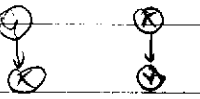
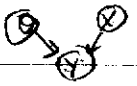
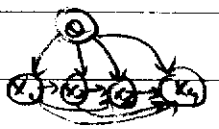
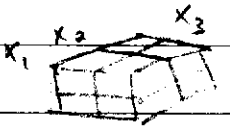
Matlab Tutorial — See online handouts...

Discriminative Learning — Skip for now...  pdf is suboptimal

ML & CML (hard)  with Bayesian Derivn. $p(y|x,\theta)$

Graphical Models — How to do ML & inference & efficiently using what we've learned?

"multinomial" learning with ML 



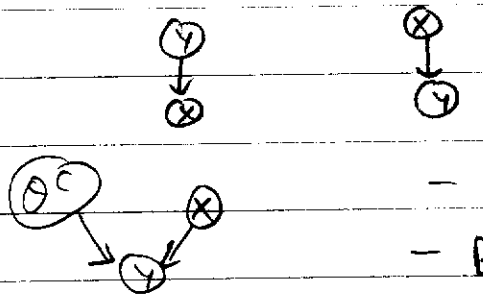
just estimate counts
count events
$\binom{0}{1}, \binom{0}{1}, \binom{1}{0} \ldots$

but, what if structured?

— Outline: Chapter 2: Directed Graphs, Undirected, Separation, Moralization, Bayes Ball, Chapter 3: Node Elimination, Chapter 8: Complete Graphs, Directed Learning, Undirected Learning, IPF

Next Week
ch. 3 & 8
Assignment #2 out tomorrow
due in 2 ~~weeks, at noon~~

Discriminative Learning



pdf is suboptimal $p(y|x,\theta)$ is used as marginal

- ML & CML
- Bayesian Derivation

$p(y|x,\theta)$ = bernoulli (two-class)

$p(y=1|x,\theta) = \mu(x)$

$p(y=0|x,\theta) = 1-\mu(x)$

$p(y|x) = \mu(x)^y (1-\mu(x))^{1-y}$

$p(y|x) = \mu^y (1-\mu)^{1-y}$

$$\mu(x) := \frac{1}{1+e^{-\eta(x)}} = \frac{1}{1+e^{-\theta^T x}}$$

$$\mu = \frac{1}{1+e^{-\eta}}$$

$$\eta = \log\left(\frac{\mu}{1-\mu}\right)$$

$$\frac{d\eta}{d\mu} = \frac{d}{d\mu}\log\left(\frac{\mu}{1-\mu}\right) = \frac{1}{\mu(1-\mu)}$$

$$\frac{d\mu}{d\eta} = \mu(1-\mu)$$

$$L(\theta) = \prod_{i=1}^{N} p(y_i|x_i;\theta) = \prod_n \mu_n^{y_n}(1-\mu_n)^{1-y_n}$$

$$l(\theta|D) = \sum_n y_n \log \mu_n + (1-y_n)\log(1-\mu_n)$$

$$\nabla_\theta l = \sum_n \frac{y_n}{\mu_n} - \frac{1-y_n}{1-\mu_n}\frac{d\mu_n}{d\eta_n}\frac{d\eta_n}{d\theta} = \sum_n \frac{y_n-\mu_n}{\mu_n(1-\mu_n)}\mu_n(1-\mu_n)x_n$$

$$= \sum_n (y_n-\mu_n)x_n := 0$$

$$\sum_n y_n x_n = \sum_n \mu_n x_n = \sum_n \theta^T x_n x_n$$

$X^T X\theta = X^T y$

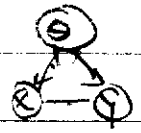
$\theta = (X^T X)^{-1} X^T y$

- Conditional Bayesian Inference & Conditional Max Likelihood

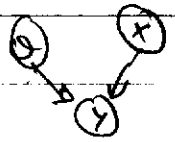
- previous def'n of $p(y|x)$ was implicit, did not specify $p(y|x)$ model as a conditioning of $p(x,y)$, i.e. $p(y|x) = \frac{p(x,y)}{\sum_y p(x,y)}$
- explicit definition of $p(x,y)$ requires different treatment

