

Readings:

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

K&F: 3.1

Introduction

Graphical Models – 10708

Carlos Guestrin

Carnegie Mellon University

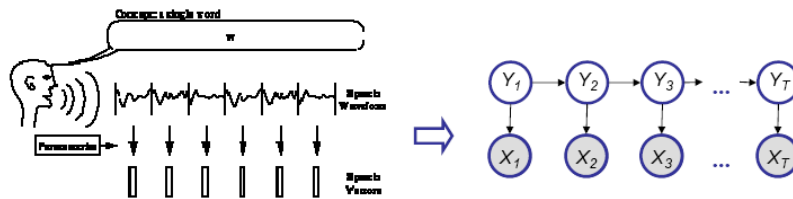
September 13th, 2006

**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

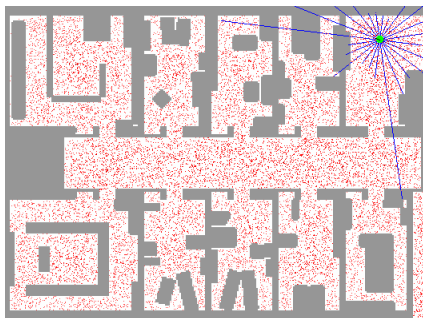


10-708 - ©Carlos Guestrin 2006

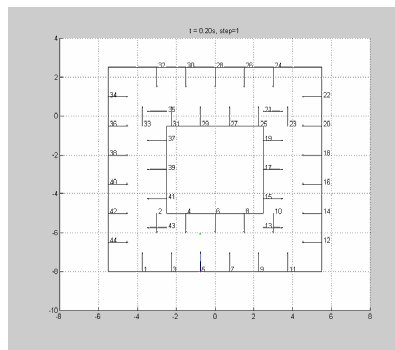
3

Tracking and robot localization

Kalman Filters



[Fox et al.]



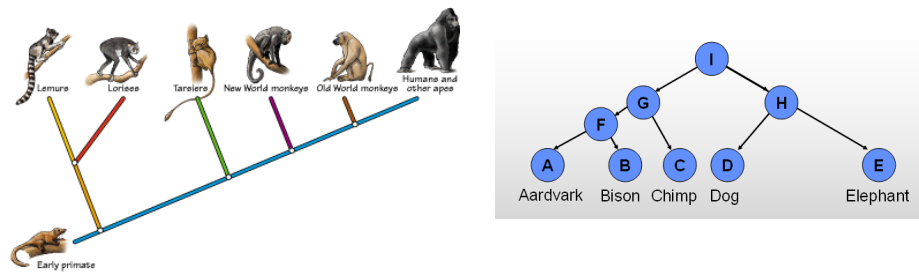
[Funiak et al.]

10-708 - ©Carlos Guestrin 2006

4

Evolutionary biology

Bayesian networks



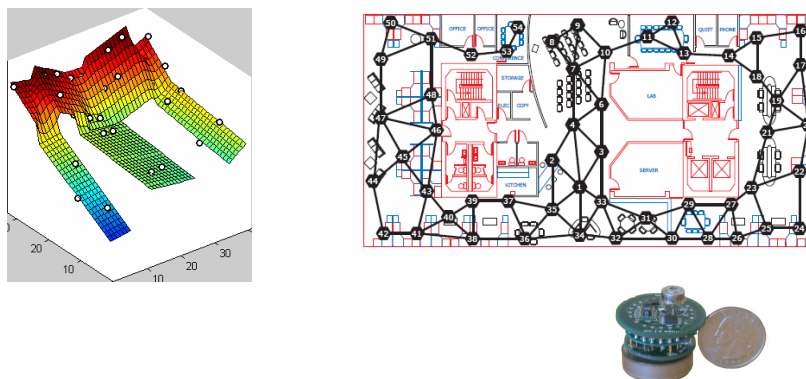
[Friedman et al.]

10-708 - ©Carlos Guestrin 2006

5

Modeling sensor data

Undirected graphical models



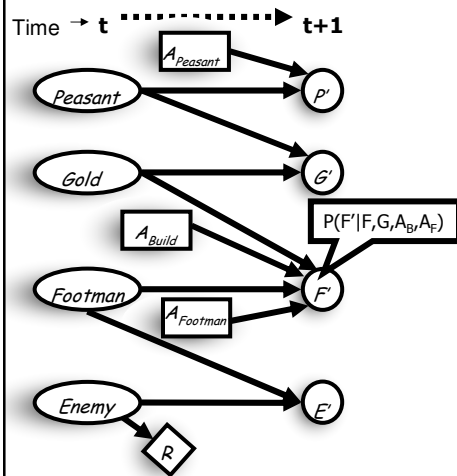
[Guestrin et al.]

