ \forall t$$

$$l(\theta_t) = Q_t(\theta_t)$$

E step sets Q_t

$$Q_t(\theta_{t+1}) \geq Q_t(\theta_t)$$

M step sets

$$\theta_{t+1} = \arg\max_\theta Q_t(\theta)$$

Usually $Q_t(\theta_{t+1}) > Q_t(\theta_t)$

When not?

$$l(\theta_{t+1}) \geq Q_t(\theta_{t+1}) \geq Q_t(\theta_t) = l(\theta_t)$$

$$\Rightarrow l(\theta_{t+1}) \geq l(\theta_t)$$

- Monotonically increases log-likelihood (at least can't decrease)

M step

- Find θ to maximize the expected complete likelihood

$$\mathbb{E}_{q_t} \log p(X, Z \mid \theta)$$

- If $p(X, Z \mid \theta)$ is in the exponential family (recall includes Gaussian, Binomial, Multinomial, Poisson… Bishop 2.4) then the log cancels the exp and M step is simple, just weighted maximum likelihood! For example, for Gaussian mixture:

$$\frac{\partial Q(\theta)}{\partial \vec{\mu}_k} = \frac{\partial}{\partial \vec{\mu}_k} \sum_{n=1}^{N} \sum_k \tau_{n,k} \log \pi_k N\left(\vec{x}_n \mid \vec{\mu}_k, \Sigma_k\right)$$

$$0 = \sum_{n=1}^{N} \tau_{n,k} \frac{\partial}{\partial \vec{\mu}_k} \left(-\frac{1}{2} \left(\vec{x}_n - \vec{\mu}_k\right)^T \Sigma_k^{-1} \left(\vec{x}_n - \vec{\mu}_k\right) \right)$$

$$\vec{\mu}_k = \frac{\sum_{n=1}^{N} \tau_{n,k} \vec{x}_n}{\sum_{n=1}^{N} \tau_{n,k}}$$

… similarly get π_k and Σ_k

EM for Max A Posteriori

- We can also do MAP instead of ML with EM (stabilizes sol'n)

new terms

$$p(\theta \mid X) = \frac{p(X \mid \theta).p(\theta)}{p(X)} \Rightarrow \log p(\theta \mid X) = \mathcal{L}(q,\theta) + KL(q \| p) + \log p(\theta) - \log p(X)$$

indep of q const

- The E-step remains the same: lock θ, optimize q

- The M-step becomes slightly different for each model

- For example, mixture of Gaussians with prior on covariance

$$\log posterior\left(\theta\right) = \sum_{n=1}^{N} \log \sum_k \pi_k N\left(\vec{x}_n \mid \vec{\mu}_k, \Sigma_k\right) + \log \prod_k p\left(\Sigma_k \mid S, \eta\right) + \ const$$

$$\log posterior\left(\theta\right) \geq \sum_{n=1}^{N} \sum_k \tau_{n,k} \log \pi_k N\left(\vec{x}_n \mid \vec{\mu}_k, \Sigma_k\right) + \sum_k \log p\left(\Sigma_k \mid S, \eta\right) + const$$

- Updates on π and μ stay the same, only Σ is:

$$\Sigma_k \leftarrow \frac{1}{\sum_{n=1}^{N} \tau_{n,k} + \eta} \left(\sum_{n=1}^{N} \tau_{n,k} \left(\vec{x}_n - \vec{\mu}_k\right)\left(\vec{x}_n - \vec{\mu}_k\right)^T + \eta S \right)$$

- Typically, we use the identity matrix I for S and a small eta.