here $p(x,y|\theta)$ is computable yet we say θ only parametrizes the distrib. conditionally

joint inference



conditional



want conditional from data:

$$p(y|x) := p(y|x,X,Y) = \int p(y,\theta|x,X,Y)\,d\theta$$

$$= \int p(y|x,\theta,X,Y)\,p(\theta|x,X,Y)\,d\theta$$

$$= \int p(y|x,\theta)\{p(\theta|X,Y)\}\,d\theta$$

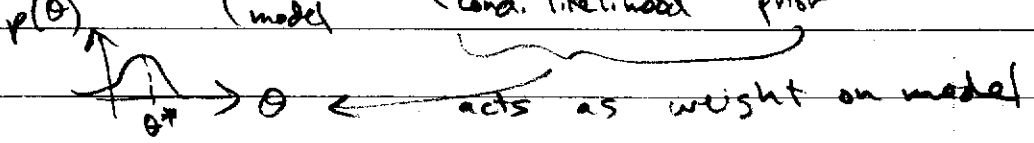
$$= \int p(y|x,\theta)\left\{\frac{p(Y|\theta,X)\,p(\theta,X)}{p(X,Y)}\right\}d\theta$$

$$= \int p(y|x,\theta)\left\{\frac{p(Y|\theta,X)\,p(X|\theta)\,p(\theta)}{p(X,Y)}\right\}d\theta$$

$$= \int p(y|x,\theta)\left\{\frac{p(Y|\theta,X)\,p(X)\,p(\theta)}{p(X,Y)}\right\}d\theta$$

$$\propto \int p(y|x,\theta)\;p(Y|\theta,X)\;p(\theta)\;d\theta$$

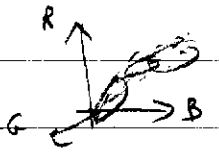
↑ model ↑ cond. likelihood ↑ prior



$p(y|x) \approx p(y|x,\theta^*)$ $\theta^* := \underset{\theta}{\text{argmax}}\; p(Y|\theta,X)\,p(\theta)$ ← conditional MAP

$p(y|x) \approx p(y|x,\theta^*)$ $\theta^* = \underset{\theta}{\text{argmax}}\; p(Y|\theta,X)$ ← conditional ML

- Assignment #2, EM, due date, 4 datasets, Matlab, C, ...
- EM motivation:

Skin RGB 

← $P(R,G,B)$

$P(X,Y)$ → 

- Matlab Tutorial – 4 functions
  – 4 datasets  digits, how represented lexigraphically as vectors $\vec{x}$

help fn
load filename
print -depsc filename

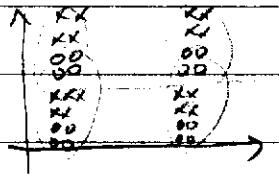
- Review ML & CML — implicit vs. explicit

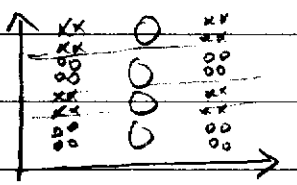
$l(\theta) = \sum_i \log p(x_i, y_i | \theta)$ ← splits data set $\quad p(y|x)$ $\quad p(y|x) = \frac{p(x,y)}{\sum_y p(x,y)}$

$cl(\theta) = \sum_i \log p(y_i | x_i, \theta)$

$= \sum_i \log p(x_i, y_i | \theta) - \sum_i \log \sum_y p(x_i, y | \theta)\, d\theta$

↑ log-sum again ∴ EM-like but negated, even worse repulsion term

example 



← used in speech rec.

