Machine Learning

4771

Instructor: Tony Jebara

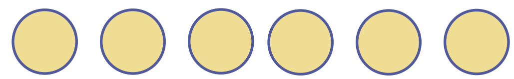
Topic 15

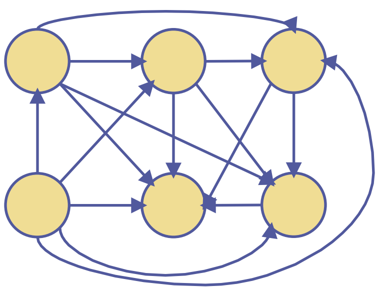
- Graphical Models
- Maximum Likelihood for Graphical Models
- Testing for Conditional Independence & D-Separation
- Bayes Ball

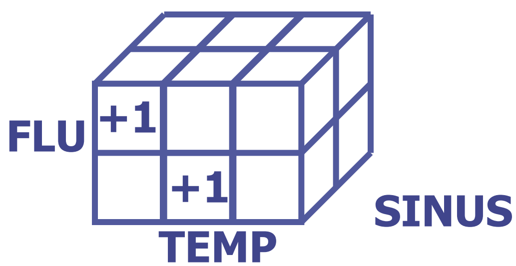
Learning Fully Observed Models

- Easiest scenario: we have observed all the nodes
- Want to learn the probability tables from data…
- Have N iid patients:

PATIENT	FLU	FEVER	SINUS	TEMP	SWELL	HEAD
1	Y	Y	N	L	Y	Y
2	N	N	N	M	N	Y
3	Y	N	Y	H	Y	N
4	Y	N	Y	M	N	N

- Simplest case: least general graph 
  handle each dim individually as Bernoulli/Multinomial
- 2nd Simplest case: most general, count each entry in pdf





Divide by total count

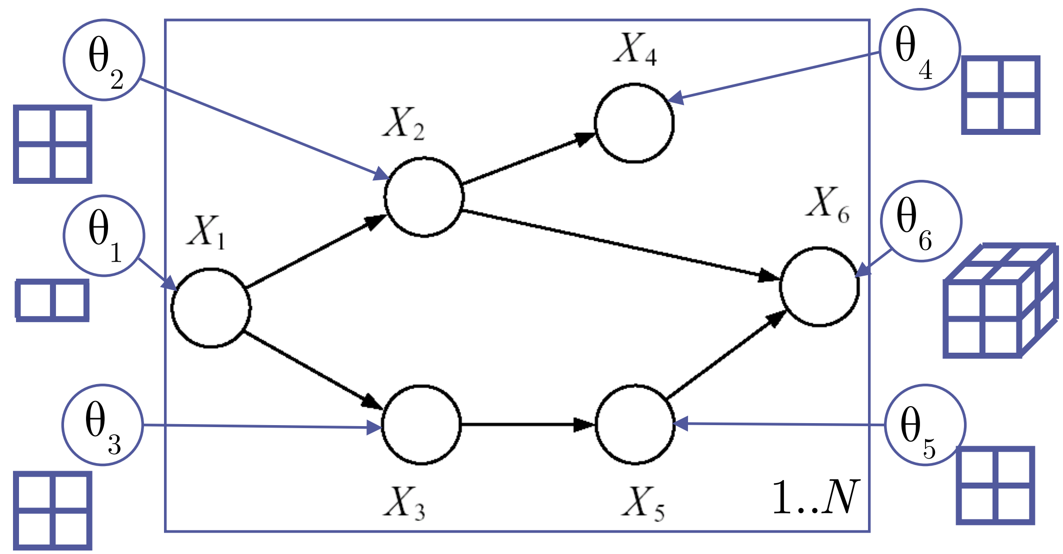
Since $\sum_{x_1} \cdots \sum_{x_6} p\left(x\right) = 1$

- What about learning graphs in between?

