Readings:

Review: K&F: *2.1*, 2.5, 2.6

K&F: 3.1

Introduction

Graphical Models – 10708

Carlos Guestrin

Carnegie Mellon University

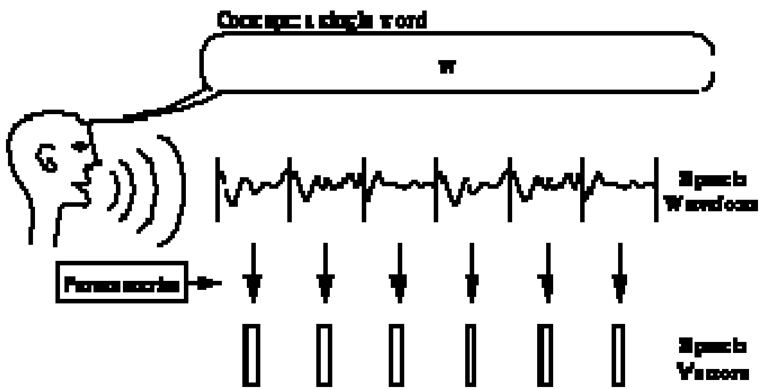
September 13th, 2006

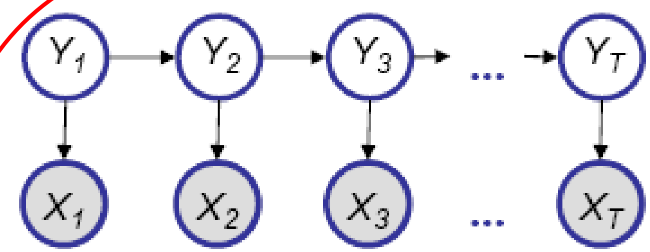
One of the most exciting developments in machine learning (knowledge representation, AI, EE, Stats,…) in the last two (or three, or more) decades…

My expectations are already high… ☺

Speech recognition

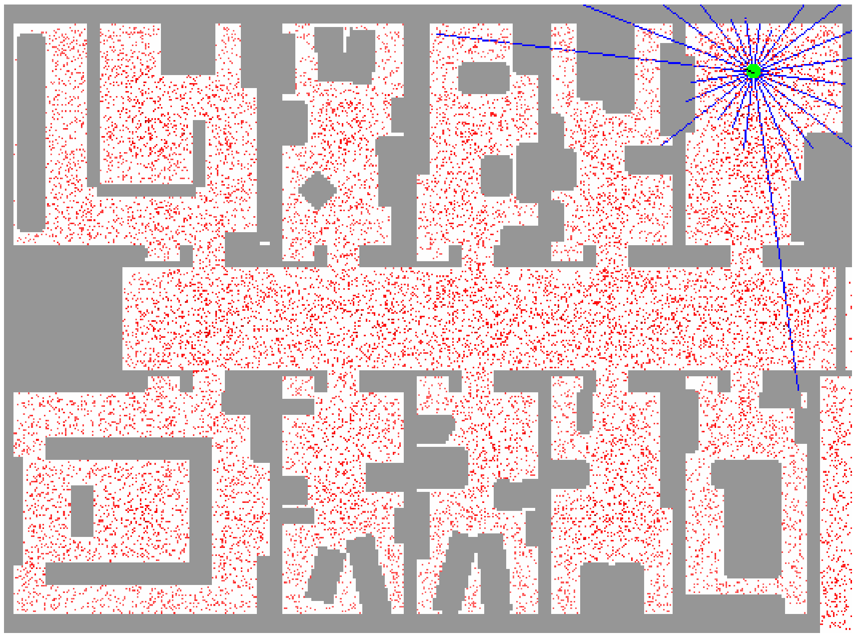
Hidden Markov models and their generalizations