- Seen ML, etc. for simple models (learning, inference, using pdf)
  have graphical models...
  ↑ discrete but easily extendable to continuous vars
- Next week: ch. 8 & 16

parent influences or causes child

X_1 & X_2 are dependent

Graphical Models — Directed vs. Undirected

- simplest case: Bernoulli: x [Heads | Tails] [0.7 | 0.3] $\rightarrow p(x)$

  multi dims: multinomial $x \in \{A, B, C, \ldots\}$

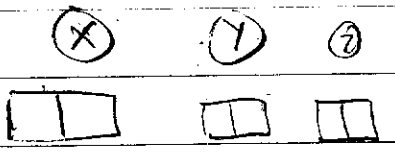
ML model was just counts (normalized)

Bayes model was counts + prior

- if we observe independent events

  do independent counts

i.e. rolling 3 dice at same time independent



$p(X = \text{Heads}) = \frac{\#\ X=\text{heads}}{\text{total}}$

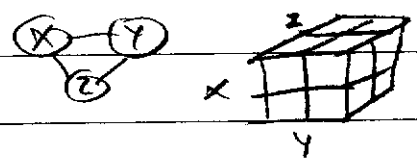
$p(X = \text{Heads}) = \frac{\#\ Y=\text{heads}}{\text{total}}$

if we observe dependent counts:

$p(X=H, Y=H) = \frac{\#\ X=H \ \&\ Y=H}{\text{total}}$

$p(X=H, Y=T) = \frac{\#\ X=H \ \&\ Y=T}{\text{total}}$

⋮



similarly

⋮

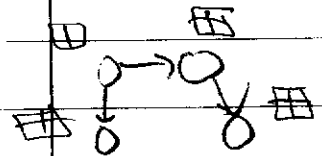
get tables of size C^M for M vars.

- Graphical Models: how to learn parameters?

  how to do inference?

  EFFICIENCY → storage / memory

  → computation



how to fill tables from data?

already saw this

- Graph is called $G = (X, E)$ or (V, E)
  - X: nodes (random vars), E: edges
  - V: vertices (random vars), E: edges

$X = \{X_1, \ldots, X_n\}$ $\quad E = \{(X_i, X_j) : i \neq j\}$

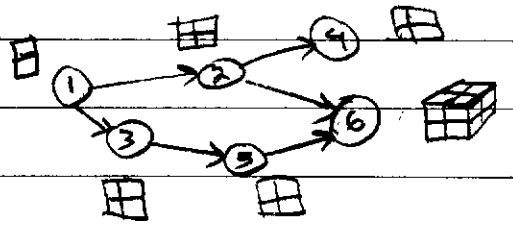
$X_c = \{X_1, X_3\}$ $\quad c$ is a subset

if G is <u>Acyclic</u>  then $p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | x_{\pi_i})$

graph specifies pdf qualitatively / structurally

still need to specify tables quantitatively with numerical

∴ graph ≡ family of possible pdfs.



1: fly
2: fever
3: sinus infection
4: temperature
5: sinuses swell
6: headache

<u>Topological Graph</u>: order the nodes $1 \ldots n$ such that all nodes in X_{π_i} appear before X_i

If topological $\{ X_i \perp\!\!\!\perp X_{v_i} | X_{\pi_i} \quad \forall i \}$

v_i: all vars appearing before X_i

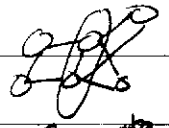
in other words $p(x_i | x_{v_i}, x_{\pi_i}) = p(x_i | x_{\pi_i})$

Example above: $p(x_1, \ldots, x_6) = p(x_1) p(x_2|x_1) p(x_3|x_1) p(x_4|x_2) p(x_5|x_3) p(x_6|x_2, x_5)$

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

$\therefore X_4 \perp\!\!\!\perp X_1, X_3 | X_2$ read off the graph

- Trickier case: $X_1 \perp X_6 \mid X_2, X_3$? harder to read off graph
- Cumbersome to compute $p(X_1 \mid X_6, X_2, X_3)$
- Want fast algorithm to check cond. independence
- Intuition: "separation" or "blocking" could imply cond. indep?