Maximum Likelihood CPTs

- Each conditional probability table θ_i part of our parameters
- Given table, have pdf

$$p\left(X_U \mid \theta\right) = \prod_{i=1}^{M} p\left(x_i \mid \pi_i, \theta_i\right)$$



- Have M variables:

$$X_U = \left\{x_1, \ldots, x_M\right\}$$

- Have N x M dataset:

$$\mathfrak{D} = \left\{X_{U,1}, \ldots, X_{U,N}\right\}$$

- Maximum likelihood:

$$\begin{aligned}
\theta^* &= \arg\max_\theta \log p\left(\mathfrak{D} \mid \theta\right) \\
&= \arg\max_\theta \sum_{n=1}^{N} \log p\left(X_{U,n} \mid \theta\right) \\
&= \arg\max_\theta \sum_{n=1}^{N} \log \prod_{i=1}^{M} p\left(x_{i,n} \mid \pi_{i,n} \theta_i\right) \\
&= \arg\max_\theta \sum_{n=1}^{N} \sum_{i=1}^{M} \log p\left(x_{i,n} \mid \pi_{i,n} \theta_i\right)
\end{aligned}$$

each θ_i appears independently, can do ML for each CPT alone! efficient storage & efficient learning

Maximum Likelihood CPTs

$$\delta\left(X_{U,n}, X_{U,m}\right) = \begin{cases} 1 & \text{if } X_{U,n} = X_{U,m} \\ 0 & \text{otherwise} \end{cases}$$

$$m\left(x_i\right) = \sum_{n=1}^{N} \delta\left(x_i, x_{i,n}\right)$$

$$m\left(X_U\right) = \sum_{n=1}^{N} \delta\left(X_U, X_{U,n}\right)$$

$$m\left(X_C\right) = \sum_{X_{U \setminus C}} m\left(X_U\right)$$

Counts: # of times what's in the bracket appeared in data, for example:

$$N = \sum_{x_1} m\left(x_1\right) = \sum_{x_1}\left(\sum_{x_2} m\left(x_1, x_2\right)\right) = \sum_{x_1}\left(\sum_{x_2}\left(\sum_{x_3} m\left(x_1, x_2, x_3\right)\right)\right)$$

- So... $l\left(\theta\right) = \sum_{n=1}^{N} \log p\left(X_{U,n} \mid \theta\right)$

$$= \sum_{n=1}^{N} \log \prod_{X_U} p\left(X_U \mid \theta\right)^{\delta\left(X_U, X_{U,n}\right)}$$

$$= \sum_{n=1}^{N} \sum_{X_U} \delta\left(X_U, X_{U,n}\right) \log p\left(X_U \mid \theta\right)$$

$$= \sum_{X_U} m\left(X_U\right) \log p\left(X_U \mid \theta\right) = \sum_{X_U} m\left(X_U\right) \log \prod_{i=1}^{M} p\left(x_i \mid \pi_i, \theta_i\right)$$

$$= \sum_{X_U} \sum_{i=1}^{M} m\left(X_U\right) \log p\left(x_i \mid \pi_i, \theta_i\right)$$

Maximum Likelihood CPTs

- Continuing: $l(\theta) = \sum_{X_U} \sum_{i=1}^{M} m(X_U) \log p(x_i \mid \pi_i, \theta_i)$

$$= \sum_{i=1}^{M} \sum_{x_i, \pi_i} \sum_{X_{U \backslash x_i \backslash \pi_i}} m(X_U) \log p(x_i \mid \pi_i, \theta_i)$$

$$= \sum_{i=1}^{M} \sum_{x_i, \pi_i} m(x_i, \pi_i) \log p(x_i \mid \pi_i, \theta_i)$$

- Define: $\theta(x_i, \pi_i) = p(x_i \mid \pi_i, \theta_i)$ Constraint: $\sum_{x_i} \theta(x_i, \pi_i) = 1$
- Now have above with Lagrange multipliers:

$$l(\theta) = \sum_{i=1}^{M} \sum_{x_i} \sum_{\pi_i} m(x_i, \pi_i) \log \theta(x_i, \pi_i) - \sum_{i=1}^{M} \sum_{\pi_i} \lambda_{\pi_i} \left(\sum_{x_i} \theta(x_i, \pi_i) - 1 \right)$$

$$\frac{\partial l(\theta)}{\partial \theta(x_i, \pi_i)} = \frac{m(x_i, \pi_i)}{\theta(x_i, \pi_i)} - \lambda_{\pi_i} = 0 \quad \rightarrow \quad \theta(x_i, \pi_i) = \frac{m(x_i, \pi_i)}{\lambda_{\pi_i}}$$