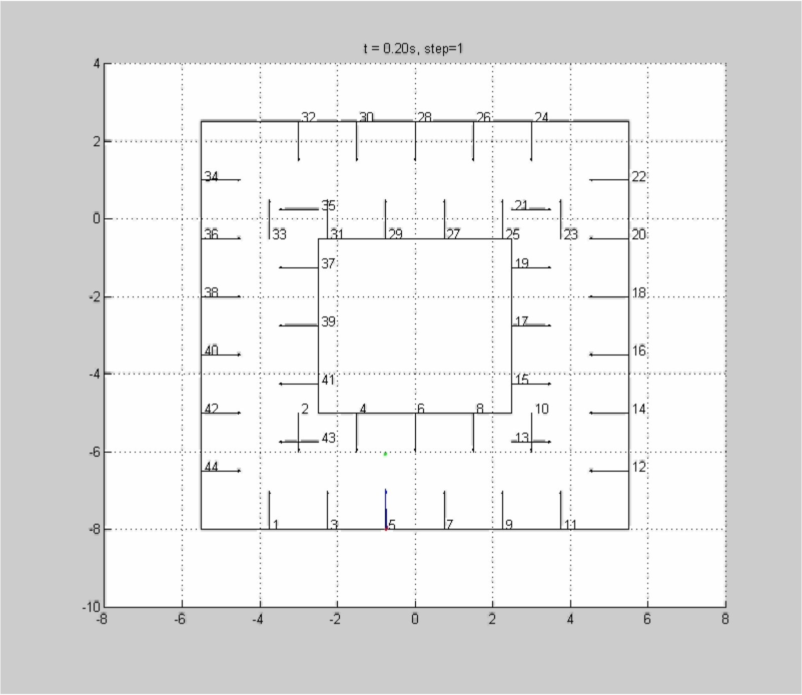


Tracking and robot localization

Kalman Filters



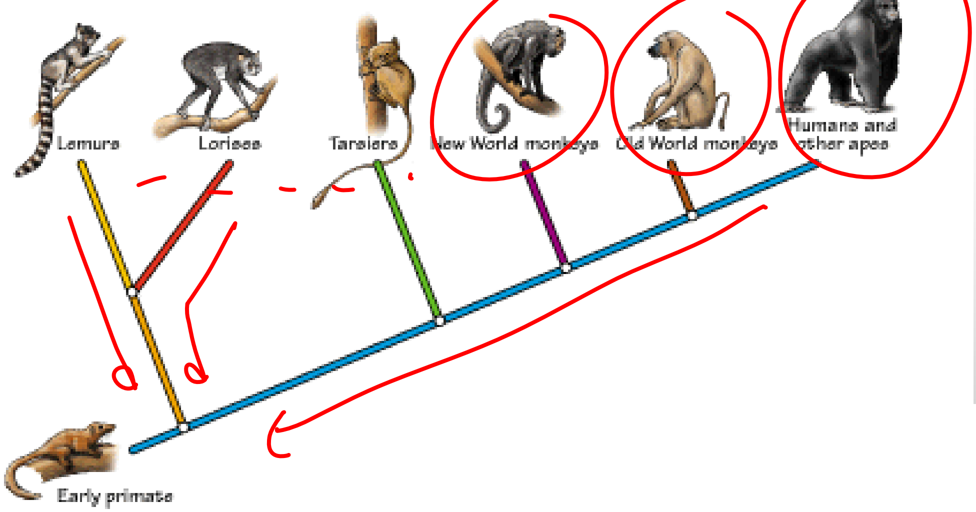
[Fox et al.]

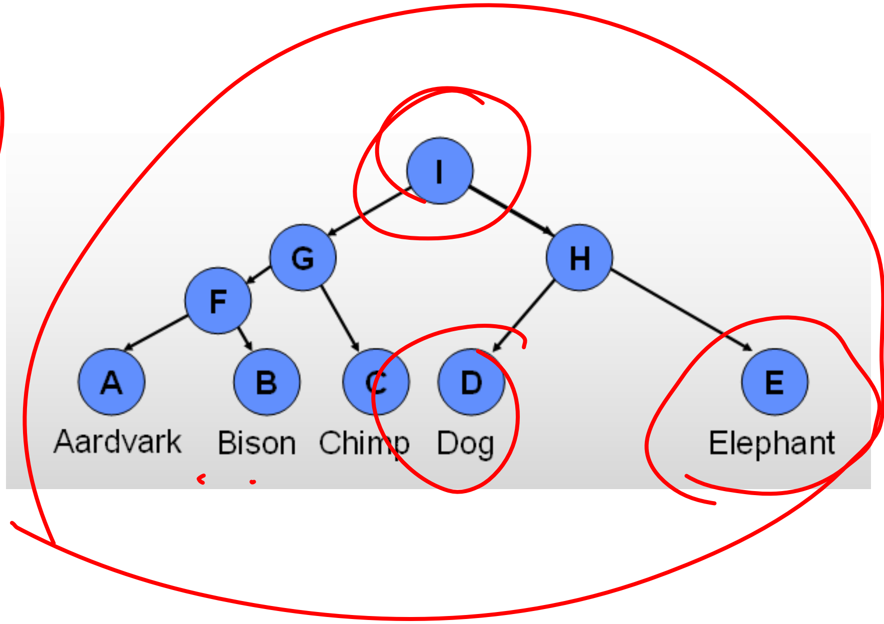


[Funiak et al.]

Evolutionary biology

Bayesian networks



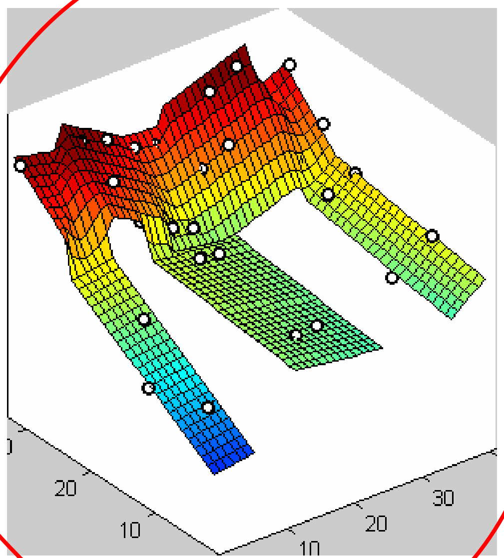


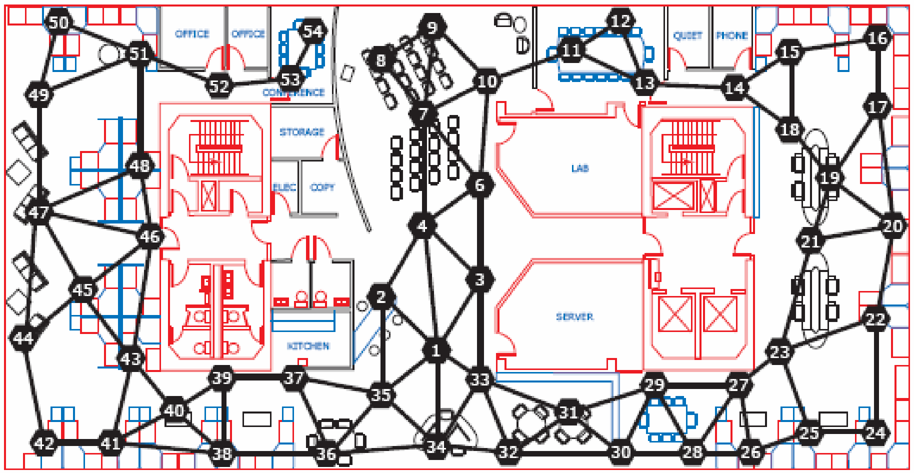
[Friedman et al.]