Intro to Graphical Models

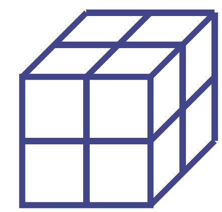
- Structuring Probability Functions for Storage
- Structuring Probability Functions for Inference
- Basic Graphical Models
- Graphical Models
- Parameters as Nodes

Structuring PDFs for Storage

- Probability tables quickly grow if p has many variables

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

- For D true/false "medical" variables $\mathit{table\,size} = 2^D\ ?$



- Exponential blow-up of storage size for the probability

- If variables are independent (Naïve Bayes assumption) then much more efficient

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

0.73	0.27

0.2	0.8

0.54	0.46

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

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 known/found variables X_f and unknown variables X_u predict queried variables X_q

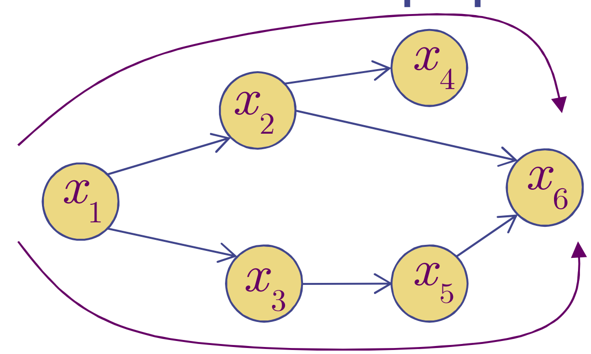
- Classical approach: truth tables (slow) or logic networks

- 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 ^ 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)

x3=0

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

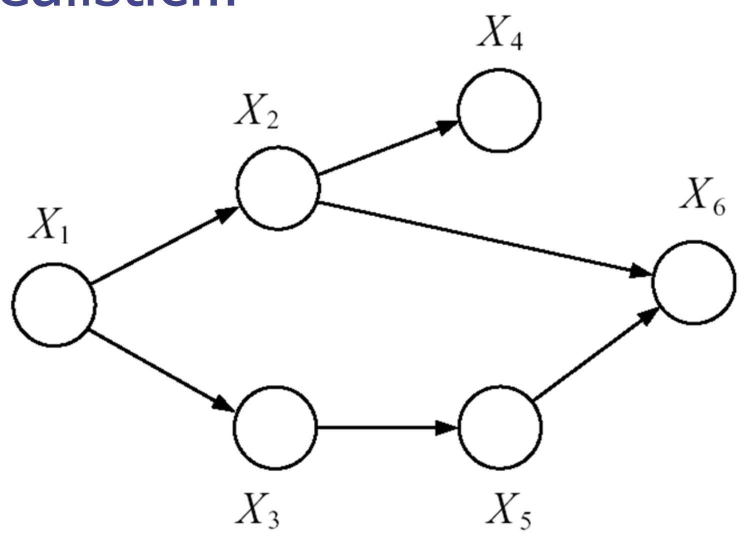
x3=1

	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

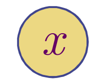
- Independence assumptions make probability tables smaller
- But real events in the world not completely independent!
- Complete independence is unrealistic…



- 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

$$p(x,y) = p(x)p(y)$$

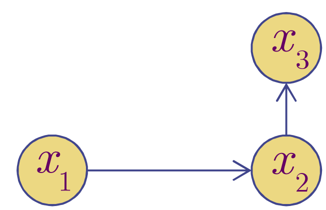
- **Dependent:** link

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

- **Arrow:** from parent to child (like causality, not exactly)
- **Child:** destination of arrow, response
- **Parent:** root of arrow, trigger

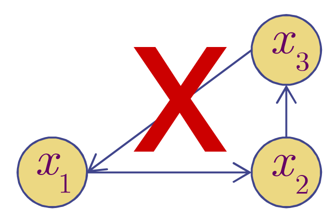
$$\mathit{parents\,of\,child\,i} = pa_i = \pi_i$$

- **Graph:** dependence/independence
- **Graph:** shows factorization of joint distribution as the products of conditionals



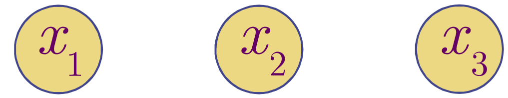
$$p\left(x_1,\ldots,x_n\right) = \prod_{i=1}^{n} p\left(x_i \mid pa_i\right) = \prod_{i=1}^{n} p\left(x_i \mid \pi_i\right)$$