10-708 - ©Carlos Guestrin 2006

6

Planning under uncertainty

Dynamic Bayesian networks
Factored Markov decision problems



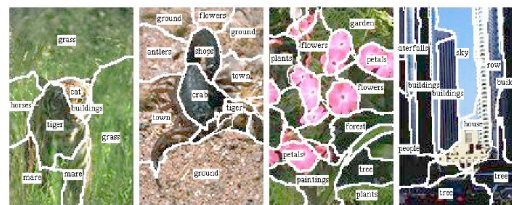
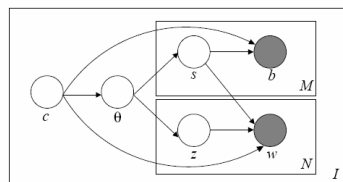
[Guestrin et al.]

10-708 - ©Carlos Guestrin 2006

7

Images and text data

Hierarchical Bayesian models



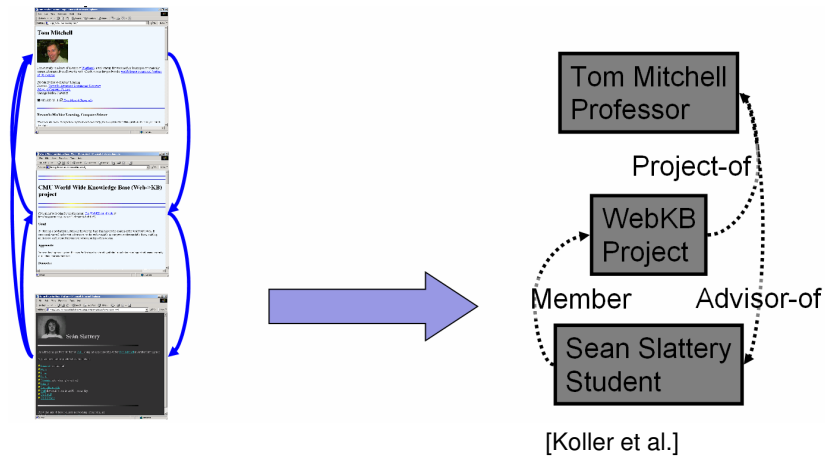
[Barnard et al.]

10-708 - ©Carlos Guestrin 2006

8

Structured data (text, webpages,...)

Probabilistic relational models



10-708 – ©Carlos Guestrin 2006

9

And many

many

many

many

many

more...

10-708 – ©Carlos Guestrin 2006

10

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 ☺**

10-708 – ©Carlos Guestrin 2006

11

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”

10-708 – ©Carlos Guestrin 2006

12




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?

10-708 - ©Carlos Guestrin 2006

13

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 

10-708 - ©Carlos Guestrin 2006

14

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>

10-708 – ©Carlos Guestrin 2006

15

Text Books

- *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.

10-708 – ©Carlos Guestrin 2006

16

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

10-708 – ©Carlos Guestrin 2006

17

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 ☺

10-708 – ©Carlos Guestrin 2006

18

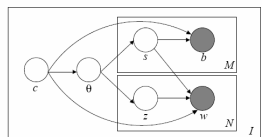
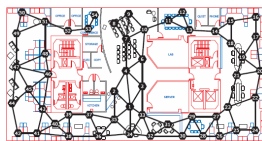
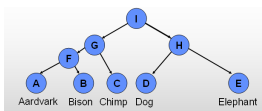
Enjoy!

- 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...

10-708 - ©Carlos Guestrin 2006

19

What are the fundamental questions of graphical models?

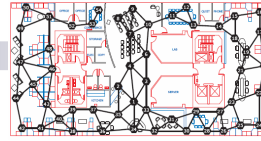
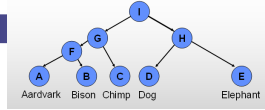


- Representation:
 - What are the types of models?
 - What does the model mean/imply/assume? (Semantics)
- Inference:
 - How do I answer questions/queries with my model?
- Learning:
 - What model is the right for my data?

10-708 - ©Carlos Guestrin 2006

20

More details???



- Representation:
 - Graphical models represent exponentially large probability distributions compactly
 - **Key concept:** *Conditional Independence*
- 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?
- Learning:
 - What are the right/good parameters for the model?
 - How do I obtain the structure of the model?