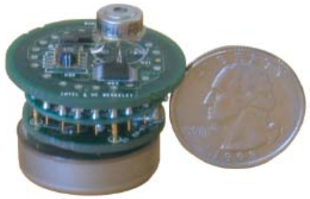
Modeling sensor data

Undirected graphical models

Markov Networks



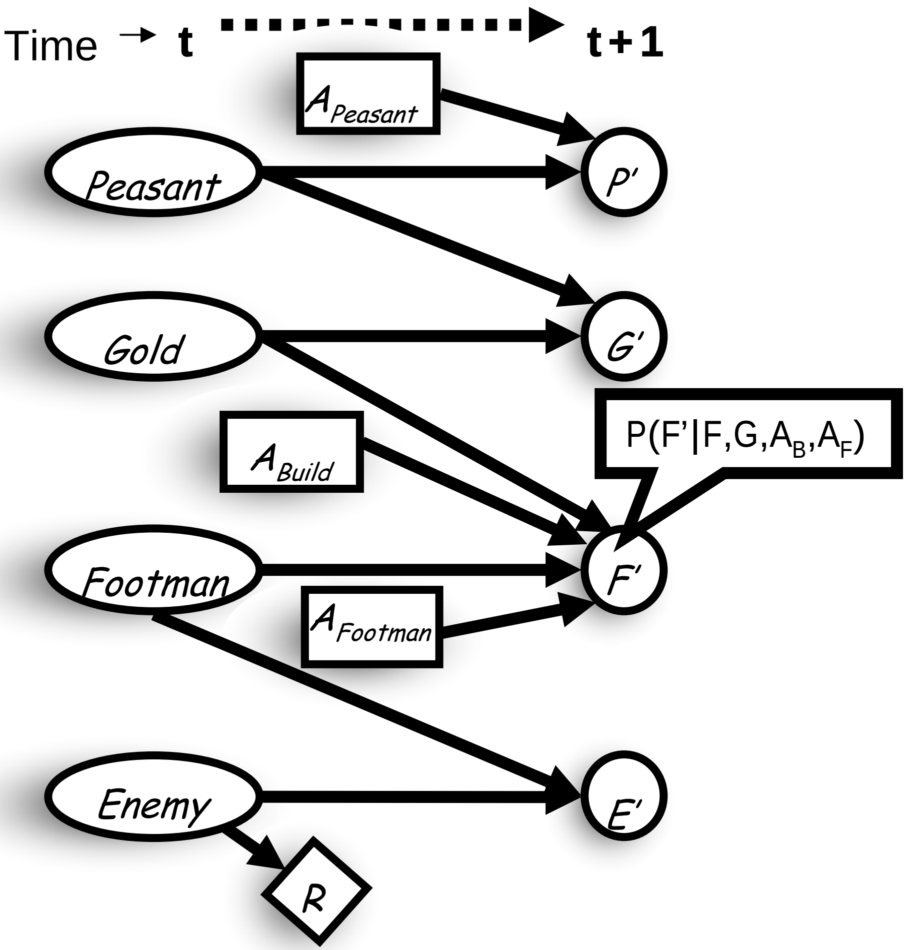


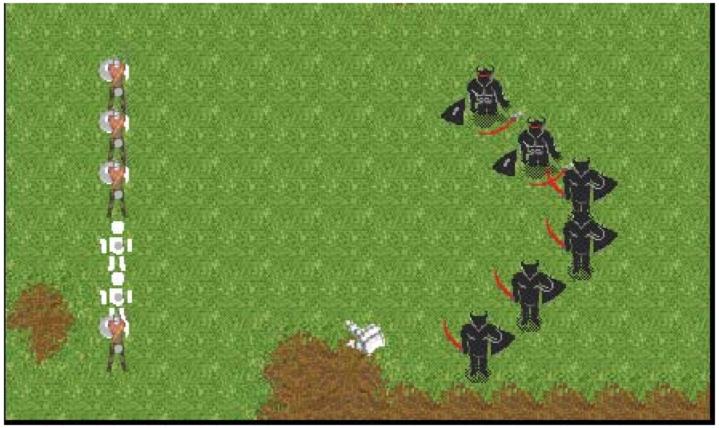


[Guestrin et al.]

Planning under uncertainty

- Dynamic Bayesian networks
- Factored Markov decision problems

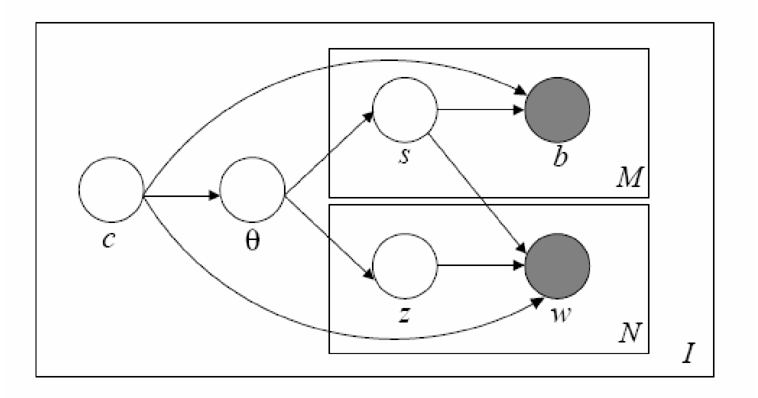


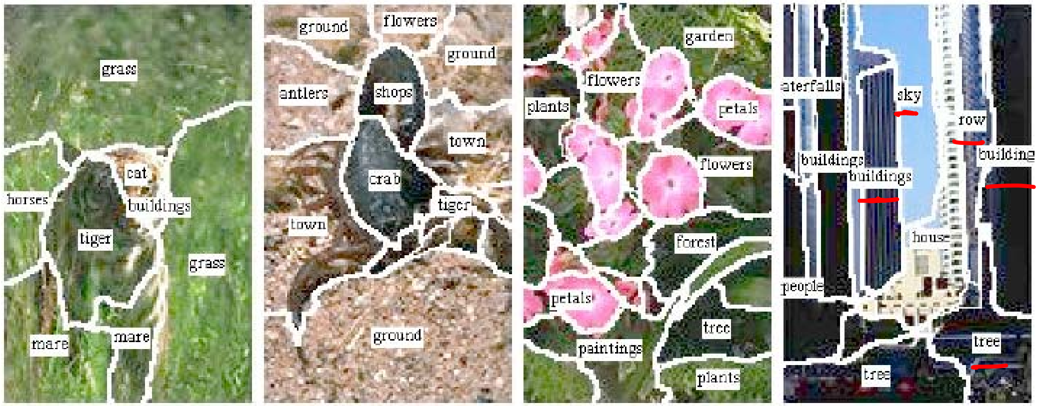


[Guestrin et al.]