- **DAG:** directed acyclic graph



Basic Graphical Models

- Independence: all nodes are unlinked

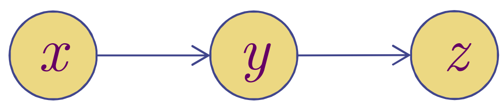


- Shading: variable is 'observed', condition on it moves to the right of the bar in the pdf

- Examples of simplest conditional independence situations…

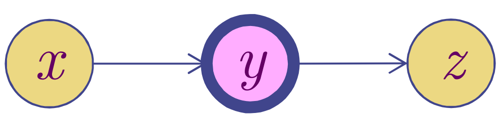
$$p\left(x_1,\ldots,x_n\right) = \prod_{i=1}^{n} p\left(x_i \mid pa_i\right) = \prod_{i=1}^{n} p\left(x_i \mid \pi_i\right)$$

1) Markov chain:



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

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



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

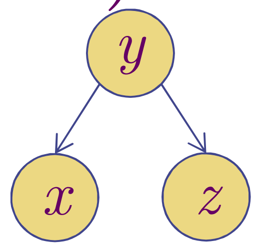
$$p\left(x \mid y,z\right) = \frac{p\left(x,y,z\right)}{p\left(y,z\right)} = p\left(x \mid y\right)$$

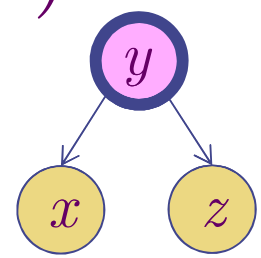
"x is conditionally independent of z given y"

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



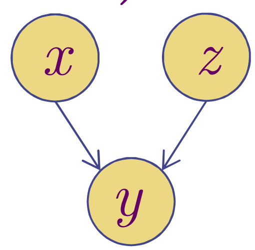


$x \perp\!\!\!\perp z \mid y$

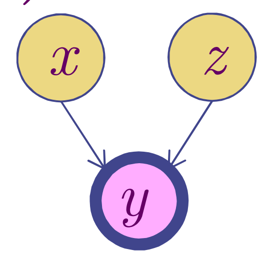
3) 2 Causes, 1 effect: $p(x,y,z) = p(x)p(z)p(y \mid x,z)$

x = aliens invade
y = mankind wiped out
z = giant asteroid hits

Explaining away…



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



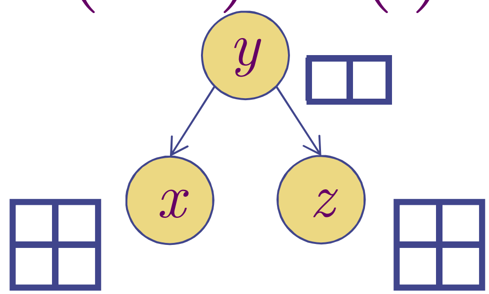
$x \cancel{\perp\!\!\!\perp} z \mid y$

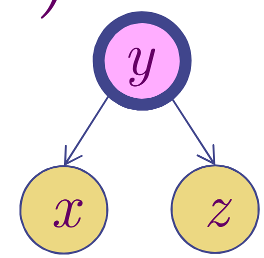
- For discrete variables, each conditional is a mini-table (Multinomial or Bernoulli conditioned on parents)

Basic Graphical Models

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

y = flu
x = sore throat
z = temperature



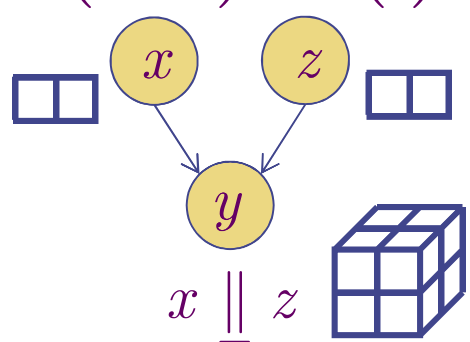


$x \perp\!\!\!\perp z \mid y$

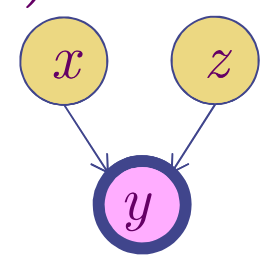
3) 2 Causes, 1 effect: $p\left(x,y,z\right) = p\left(x\right)p\left(z\right)p\left(y \mid x,z\right)$

