Compact Representation of Distributions
Directed Acyclic Graphical Models
Bayes-ball Algorithm

A6: - typo in Fenchel dual:
$$D(y) = -f^*(-y^*) - g^*(A^T y)$$
- extra hints now included
- Due Monday!

Midterm: - cancel class monday: 3pm in class!
- Midterm November 12, 17 TBD
- tutorial moved? TBD

Compact Representions of Joint Distributions

Notation for today:
- we'll use x_j for $(x_i)_j$ — training example i; vector, pick j

Consider $x \in \{0,1\}^d$ (binary vectors)

We want to model $p(x)$ (joint distribution)

- why? scientific discovery, outlier detection, $p(x_i \mid y_i)$ in generative model, $p(x_A \mid x_B)$

- previous: naive bayes, Gaussian, Mixture Models
- today: "graphical" models

Basic Idea behind directic acyclic graphical models:
- use product rule repeatedly:

$$p(x) = p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_2, x_1) \cdots \underbrace{p(x_d \mid x_{1:(d-1)})}_{2^{D-1} \text{ parameters}}$$
$$= \prod_{j=1}^{d} p(x_j \mid x_{1:(j-1)})$$

- too many parameters

Solutions:
- "parsimonious" parameterization
$$p(x_j \mid x_{1:(j-1)}) = f(w^T x_{1:(j-1)})$$
D parameters

Can also do both

- conditional independence

eg. naive bayes: $x_j \perp x_{1:(j-1)} \mid y_i \Rightarrow p(x_j \mid x_{1:(j-1)}, y_i) = p(x_j \mid y_i)$

useful for time: markov chain: $x_j \perp x_{1:(j-2)} \mid x_{j-1} \Rightarrow p(x_j \mid x_{1:(j-1)}) = p(x_j \mid x_{j-1})$

General: $x_j \perp x_{1:(j-1) \setminus \pi(j)} \mid \pi(j)$ "parents"

Directed Acyclic Graphical (DAG) models

"Bayesian Networks"
"Belief Networks"
"Causal Networks" → (you need a justification to interperet edges causally)

Graph $G = (V, E)$ — "vertices", "edges"

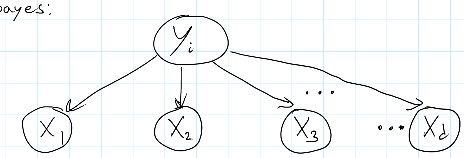
V: random variables x_j

E: $x_{\pi(j)} \to x_j$

$\pi(j)$: parents of j

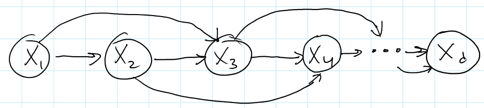
$$p(x) = \prod_{j=1}^{d} p(x_j \mid x_{\pi(j)})$$

Naive Bayes:



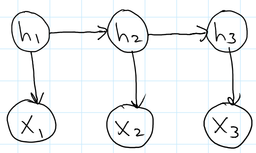
Markov Chain:

2nd Order Markov Chain:

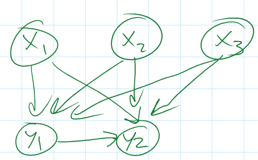


$$p\left(x_j \mid x_{j-1}, x_{j-2}\right)$$

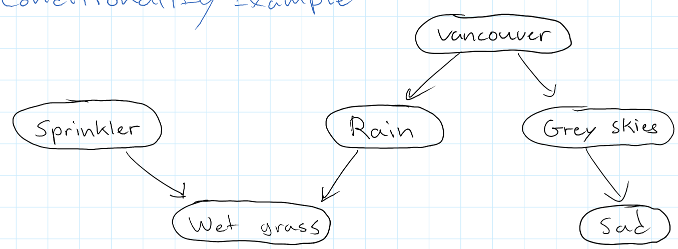
Hidden Markov Model: (kalman filtering)



Conditional DAG:



Conditionality Example



$$x_j \to x_{j+1}$$

Learning

- fit each $p(x_j \mid x_{\pi(j)})$ independently (parents)
- "inference" computing $p(x)$; computing $p(x_j \mid x_{\pi(j)})$ } easy

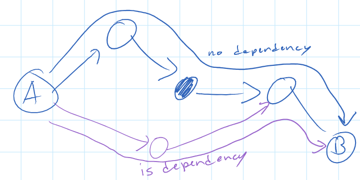
computing $p(x_j \mid x_{j+1})$ } #P - Hard

Conditional Independence Properties

- We can use the graph to check whether or not

$$X_A \perp X_B \mid X_B$$

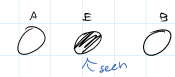
follows from conditional independence assumptions.



$X_A \perp X_B \mid X_E \iff$ A and B are "d-separated" given E

A and B are d-separated if for all paths 'P' between A and B, at least one of the following holds:

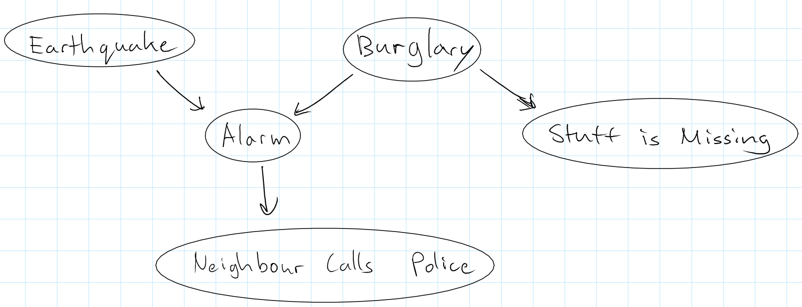
1. P includes a "chain": $A \rightarrow E \rightarrow B$



2. P includes a "fork": $A \leftarrow E \rightarrow B$

note: no E here

3. P contains a "collider": $A \rightarrow C \leftarrow B$, where C and all its descendents are unobserved



Earthquake $\not\perp$ call — dependent: increase chances

Earthquake $\perp$ call | Alarm — independant

Alarm $\not\perp$ Stuff Missing — dependent

Alarm $\perp$ Stuff Missing | Buglary — independent

Earthquake $\perp$ Buglary — independent

Earthquake $\not\perp$ Buglary | Alarm — dependent "explaining away"

Earthquake $\not\perp$ Buglary | Call — dependent (descendant of alarm)

Buglary $\not\perp$ Call

Call $\not\perp$ stuff is missing

Gaussian DAGs ⊛

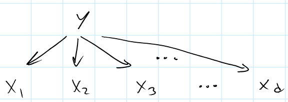
$$p(x_j \mid x_{\pi(j)}) \sim \mathcal{N}\left((\mu_j + \bar{w}_j^T x_{\pi(j)}), (\sigma_j^2)\right)$$

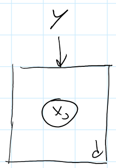
$$\iff x \sim \mathcal{N}(\mu, \Sigma)$$

$$\Sigma^{-1} = LL^T$$

non-zero pattern of L comes from G

Plate Notation





d times

x_j up to d