Images and text data

Hierarchical Bayesian models

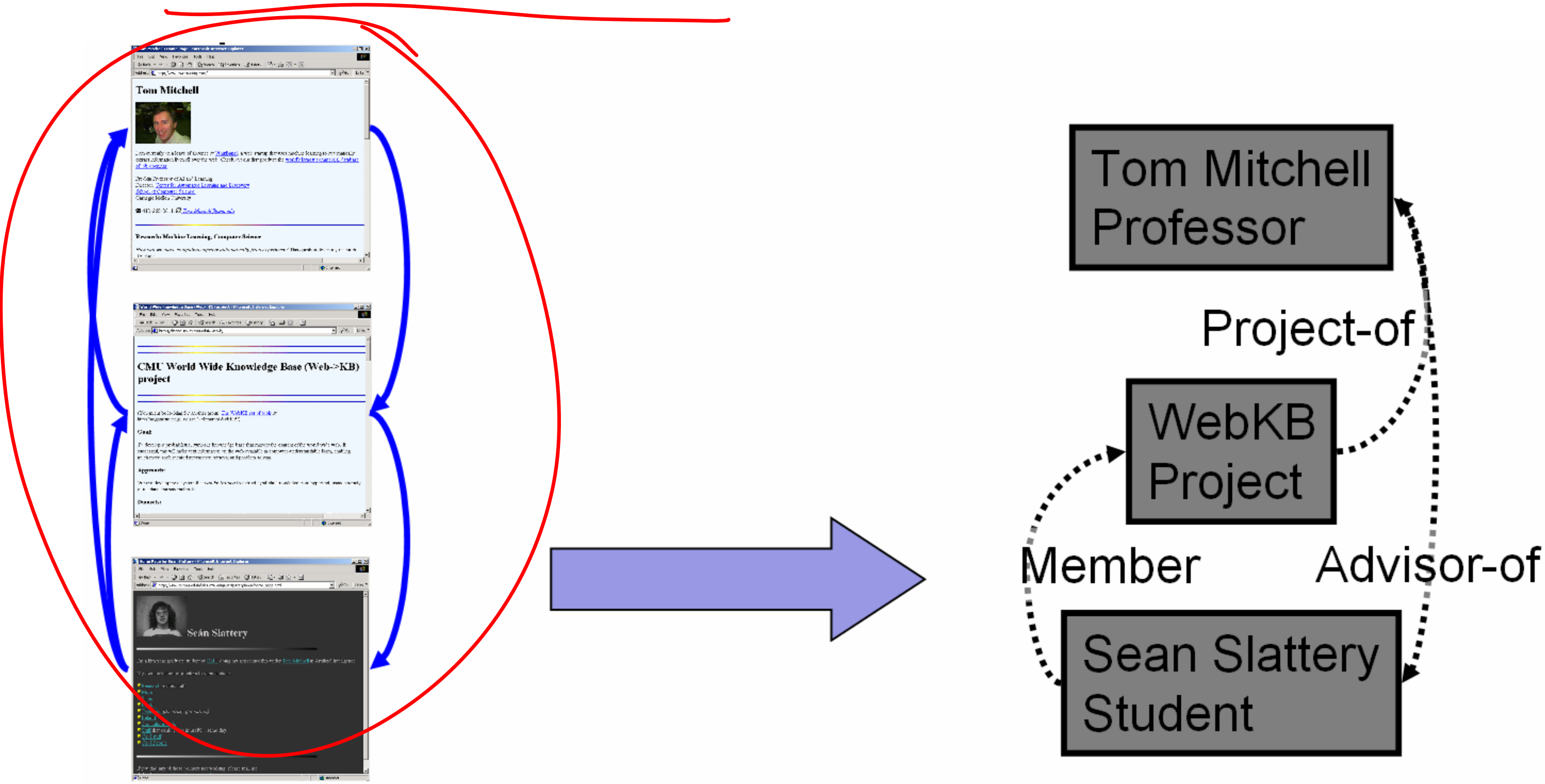




[Barnard et al.]

Structured data (text, webpages,…)

Probabilistic relational models



[Koller et al.]

And many

many

many

many

many

more…

Syllabus

- Covers a wide range of Probabilistic Graphical Models topics – from basic to state-of-the-art
- You will learn about the methods you heard about:
  - Bayesian networks, Markov networks, factor graphs, decomposable models, junction trees, parameter learning, structure learning, semantics, exact inference, variable elimination, context-specific independence, approximate inference, sampling, importance sampling, MCMC, Gibbs, variational inference, loopy belief propagation, generalized belief propagation, Kikuchi, Bayesian learning, missing data, EM, Chow-Liu, structure search, IPF for tabular MRFs, Gaussian and hybrid models, discrete and continuous variables, temporal and template models, hidden Markov Models, Forwards-Backwards, Viterbi, Baum-Welch, Kalman filter, linearization, switching Kalman filter, assumed density filtering, DBNs, BK, Relational probabilistic models, Causality,…
- Covers algorithms, theory and applications
- **It's going to be fun and hard work ☺**

Prerequisites

- 10-701 – Machine Learning, especially:
  - Probabilities
    - Distributions, densities, marginalization…
  - Basic statistics
    - Moments, typical distributions, regression…
- Algorithms
  - Dynamic programming, basic data structures, complexity…
- Programming
  - Matlab will be very useful
- We provide some background, but the class will be fast paced

- Ability to deal with “abstract mathematical concepts”

Review Sessions

- Very useful!
  - Review material
  - Present background
  - Answer questions
- Thursdays, 5:00-6:30 in Wean Hall 4615A
- First recitation is **tomorrow**
  - Review of probabilities & statistics
- Sometimes this semester: Especial recitations on Mondays 5:30-7pm in Wean Hall 4615A
  - Cover special topics that we can't cover in class
  - These are optional, but you are here to learn… ☺

- Do we need a Matlab review session?

Staff

- Two Great TAs: Great resource for learning, interact with them!
  - Khalid El-Arini <kbe@cs.cmu.edu>

  - Ajit Paul Singh <ajit@cs.cmu.edu>

- Administrative Assistant
  - Monica Hopes, Wean 4619, x8-5527, meh@cs.cmu.edu

First Point of Contact for HWs

- To facilitate interaction, a TA will be assigned to each homework question – This will be your "first point of contact" for this question
  - But, you can always ask any of us
  - (Due to logistic reasons, we will only start this policy for HW2)
- For e-mailing instructors, always use:
  - 10708-instructors@cs.cmu.edu
- For announcements, subscribe to:
  - 10708-announce@cs
  - https://mailman.srv.cs.cmu.edu/mailman/listinfo/10708-announce

Text Books

Something Else

- *Primary*: Daphne Koller and Nir Friedman, ~~**Bayesian Networks and Beyond**~~, in preparation. These chapters are part of the course reader. You can purchase one from Monica Hopes.
- *Secondary*: M. I. Jordan, **An Introduction to Probabilistic Graphical Models,** in preparation. Copies of selected chapters will be made available.