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

Explaining away…



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



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

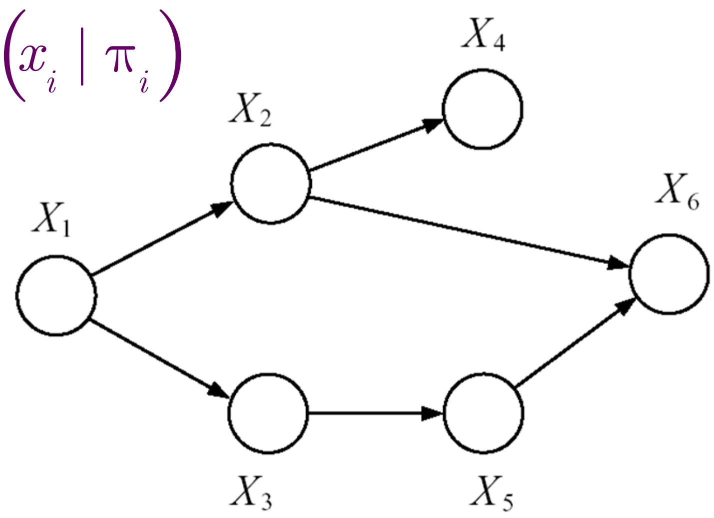
• For discrete variables, each conditional is a mini-table (Multinomial or Bernoulli conditioned on parents)

Graphical Models

- Example: factorization of the following system of variables

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

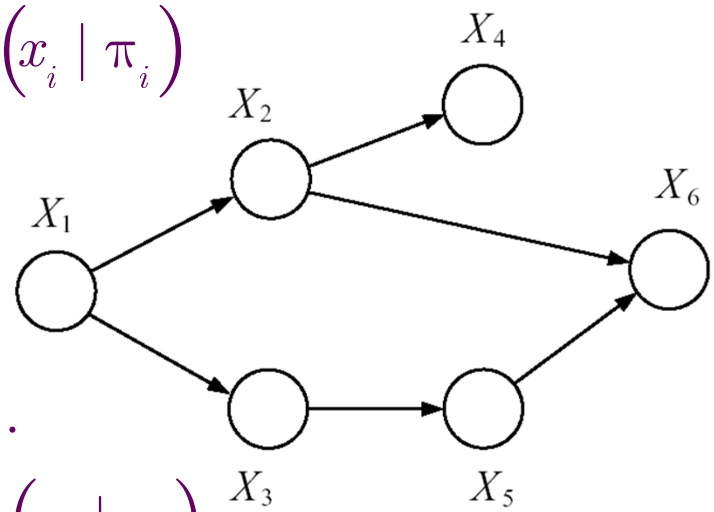
$$p\left(x_1,\ldots,x_6\right) = p\left(x_1\right)\ldots$$



Graphical Models

•Example: factorization of the following system of variables

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



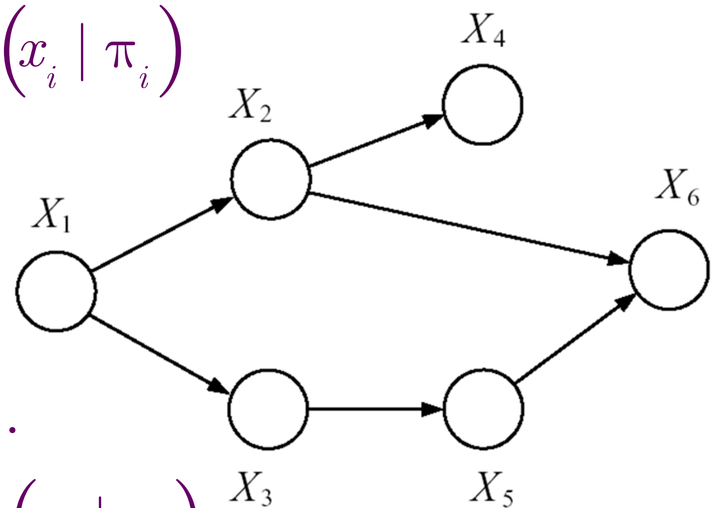
$$\begin{aligned}
p(x_1,\ldots,x_6) &= p(x_1)\ldots \\
&= p(x_1)p(x_2 \mid x_1)\ldots \\
&= p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)\ldots \\
&= p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2)\ldots \\
&= p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2)p(x_5 \mid x_3)\ldots \\
&= p(x_1)p(x_2 \mid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