10-708 – ©Carlos Guestrin 2006

21

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

10-708 – ©Carlos Guestrin 2006

22

Today

- Probabilities
 - Independence
 - Two nodes make a BN
 - Naïve Bayes
-
- Should be a review for everyone – Setting up notation for the class

10-708 – ©Carlos Guestrin 2006

23

Event spaces

- Outcome space Ω
- Measurable events S
 - Each $\alpha \in S$ is a subset of Ω
- Must contain
 - Empty event \emptyset
 - Trivial event Ω
- Closed under
 - Union: $\alpha \cup \beta \in S$
 - Complement: $\alpha \in S$, then $\Omega - \alpha$ also in S

10-708 – ©Carlos Guestrin 2006

24

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.,
 - $P(\emptyset) = 0$
 - $P(\alpha \cup \beta) = P(\alpha) + P(\beta) - P(\alpha \cap \beta)$

10-708 – ©Carlos Guestrin 2006

25

Interpretations of probability – A can of worms!

- Frequentists
 - $P(\alpha)$ is the frequency of α in the limit
 - Many arguments against this interpretation
 - What is the frequency of the event “it will rain tomorrow”?
- Subjective interpretation
 - $P(\alpha)$ is my degree of belief that α will happen
 - What the does “degree of belief mean?”
 - If I say $P(\alpha) = 0.8$, then I am willing to bet!!!

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

10-708 – ©Carlos Guestrin 2006

26

Conditional probabilities

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

10.708 – ©Carlos Guestrin 2006

27

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

- $P(\alpha \cap \beta) = P(\alpha)P(\beta|\alpha)$
- More generally:
 - $P(\alpha_1 \cap \dots \cap \alpha_k) = P(\alpha_1) P(\alpha_2|\alpha_1) \dots P(\alpha_k|\alpha_1 \cap \dots \cap \alpha_{k-1})$

10.708 – ©Carlos Guestrin 2006

28

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 even γ :
 - $P(\alpha \mid \beta \cap \gamma) = \frac{P(\beta \mid \alpha \cap \gamma)P(\alpha \mid \gamma)}{P(\beta \mid \gamma)}$

10.708 – ©Carlos Guestrin 2006

29

Most important concept: a) Independence

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

10.708 – ©Carlos Guestrin 2006

30

Most important concept:

b) Conditional independence

- Independence is rarely true, but conditionally...
- α and β **conditionally independent** given γ if
$$P(\beta|\alpha\cap\gamma)=P(\beta|\gamma)$$
 - $P \models (\alpha \perp \beta \mid \gamma)$

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

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

10-708 – ©Carlos Guestrin 2006

31

Random variable

- Events are complicated – we think about attributes
 - Age, Grade, HairColor
- Random variables formalize attributes:
 - Grade=A shorthand for event $\{\omega \in \Omega: f_{\text{Grade}}(\omega) = A\}$
- Properties of random vars, X :
 - $\text{Val}(X)$ = possible values of random var X
 - For discrete (categorical): $\sum_{i=1 \dots |\text{Val}(X)|} P(X=x_i) = 1$
 - For continuous: $\int_x p(X=x)dx = 1$

10-708 – ©Carlos Guestrin 2006

32

Marginal distribution

- Probability $P(X)$ of possible outcomes X

10-708 – ©Carlos Guestrin 2006

33

Joint distribution, Marginalization

- Two random variables – Grade & Intelligence

- Marginalization – Compute marginal over single var

10-708 – ©Carlos Guestrin 2006

34

Marginalization – The general case

- Compute marginal distribution $P(X_i)$:

$$P(X_1, X_2, \dots, X_i) = \sum_{x_{i+1}, \dots, x_n} P(X_1, X_2, \dots, X_i, x_{i+1}, \dots, x_n)$$

$$P(X_i) = \sum_{x_1, \dots, x_{i-1}} P(x_1, \dots, x_{i-1}, X_i)$$

10-708 – ©Carlos Guestrin 2006

35

Basic concepts for random variables

- Atomic outcome: assignment x_1, \dots, x_n to X_1, \dots, X_n
- Conditional probability: $P(X, Y) = P(X)P(Y|X)$
- Bayes rule: $P(X|Y) =$
- Chain rule:
 - $P(X_1, \dots, X_n) = P(X_1)P(X_2|X_1) \dots P(X_k|X_1, \dots, X_{k-1})$