Grading

- 5 homeworks (50%)
  - First one goes today!
  - Homeworks are long and hard ☺
    - please, please, please, please, please, please start early!!!
- Final project (30%)
  - Done individually or in pairs
  - Details out October 4th
- Final (20%)
  - Take home, out Dec. 1st, due Dec. 15th

Homeworks

- Homeworks are hard, start early ☺
- Due in the beginning of class
- 3 late days for the semester
- After late days are used up:
  - Half credit within 48 hours
  - Zero credit after 48 hours
- All homeworks **must be handed in**, even for zero credit
- Late homeworks handed in to Monica Hopes, WEH 4619

- Collaboration
  - You may **discuss** the questions
  - Each student writes their own answers
  - Write on your homework anyone with whom you collaborate

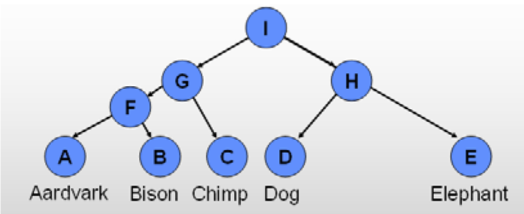
- **IMPORTANT**:
  - **We may use some material from previous years or from papers for the homeworks. Unless otherwise specified, please only look at the readings when doing your homework → You are taking this advanced graduate class because you want to learn, so this rule is self-enforced ☺**

Enjoy!

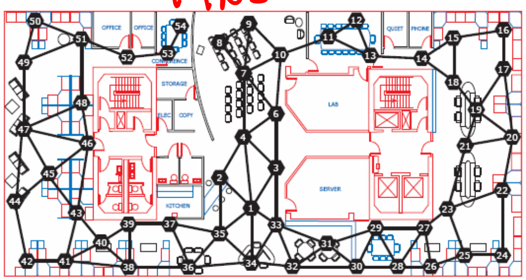
- Probabilistic graphical models are having significant impact in science, engineering and beyond
- This class should give you the basic foundation for applying GMs and developing new methods
- The fun begins…

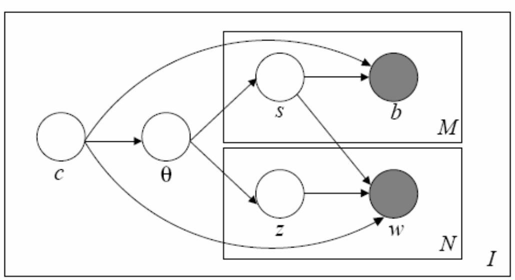
What are the fundamental questions of graphical models?

BNs



MNs

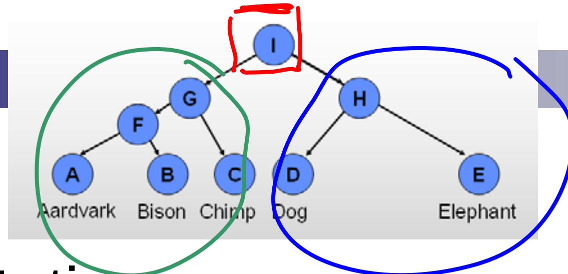


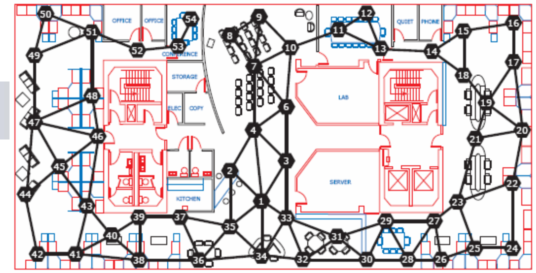


H.B.M

- Representation:
  - What are the types of models?
  - What does the model mean/imply/assume? (Semantics)
- Inference:
  - How do I answer questions/queries with my model?
- Learning:
  - What model is the right for my data?

More details???





- Representation:
  - Graphical models represent exponentially large probability distributions compactly
  - **Key concept**: *Conditional Independence*

n variables (binary) ⇒ $P(X_1, \ldots X_n)$ ∋ represented by $2^n - 1$

- Inference:
  - What is the probability of X given some observations?
  - What is the most likely explanation for what is happening?
  - What decisions should I make?

$P(B = \text{dead} \mid S = \text{false})$

- Learning:
  - What are the right/good parameters for the model?
  - How do I obtain the structure of the model?

Where do we start?

- From Bayesian networks
- "Complete" BN presentation first
  - Representation
  - Exact inference
  - Learning
  - Only discrete variables for now
- Later in the semester
  - Undirected models
  - Approximate inference
  - Continuous
  - Temporal models
  - And more…
- Class focuses on fundamentals – Understand the foundation and basic concepts

Today

- Probabilities
- Independence
- Two nodes make a BN
- Naïve Bayes

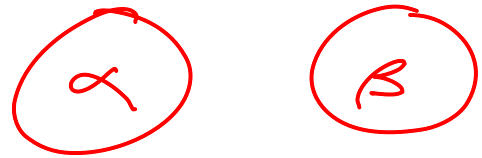
- Should be a review for everyone – Setting up notation for the class

Event spaces

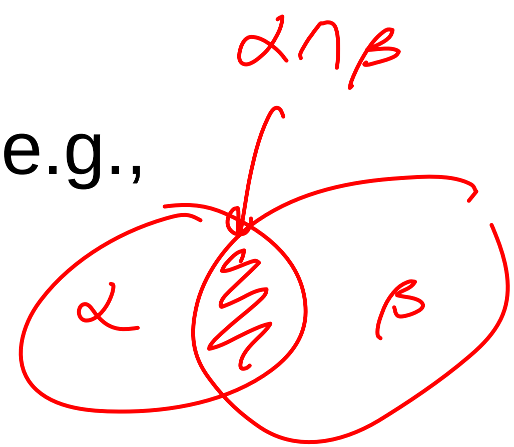
- Outcome space Ω = {1, 2, 3, 4, 5}
- Measurable events S
  - Each $\alpha \in S$ is a subset of Ω α = {1, 2}
- Must contain
  - Empty event $\emptyset$
  - Trivial event Ω
- Closed under
  - Union: $\alpha \cup \beta \in S$ {1, 2} $\in \Omega$ {2, 3} $\in \Omega$ $\Rightarrow$ {1, 2, 3} $\in \Omega$
  - Complement: $\alpha \in S$, then $\Omega - \alpha$ also in S

Probability distribution P over (Ω, S)

- $P(\alpha) \geq 0$
- $P(\Omega) = 1$
- If $\alpha \cap \beta = \emptyset$, then $P(\alpha \cup \beta) = P(\alpha) + P(\beta)$



- From here, you can prove a lot, e.g.,
  - $P(\emptyset) = 0$
  - $P(\alpha \cup \beta) = P(\alpha) + P(\beta) - P(\alpha \cap \beta)$



Interpretations of probability – A can of worms!