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

Graphical Models

- Example: factorization of the following system of variables

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



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

2^6 $\quad$ 2^1 $\quad$ 2^2 $\quad$ 2^2 $\quad$ 2^2 $\quad$ 2^2 $\quad$ 2^3

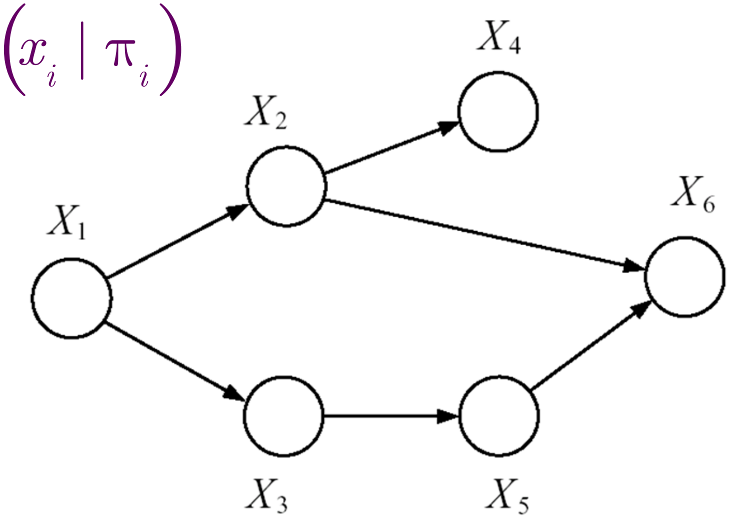
- How big are these tables (if binary variables)?

Graphical Models

- Example: factorization of the following system of variables

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

- Interpretation???



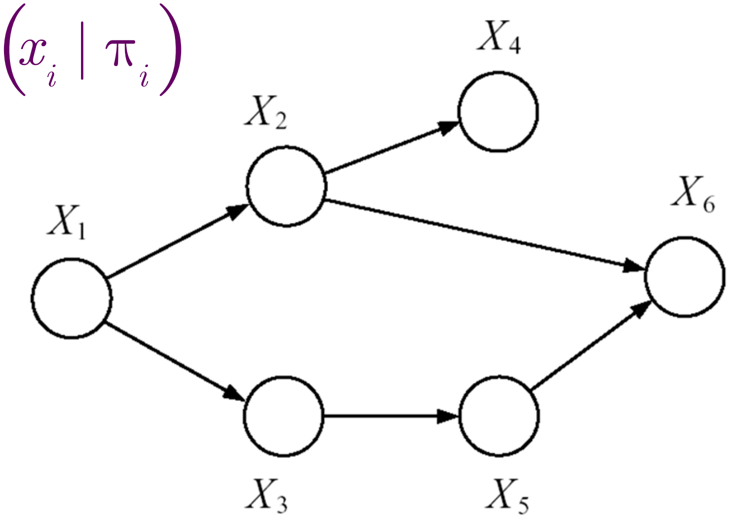
$$p\left(x_1,\ldots,x_6\right)=p\left(x_1\right)p\left(x_2 \mid x_1\right)p\left(x_3 \mid x_1\right)p\left(x_4 \mid x_2\right)p\left(x_5 \mid x_3\right)p\left(x_6 \mid x_2,x_5\right)$$