10-708 – ©Carlos Guestrin 2006

36

Conditionally independent random variables

- **Sets** of variables $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$
- \mathbf{X} is independent of \mathbf{Y} given \mathbf{Z} if
 - $P \models (\mathbf{X} \perp \mathbf{Y} | \mathbf{Z})$, $\forall \mathbf{x} \in \text{Val}(\mathbf{X}), \mathbf{y} \in \text{Val}(\mathbf{Y}), \mathbf{z} \in \text{Val}(\mathbf{Z})$
- Shorthand:
 - **Conditional independence:** $P \models (\mathbf{X} \perp \mathbf{Y} | \mathbf{Z})$
 - For $P \models (\mathbf{X} \perp \mathbf{Y} | \emptyset)$, write $P \models (\mathbf{X} \perp \mathbf{Y})$
- **Proposition:** P satisfies $(\mathbf{X} \perp \mathbf{Y} | \mathbf{Z})$ if and only if
 - $P(\mathbf{X}, \mathbf{Y} | \mathbf{Z}) = P(\mathbf{X} | \mathbf{Z}) P(\mathbf{Y} | \mathbf{Z})$

10-708 – ©Carlos Guestrin 2006

37

Properties of independence

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

10-708 – ©Carlos Guestrin 2006

38

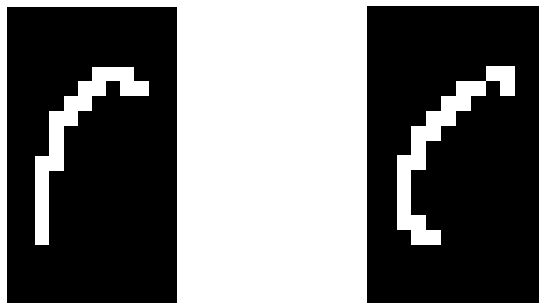
Bayesian networks

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

10-708 – ©Carlos Guestrin 2006

39

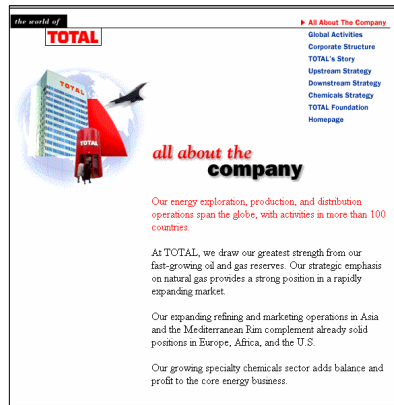
Handwriting recognition



10-708 – ©Carlos Guestrin 2006

40

Webpage classification



Company home page

vs

Personal home page

vs

Univeristy home page

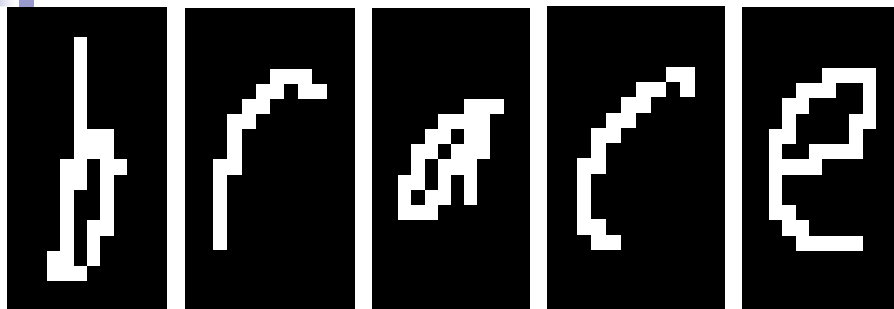
vs

...

10-708 - ©Carlos Guestrin 2006

41


Handwriting recognition 2



10-708 - ©Carlos Guestrin 2006

42

Webpage classification 2



10-708 – ©Carlos Guestrin 2006 43

Let's start on BNs...

- Consider $P(X_i)$
 - Assign probability to each $x_i \in \text{Val}(X_i)$
 - Independent parameters
- Consider $P(X_1, \dots, X_n)$
 - How many independent parameters if $|\text{Val}(X_i)|=k$?

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 $(\mathbf{X} \perp \mathbf{Y}), \forall \mathbf{X}, \mathbf{Y}$ subsets of $\{X_1, \dots, X_n\}$
- Can write
 - $P(X_1, \dots, X_n) = \prod_{i=1 \dots n} P(X_i)$
- How many independent parameters now?

10-708 – ©Carlos Guestrin 2006

45

Conditional parameterization – two nodes

- Grade is determined by Intelligence

10-708 – ©Carlos Guestrin 2006

46

Conditional parameterization – three nodes

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

10-708 – ©Carlos Guestrin 2006

47

The naïve Bayes model – Your first real Bayes Net

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

10-708 – ©Carlos Guestrin 2006

48

What you need to know

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

10-708 – ©Carlos Guestrin 2006

49

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

10-708 – ©Carlos Guestrin 2006

50