- Frequentists
  - P(α) is the frequency of α in the limit
  - Many arguments against this interpretation
    - What is the frequency of the event "it will rain tomorrow"?

flip HHTHT.... $\lim_{\#flips \to \infty} \frac{\#head}{\#flips} = \theta$

θ " P(H)

9/14/06

"Bayesians"

- Subjective interpretation
  - P(α) is my degree of belief that α will happen
  - What the …. does "degree of belief mean"?
  - If I say P(α)=0.8, then I am willing to bet!!!

I give \$80 if $\neg\alpha$

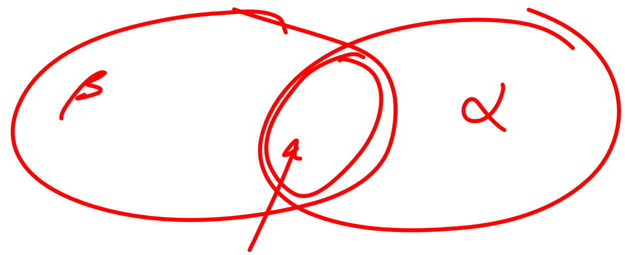
you give \$20 if α

- For this class, we (mostly) don't care what camp you are in

Conditional probabilities

- After learning that α is true, how do we feel about β?
- $P(\beta|\alpha)$

area



Two of the most important rules of the semester: 1. The chain rule

- $P(\alpha \cap \beta) = P(\alpha) P(\beta|\alpha)$ $= P(\beta) \cdot P(\alpha|\beta)$

$\{1,2\}$ $\{2,3\}$

- More generally:
  - $P(\alpha_1 \cap \ldots \cap \alpha_k) = \underline{P(\alpha_1)}\ \underline{P(\alpha_2|\alpha_1)} \cdots P(\alpha_k|\underline{\alpha_1 \cap \ldots \cap \alpha_{k-1}})$

$= P(\alpha_3) \cdot P(\alpha_7|\alpha_3) \cdot P(\alpha_4|\alpha_3, \alpha_7) \ldots$

any order

Two of the most important rules of the semester: 2. Bayes rule

- $$P(\alpha \mid \beta) = \frac{P(\beta \mid \alpha) P(\alpha)}{P(\beta)}$$

- More generally, external event γ:
  - $$P(\alpha \mid \beta \cap \gamma) = \frac{P(\beta \mid \alpha \cap \gamma) P(\alpha \mid \gamma)}{P(\beta \mid \gamma)}$$

Most important concept: a) Independence

pizza rain

- α and β ***independent***, if $P(\beta|\alpha)=P(\beta)$
  - $P \vDash (\alpha \perp \beta)$

dist. entails α indep. of β

- **Proposition:** α and β *independent* if and only if $P(\alpha \cap \beta)=P(\alpha)P(\beta)$

Most important concept: b) Conditional independence

- Independence is rarely true, but conditionally…

- α and β ***conditionally independent*** given γ if $P(\beta|\alpha\cap\gamma)=P(\beta|\gamma)$
  - $P \models (\alpha \perp \beta \mid \gamma)$

    α indep. β given γ

Proposition: $P \models (\alpha \perp \beta \mid \gamma)$ if and only if

$$P(\alpha\cap\beta \mid\gamma)=P(\alpha \mid\gamma)P(\beta \mid\gamma)$$

Random variable

- Events are complicated – we think about attributes
  - Age, Grade, HairColor
- Random variables formalize attributes:
  - Grade=A$^+$ shorthand for event $\{\omega \in \Omega : f_{Grade}(\omega) = A^+\}$

$\{A^+, A, B^+, \}$

- Properties of random vars, X:
  - Val(X) = possible values of random var X $\in \{$Works, Dead$\}$ $\in \mathbb{R}$
  - For discrete (categorical): $\sum_{i=1\ldots|Val(X)|} P(X=x_i) = 1$
  - For continuous: $\int_x p(X=x)dx = 1$

$X = x$

R.V.

Value

Marginal distribution

- Probability P(X) of possible outcomes X

$$P(Grade=A) = \sum_a \sum_h P(G=A, Age=a, H=h)$$

$$P(G, A, H)$$

Joint distribution, Marginalization

- Two random variables – Grade & Intelligence

$G = \{A, B\}$

$I = \{H, VH\}$

$P(G, I) =$

I \ G	A	B
H	0.04	0.01
VH	0.8	0.15

- Marginalization – Compute marginal over single var

$$P(G): \left\{ \begin{array}{l} P(G=A) = 0.8 + 0.04 = P(G=A, I=H) + P(G=A, I=VH) \\ \qquad = 0.84 \\ P(G=B) = 0.16 \end{array} \right.$$

Marginalization – The general case

- Compute marginal distribution $P(X_i)$:

$P(X_1, \ldots, X_n)$ if all X_i binary?

$$\underline{P(X_1, X_2, \ldots, X_i)} = \underbrace{\sum_{x_{i+1},\ldots,x_n}}_{2^{n-i}} P(X_1, X_2, \ldots, X_i, x_{i+1}, \ldots, x_n)$$

2^{n-i} for each $x_1 \ldots x_i$

$$P(X_i) = \underbrace{\sum_{x_1,\ldots,x_{i-1}}}_{2^{i-1}} P(x_1, \ldots, x_{i-1}, X_i)$$

2^{i-1} for each x_i

Basic concepts for random variables

- Atomic outcome: assignment $x_1,\ldots,x_n$ to $X_1,\ldots,X_n$
- Conditional probability: $P(X,Y)=P(X)P(Y|X)$