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

Graphical Models

- Example: factorization of the following system of variables

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



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

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

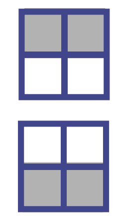
$2^6 \qquad 2^1 \qquad 2^2 \qquad 2^2 \qquad 2^2 \qquad 2^2 \qquad 2^3$

Graphical Models

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

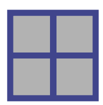
p(x) **2-1**

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



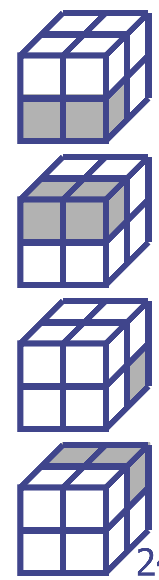
p(x,y) **4-1**

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

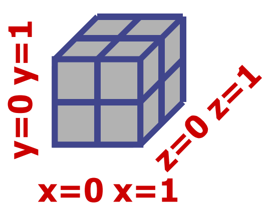


p(x|y) **4-2**

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



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



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

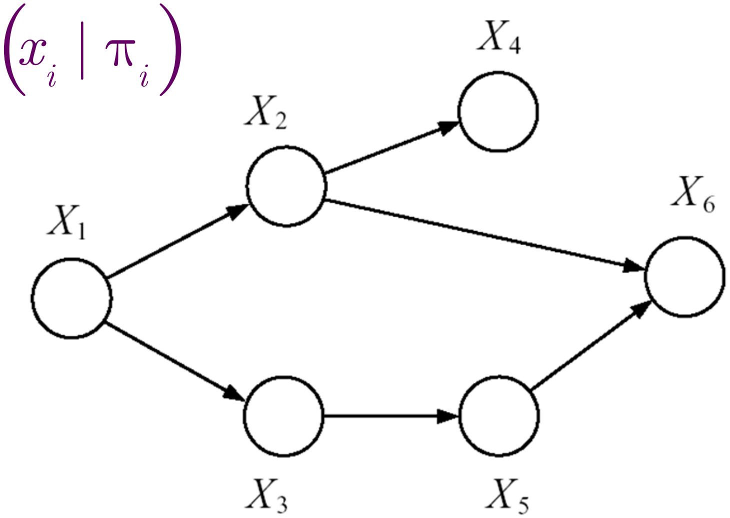
p(x|y,z) **8-4**

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

Graphical Models

- Example: factorization of the following system of variables

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



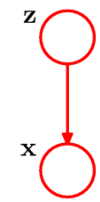
- Interpretation
  1: flu
  2: fever
  3: sinus infection
  4: temperature
  5: sinus swelling
  6: headache

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

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

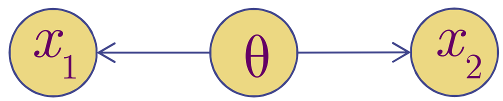
63 vs. 13 degrees of freedom

Mixture model
$p(x,z)=p(z)p(x|z)$

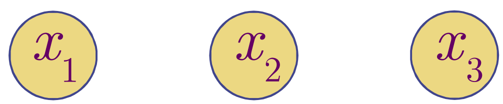


Parameters as Nodes

- Consider the model variable θ ALSO as a random variable

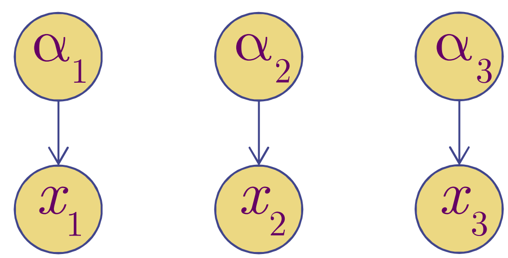


- But would need a prior distribution $P(\theta)$… ignore for now
- Recall: Naïve Bayes, probabilities are independent