- Plug constraint: $\sum_{x_i} \frac{m(x_i, \pi_i)}{\lambda_{\pi_i}} = 1 \quad \rightarrow \quad \lambda_{\pi_i} = \sum_{x_i} m(x_i, \pi_i) = m(\pi_i)$

- Final solution (trivial!): $\boxed{\theta(x_i, \pi_i) = \frac{m(x_i, \pi_i)}{m(\pi_i)}}$

Maximum Likelihood CPTs

- Continuing: $l(\theta) = \sum_{X_U} \sum_{i=1}^{M} m(X_U) \log p(x_i \mid \pi_i, \theta_i)$

$$= \sum_{i=1}^{M} \sum_{x_i, \pi_i} \sum_{X_{U \setminus x_i \setminus \pi_i}} m(X_U) \log p(x_i \mid \pi_i, \theta_i)$$

$$= \sum_{i=1}^{M} \sum_{x_i, \pi_i} m(x_i, \pi_i) \log p(x_i \mid \pi_i, \theta_i)$$

- Define: $\theta(x_i, \pi_i) = p(x_i \mid \pi_i, \theta_i)$ Constraint: $\sum_{x_i} \theta(x_i, \pi_i) = 1$
- Now have above with Lagrange multipliers:

$$l(\theta) = \sum_{i=1}^{M} \sum_{x_i} \sum_{\pi_i} m(x_i, \pi_i) \log \theta(x_i, \pi_i) - \sum_{i=1}^{M} \sum_{\pi_i} \lambda_{\pi_i} \left(\sum_{x_i} \theta(x_i, \pi_i) - 1 \right)$$

$$\frac{\partial l(\theta)}{\partial \theta(x_i, \pi_i)} = \frac{m(x_i, \pi_i)}{\theta(x_i, \pi_i)} - \lambda_{\pi_i} = 0 \;\rightarrow\; \theta(x_i, \pi_i) = \frac{m(x_i, \pi_i)}{\lambda_{\pi_i}}$$

- Plug constraint: $\sum_{x_i} \frac{m(x_i, \pi_i)}{\lambda_{\pi_i}} = 1 \;\rightarrow\; \lambda_{\pi_i} = \sum_{x_i} m(x_i, \pi_i) = m(\pi_i)$
- Final solution (trivial!): $\boxed{\theta(x_i, \pi_i) = \frac{m(x_i, \pi_i) + \varepsilon}{m(\pi_i) + \varepsilon |x_i|}}$

MAP VERSION

Maximum Likelihood CPTs

- Let's try an example:
- Compute the cpt

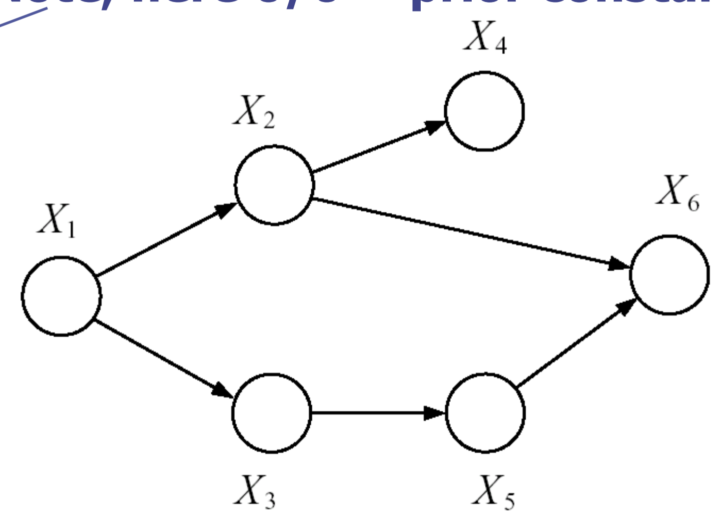
PATIENT	FLU	FEVER	SINUS	TEMP	SWELL	HEAD
1	Y	Y	N	L	Y	Y
2	N	N	N	M	N	Y
3	Y	N	Y	H	Y	N
4	Y	N	Y	M	N	N

$$p\left(x_3 \mid x_1\right)$$

- Using the formula:

$$\theta\left(x_i, \pi_i\right) = \frac{m\left(x_i, \pi_i\right)}{m\left(\pi_i\right)}$$