Separate Left from Right nodes

i.e. X_2 & X_3 separate path from X_1 to X_6 so indep?

this definition is true for undirected graphs

for directed, need more sophisticated algorithm which uses arrows to determine "d-separation".

- Bayes Ball Algorithm: efficiently tests cond. indep. statements for directed graphs.

"side note": an edge doesn't necessarily imply dependence

i.e. could have a table with numerical entries that are independent,

i.e. $p(x,y) =$  where $x \perp\!\!\!\perp y$

So conditional indep. tests for a property that must hold for this graph. But, due to numerical entries some extra independence properties could exist we won't detect.

3 types of subgraphs:

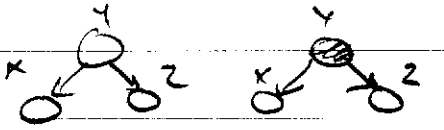
causal chain

$p(x,y,z) = p(x)\,p(y|x)\,p(z|y)$ $\quad x \perp\!\!\!\perp z \mid y$

$p(z|x,y) = p(z|y)$ conditioning on y blocks path from x to z

x = trip
y = fall
z = bruise

one cause multi effects



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

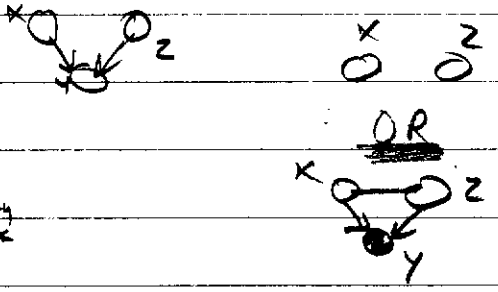
$p(x,z|y) = p(x|y)p(z|y)$

$X \perp\!\!\!\perp Z \mid y$

x = sore throat
y = flu
z = temperature

cond. on y "blocks" path from x to z

multi cause one effect



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

$X \perp\!\!\!\perp Z$

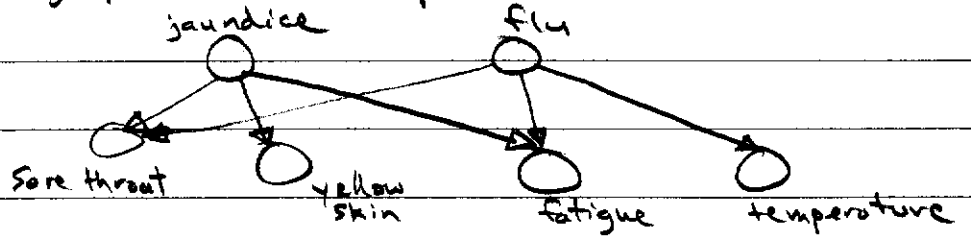
x = rain
y = wet driveway
z = car oil leak

cond. on y "unblocks" path from x to z

"V" structure, "explaining away"

- General graphs are composed of combos of the above

i.e. multicause multieffect

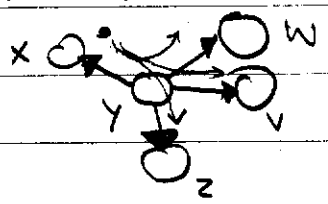


- need to compute d-separation for cond. indep.
- Bayes Ball Alg asks if $X_A \perp\!\!\!\perp X_B \mid X_C$

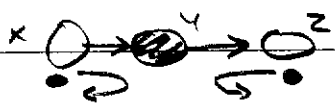
i.e. for a directed graph do nodes X_C d-separate nodes X_A & X_B

1) Shade nodes X_C (i.e. observed)
2) place a ball at each node in X_A
3) bounce around the graph according to some rules to see if any of the balls reaches X_B
4) if no balls reach X_B, then $X_A \perp\!\!\!\perp X_B \mid X_C$ IS TRUE
   else FALSE