  $P(Y|X) : P(Y=y \mid \cancel{X}=x)$
- Bayes rule: $P(X|Y)= \dfrac{P(Y|X) \cdot P(X)}{P(Y)}$
- Chain rule:
  - $P(X_1,\ldots,X_n) = \underline{P(X_1)}\,\underline{P(X_2|X_1)}\cdots P(X_k|X_1,\ldots,X_{k-1})\cdots \underline{P(X_n|X_1,\ldots,X_{n-1})}$

Conditionally independent random variables

$\in \{G, I\}$ $\in \{\text{Admission}\}$ $\in \{SAT\}$

- **Sets** of variables **X**, **Y**, **Z**
- X is independent of Y given Z if

  $P(X=x \mid Y=y, Z=z) = P(X=x \mid Z=z)$
  - $P \vDash (\mathbf{X}=\mathbf{x} \perp \mathbf{Y}=\mathbf{y} \mid \mathbf{Z}=\mathbf{z}),\ \forall\ \mathbf{x} \in Val(\mathbf{X}),\ \mathbf{y} \in Val(\mathbf{Y}),\ \mathbf{z} \in Val(\mathbf{Z})$

- Shorthand:
  - **Conditional independence:** $P \vDash (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ X indep. Y given Z
  - For $P \vDash (\mathbf{X} \perp \mathbf{Y} \mid \emptyset)$, write $P \vDash (\mathbf{X} \perp \mathbf{Y})$

    pizza rain

- **Proposition:** *P* statisfies $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ if and only if
  - $P(\mathbf{X},\mathbf{Y}|\mathbf{Z}) = P(\mathbf{X}|\mathbf{Z})\ P(\mathbf{Y}|\mathbf{Z})$

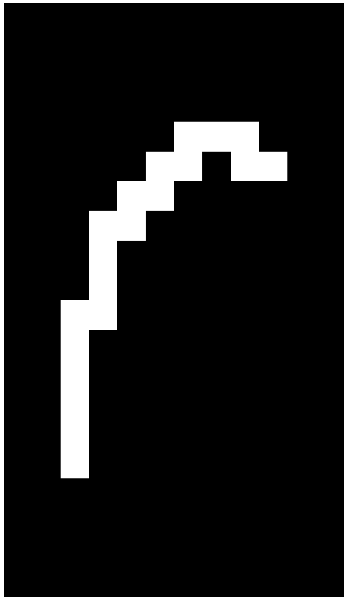
Properties of independence

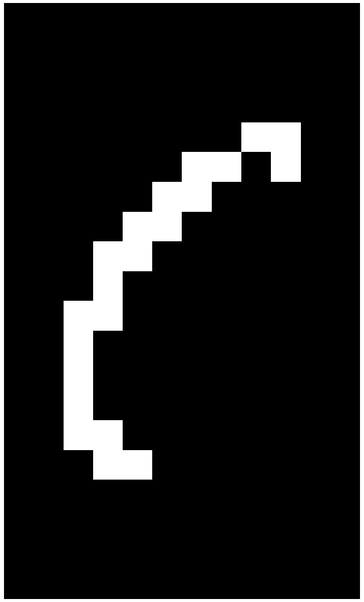
- **Symmetry:**
  - $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) \Rightarrow (\mathbf{Y} \perp \mathbf{X} \mid \mathbf{Z})$
- **Decomposition:**
  - $(\mathbf{X} \perp \mathbf{Y},\mathbf{W} \mid \mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$
- **Weak union:**
  - $(\mathbf{X} \perp \mathbf{Y},\mathbf{W} \mid \mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z},\mathbf{W})$
- **Contraction:**
  - $(\mathbf{X} \perp \mathbf{W} \mid \mathbf{Y},\mathbf{Z})$ & $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y},\mathbf{W} \mid \mathbf{Z})$
- **Intersection:**
  - $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{W},\mathbf{Z})$ & $(\mathbf{X} \perp \mathbf{W} \mid \mathbf{Y},\mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y},\mathbf{W} \mid \mathbf{Z})$
  - Only for positive distributions!
  - $P(\alpha)>0, \forall\alpha, \alpha\neq\emptyset$
- **Notation:** $I(P)$ – independence properties entailed by P

Bayesian networks

- One of the most exciting recent advancements in statistical AI
- Compact representation for exponentially-large probability distributions
- Fast marginalization too
- Exploit conditional independencies