Note, here 0/0 = prior constant



$m\left(x_3, x_1\right)$:

	$x_1 = 0$	$x_1 = 1$
$x_3 = 0$	1	1
$x_3 = 1$	0	2

$m\left(x_1\right)$:

1	3

$p\left(x_3 \mid x_1\right)$:

1	1/3
0	2/3

Efficient, only count over subset of variables in p(X_B|X_A) Not all p(x_1,…,x_M)

Conditional Dependence Tests

- Another thing we would like to do with a graphical model:
  Check conditional independencies…
  "Is Temperature Indep. of Flu Given Fever?"
  "Is Temperature Indep. of Sinus Infection Given Fever?"
- Try computing & simplify marginals of p(x)

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

$$p(x_4 \mid x_1, x_2, x_3) = \frac{p(x_1, x_2, x_3, x_4)}{p(x_1, x_2, x_3)} = \frac{\sum_{x_5}\sum_{x_6} p(X)}{\sum_{x_4}\sum_{x_5}\sum_{x_6} p(X)}$$

$$= \frac{p(x_1)\,p(x_2 \mid x_1)\,p(x_3 \mid x_1)\,p(x_4 \mid x_2)}{p(x_1)\,p(x_2 \mid x_1)\,p(x_3 \mid x_1)}$$

$$= p(x_4 \mid x_2) \longleftrightarrow x_4 \perp\!\!\!\perp x_1, x_3 \mid x_2$$

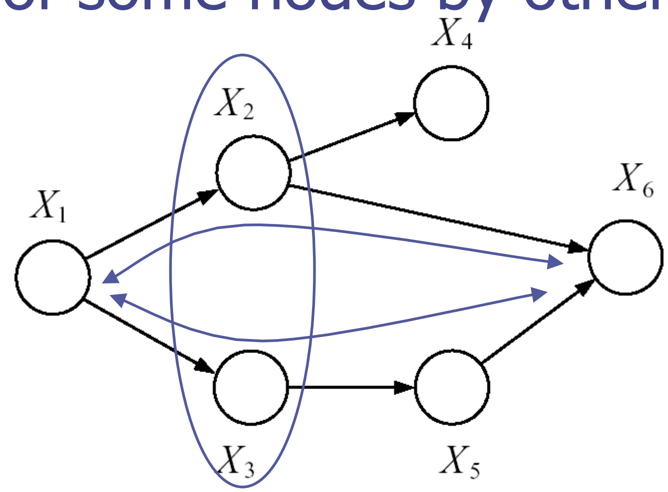
- In this case it was easy, what if checking: $x_1 \perp\!\!\!\perp x_6 \mid x_2, x_3$
- Hard to compute $p(x_1 \mid x_2, x_3, x_6)$ want <u>efficient</u> algorithm…

D-Separation & Bayes Ball

- There is a graph algorithm for checking independence
- Intuition: separation or blocking of some nodes by others
- Example:

if nodes x_2, x_3 "block" path from x_1 to x_6 we might say that

$$x_1 \perp\!\!\!\perp x_6 \mid x_2, x_3$$



- This is not exact for directed graphs (true for Undirected)
- We need more than just simple Separation
- Need D-Separation (directed separation)
- D-Separation is computed via the Bayes Ball algorithm
- Use to prove general statements over subsets of vars:

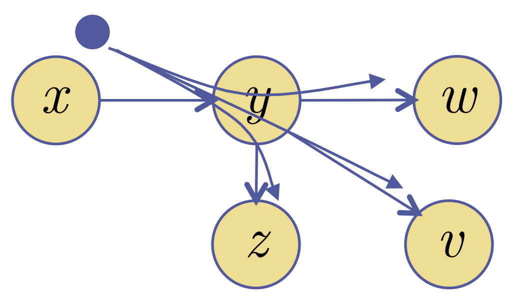
$$X_A \perp\!\!\!\perp X_B \mid X_C$$

Bayes Ball Algorithm

$$X_A \perp\!\!\!\perp X_B \mid X_C$$

- The algorithm:
  1) Shade nodes X_C
  2) Place a ball at each node in X_A
  3) Bounce balls around graph according to some *rules*
  4) If no balls reach X_B, then $X_A \perp\!\!\!\perp X_B \mid X_C$ is true (else false)

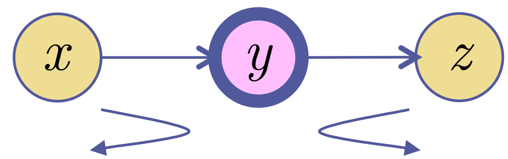
  Balls can travel along/against arrows
  Pick any incoming & outgoing path
  Test each to see if ball goes through or bounces back

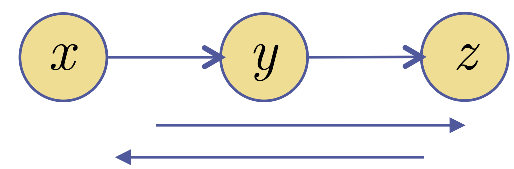


Look at canonical sub-graphs & leaf cases for rules…

Bayes Ball Algorithm

1) Markov Chain:



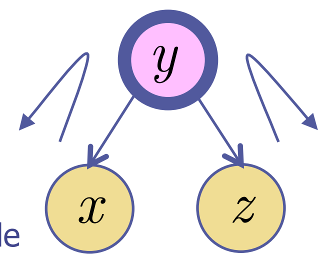


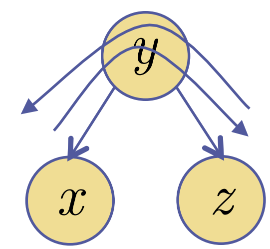
Only care about the shading of middle node

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

Go Through $x \cancel{\perp\!\!\!\perp} z$

2) Two Effects:



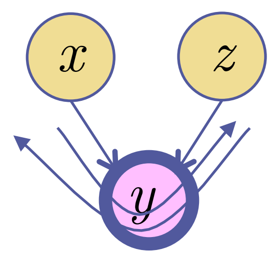


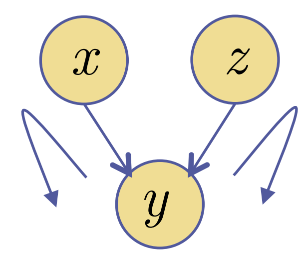
Only care about the shading of middle node

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

Go Through $x \cancel{\perp\!\!\!\perp} z$

3) Two Causes (V):





Only care about the shading of middle node

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

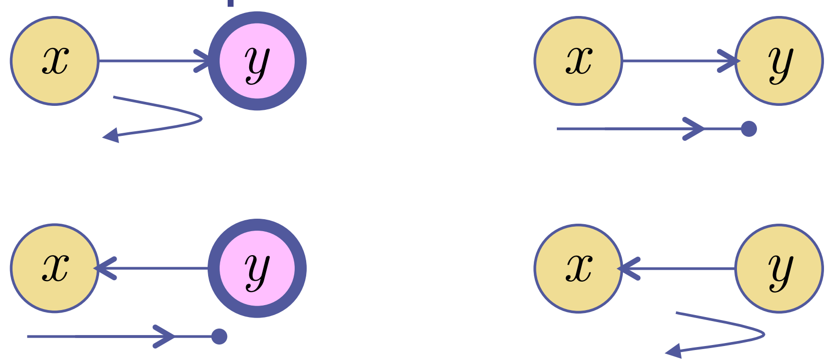
Bounce back $x \perp\!\!\!\perp z$

Bayes Ball Algorithm

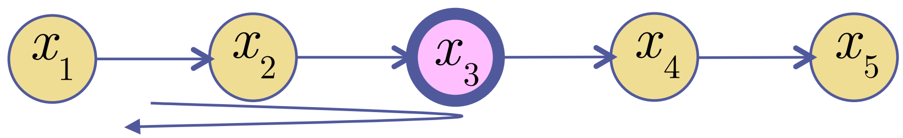
- Also need to look at special 'leaf' cases:

Bounces back — Ball is stopped

Ball is stopped — Bounces back

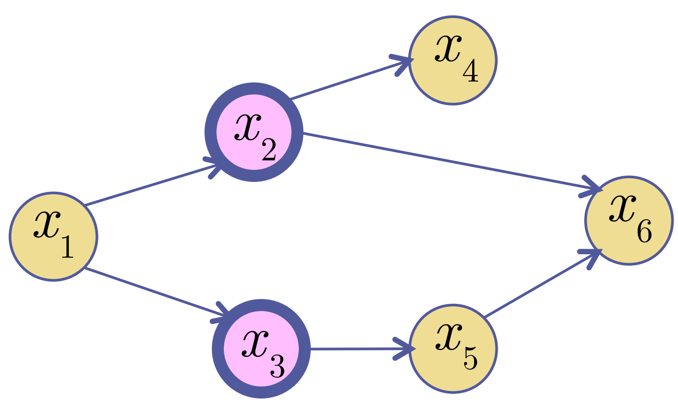


- Example:



$$x_1 \perp\!\!\!\perp x_5 \mid x_3$$

- Example:



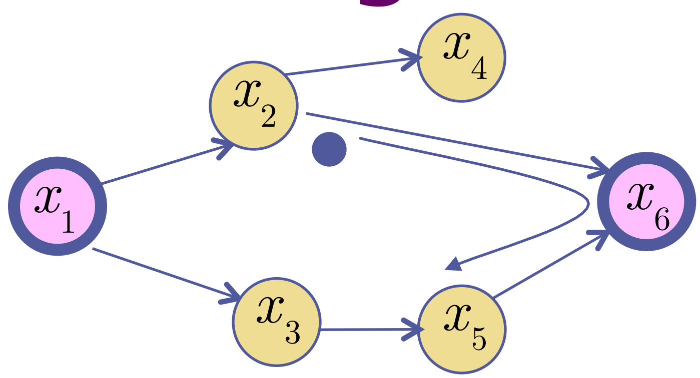
X_1, X_2, X_4 Stopped
X_1, X_2, X_6 Stopped
X_1, X_3, X_5 Stopped

$$\therefore \quad x_1 \perp\!\!\!\perp x_6 \mid \left\{ x_2, x_3 \right\}$$

Flu is independent of headache given fever & sinus infection!

Bayes Ball Algorithm

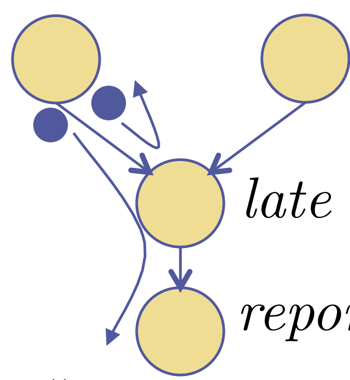
- Example:

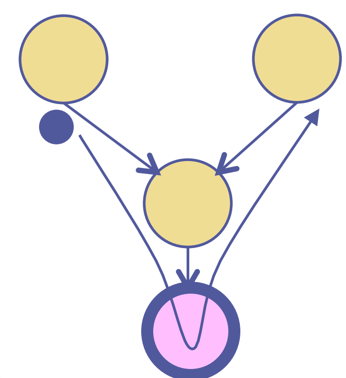


X_2, X_6, X_5 Goes Through Because of V-structure

$$\therefore \quad x_2 \not\perp\!\!\!\perp x_3 \mid \{x_1, x_6\}$$

- Example:





$$aliens \perp\!\!\!\perp watch \qquad aliens \not\perp\!\!\!\perp watch \mid report$$

Ball bounces back from report leaf and goes to right if report is shaded. Bob is waiting for Alice but can't know if she is late. Instead a security guard says if she is. She can be late if aliens abduct her or Bob's watch is ahead (daylight savings time). Guard reports she is late. If watch is ahead, p(alien=true) goes down, they are dependent.