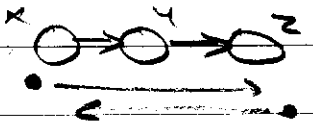
Rules: - Balls can travel along/against arrows
- pick any path & any outgoing path & test each at a time:



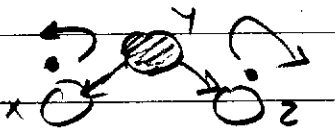
look at 3 canonical subgraphs



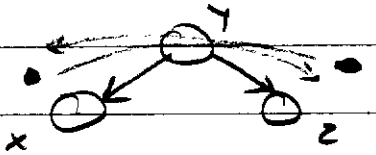
ball bounces back since $x \not\parallel z | y$



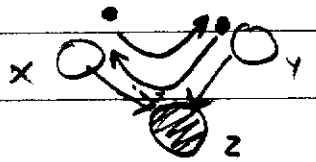
ball goes through since $x \not\parallel z$



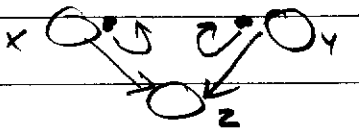
bounce back



go through

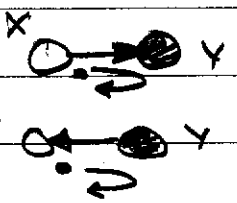


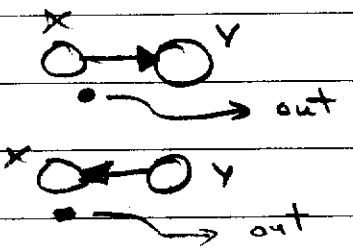
goes through



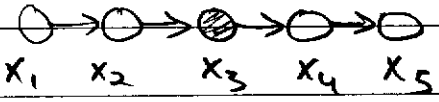
bounces back

if node Y is a leaf

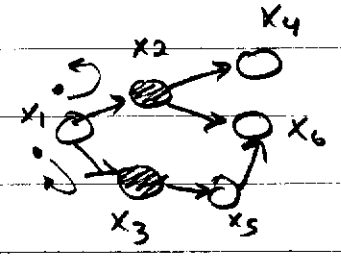




example



example



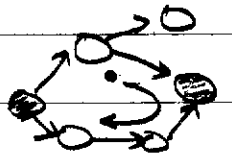
$X_1\ X_2\ X_4$ fails
$X_1\ X_2\ X_6$ fails
$X_1\ X_3\ X_5$ fails

X_1 to X_5 blocked by X_3

X_1 to X_2 blocked

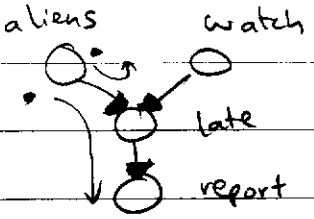
$\therefore\ X_1 \perp\!\!\!\perp X_6 \mid \{X_2, X_3\}$

But $X_2 \not\perp\!\!\!\perp X_3 \mid \{X_1, X_6\}$ false

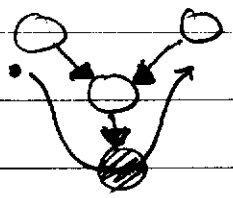


because of "V" structure

example



aliens $\perp\!\!\!\perp$ watch



aliens $\not\perp\!\!\!\perp$ watch | report

- 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 or Bob's watch is ahead (day light saving)
- report of her being late connects probs of aliens & watch. If watch is ahead, p(alien=true) ↓

Final Note:
- graph implies $\prod_i p(x_i \mid x_{\pi_i})$
- numerical can imply more
- set of pdfs implied by graph
- set of pdf's that satisfy all cond. indep. properties given by Bayes Ball

i.e. $X_A \perp\!\!\!\perp X_C \mid X_D$ ⇒ all pdfs in graph satisfy this
$X_A \not\perp\!\!\!\perp X_C \mid X_B$ ⇒ some pdfs in graph don't satisfy this