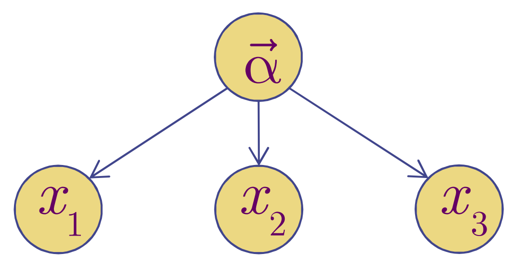
- Text: Multivariate Bernoulli

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



- Text: Multinomial

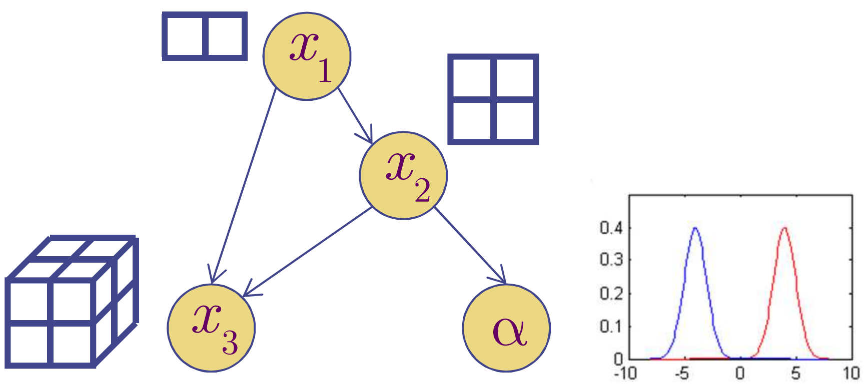
$$p\left(X \mid \vec{\alpha}\right)=\frac{\left(\sum_{m=1}^{M} X_m\right)!}{\prod_{m=1}^{M} X_m!}\prod_{m=1}^{M}\alpha_m^{X_m}$$



Continuous Conditional Models

- In previous slide, θ and α were a random variable in graph
- But, θ and α 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

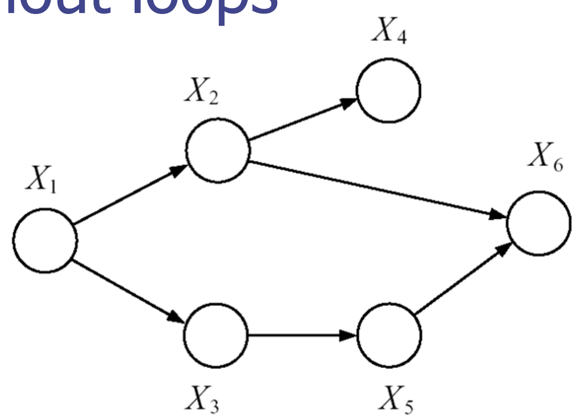


- Most popular continuous distribution = Gaussian

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:

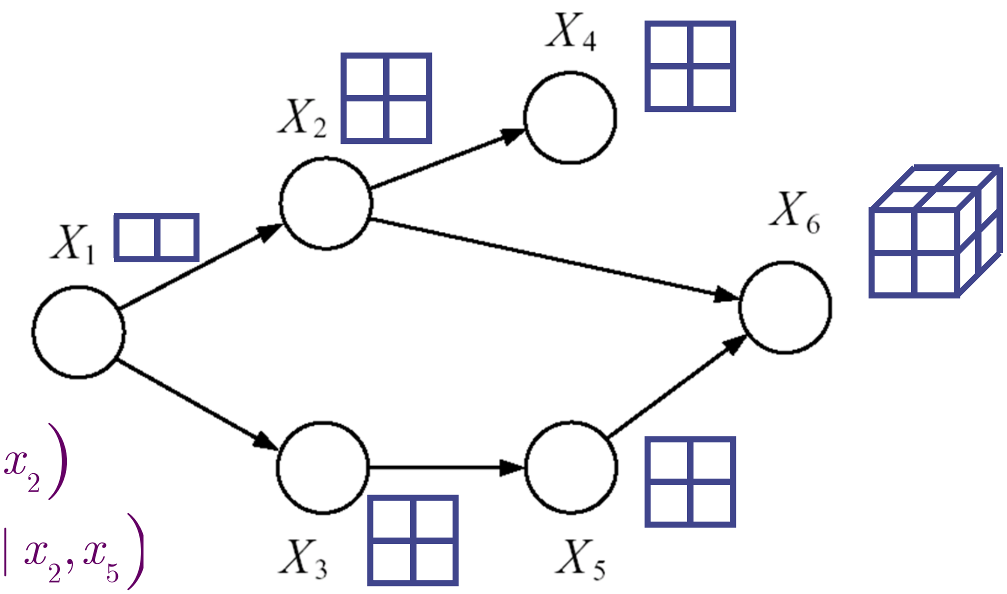
$$G = \{X, E\} = \{\texttt{nodes/randomvars,edges}\}$$
$$X = \{x_1, \ldots, x_M\}$$
$$E = \{(x_i, x_j) : i \neq j\}$$
$$X_c = \{x_1, x_3, x_4\} = subset$$