Handwriting recognition





Webpage classification

→

Company home page

vs

Personal home page

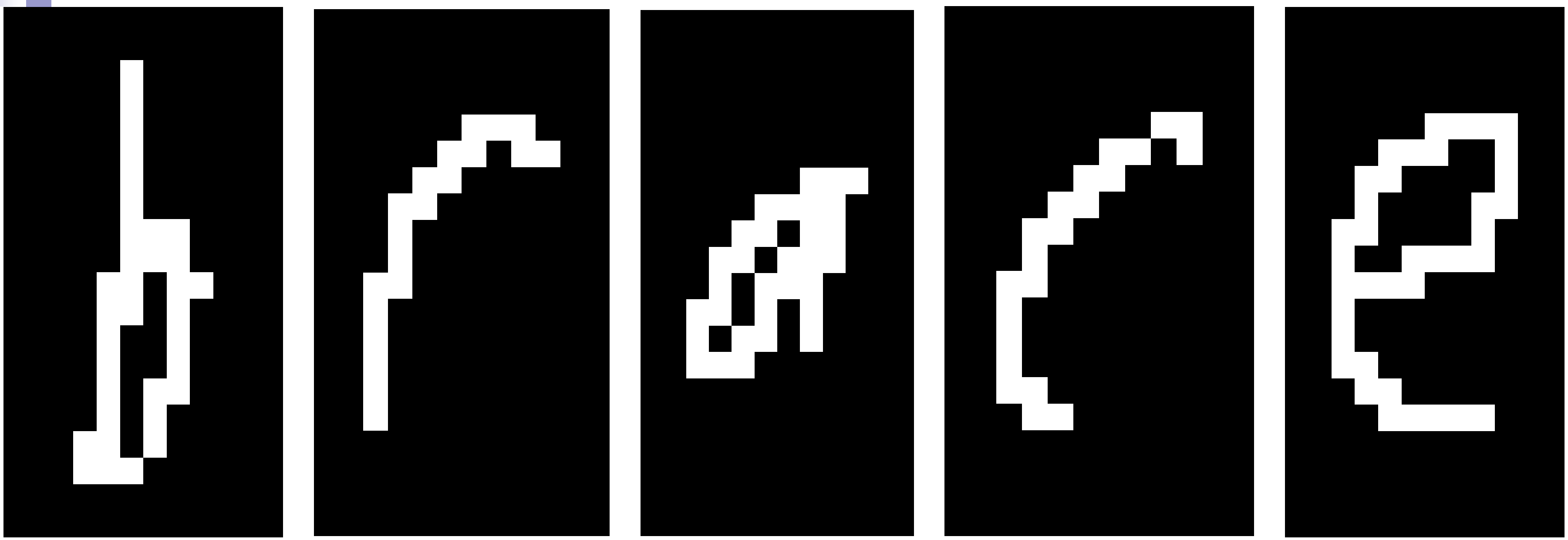
vs

Univeristy home page

vs

…

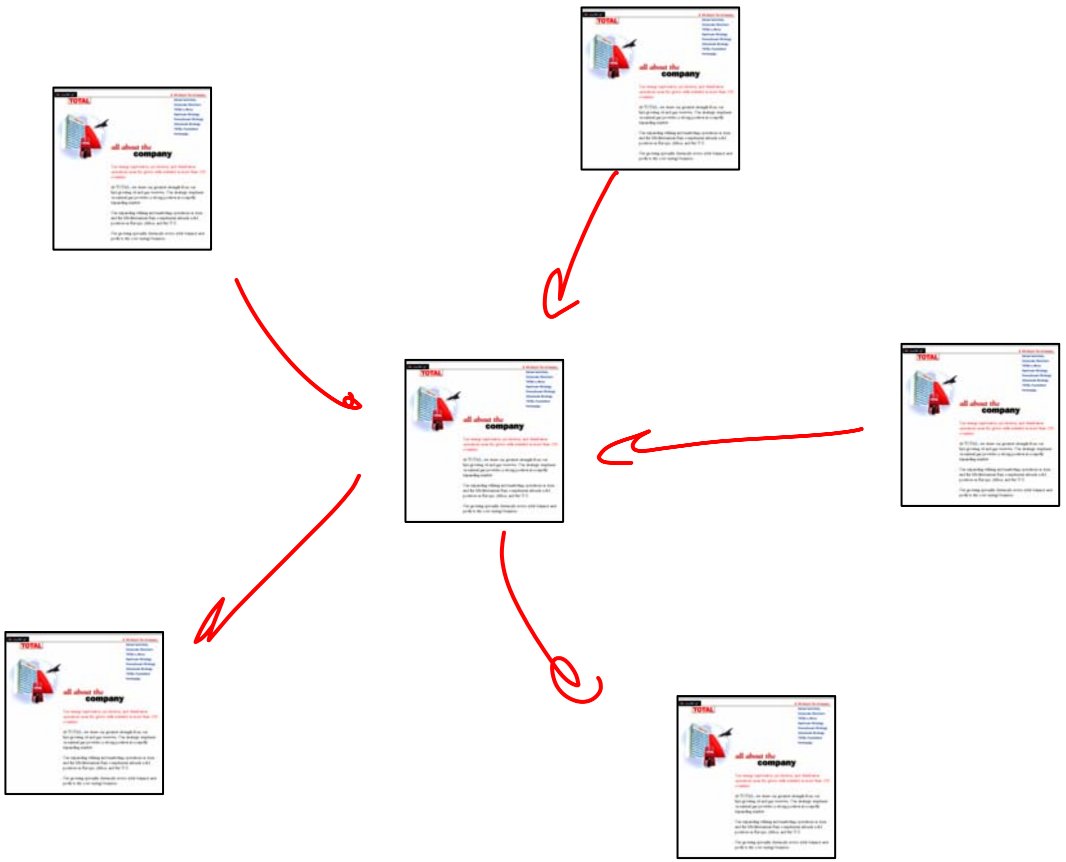
Handwriting recognition 2



$X_1 = \{a \ldots z\}$ $X_2 = \{a \ldots z\}$ X_3 X_4 X_5

$$P(X_2 = r) = \sum_{X_1 = a,\ldots,z} \; \sum_{x_3, x_4, x_5} P(X_1 = x_1, X_2 = r, x_3, x_4, x_5)$$

Webpage classification 2



Let's start on BNs…

- Consider $P(X_i)$ ← $k-1$ parms.
  - Assign probability to each $x_i \in Val(X_i)$
  - $k-$ Independent parameters
- Consider $P(X_1,\ldots,X_n)$
  - How many independent parameters if $|Val(X_i)|=k$?

$k^n - 1$ parameters

What if variables are independent?

- What if variables are independent?
  - $(X_i \perp X_j)$, $\forall$ i,j
  - Not enough!!! (See homework 1 ☺)
  - Must assume that $(\mathbf{X} \perp \mathbf{Y})$, $\forall$ **X**,**Y** subsets of $\{X_1,\ldots,X_n\}$

$(\{X_1, X_3\} \perp \{X_7, X_9\})$

- Can write
  - $P(X_1,\ldots,X_n) = \prod_{i=1\ldots n} P(X_i)$
- How many independent parameters now?

$n \cdot (K-1)$ params

Conditional parameterization – two nodes

- Grade is determined by Intelligence

Conditional parameterization – three nodes

- Grade and SAT score are determined by Intelligence
- $(G \perp S \mid I)$

The naïve Bayes model – Your first real Bayes Net

- Class variable: C
- Evidence variables: $X_1,\ldots,X_n$
- assume that $(\mathbf{X} \perp \mathbf{Y} \mid C)$, $\forall\, \mathbf{X},\mathbf{Y}$ subsets of $\{X_1,\ldots,X_n\}$

What you need to know

- Basic definitions of probabilities
- Independence
- Conditional independence
- The chain rule
- Bayes rule
- Naïve Bayes

Next class

- We’ve heard of Bayes nets, we’ve played with Bayes nets, we’ve even used them in your research
- Next class, we’ll learn the semantics of BNs, relate them to independence assumptions encoded by the graph