- 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)

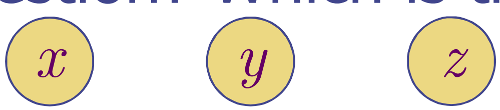
Graphical Models

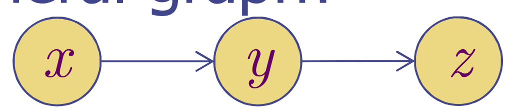
- Graph factorizes probability: $p\left(x_1,\ldots,x_n\right) = \prod_{i=1}^{n} p\left(x_i \mid \pi_i\right)$
- Topological graph: nodes are in order so that parents π come before children



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

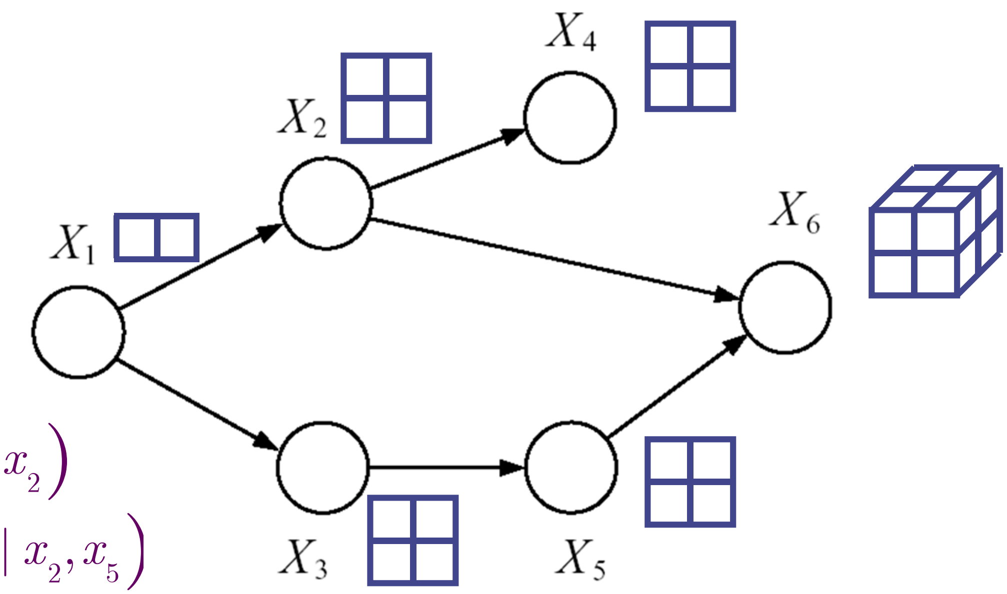
- Question? Which is the more general graph?





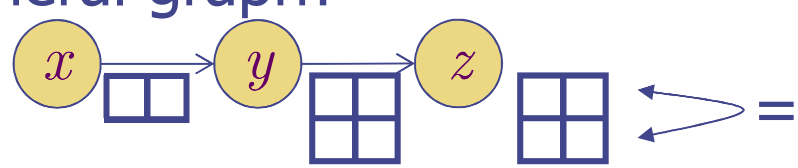
Graphical Models

- 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 π 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