



# The Basics

Graphical Models – 10708

Carlos Guestrin

Carnegie Mellon University

September 12<sup>th</sup>, 2005

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

# Probability distribution $P$ over $(\Omega, \mathcal{S})$

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

- $P(\alpha)$  is the frequency of  $\alpha$  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  $\alpha$  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

# Conditional probabilities



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

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

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

- More generally:

- $P(\alpha_1, \dots, \alpha_k) = P(\alpha_1) P(\alpha_2|\alpha_1) \dots P(\alpha_k|\alpha_1 \cap \dots \cap \alpha_{k-1})$



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

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

- More generally:

- $$P(\alpha | \beta \cap \gamma) = \frac{P(\beta | \alpha \cap \gamma)P(\alpha | \gamma)}{P(\beta | \gamma)}$$

# Most important concept:

## a) Independence

- $\alpha$  and  $\beta$  *independent*, if  $P(\beta|\alpha)=P(\beta)$ 
  - $P \models (\alpha \perp \beta)$
  
- **Proposition:**  $\alpha$  and  $\beta$  *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...
- $\alpha$  and  $\beta$  ***conditionally independent*** given  $\gamma$  if
$$P(\beta|\alpha\cap\gamma)=P(\beta|\gamma)$$
  - $P \models (\alpha \perp \beta | \gamma)$

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

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

# Random variable


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

# Marginal distribution



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

# Joint distribution, Marginalization



- Two random variables – Grade & Intelligence
- Marginalization – Compute marginal over single var

# Marginalization – The general case

- Compute marginal distribution  $P(X_i)$ :

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

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

# 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:
  - $P(X_1, \dots, X_n) = P(X_1)P(X_2|X_1) \cdots P(X_k|X_1, \dots, X_{k-1})$



# Conditionally independent random variables

- **Sets** of variables  $\mathbf{X}$ ,  $\mathbf{Y}$ ,  $\mathbf{Z}$
- $X$  is independent of  $Y$  given  $Z$  if
  - $P \models (\mathbf{X}=\mathbf{x} \perp \mathbf{Y}=\mathbf{y} | \mathbf{Z}=\mathbf{z}), \forall \mathbf{x} \in \text{Val}(\mathbf{X}), \mathbf{y} \in \text{Val}(\mathbf{Y}), \mathbf{z} \in \text{Val}(\mathbf{Z})$
- 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
  - $P(\mathbf{X}, \mathbf{Y} | \mathbf{Z}) = P(\mathbf{X} | \mathbf{Z}) P(\mathbf{Y} | \mathbf{Z})$

# Properties of independence

- **Symmetry:**

- $(X \perp Y \mid Z) \Rightarrow (Y \perp X \mid Z)$

- **Decomposition:**

- $(X \perp Y, W \mid Z) \Rightarrow (X \perp Y \mid Z)$

- **Weak union:**

- $(X \perp Y, W \mid Z) \Rightarrow (X \perp Y \mid Z, W)$

- **Contraction:**

- $(X \perp W \mid Y, Z) \& (X \perp Y \mid Z) \Rightarrow (X \perp Y, W \mid Z)$

- **Intersection:**

- $(X \perp Y \mid W, Z) \& (X \perp W \mid Y, Z) \Rightarrow (X \perp Y, W \mid Z)$

- Only for positive distributions!

- $P(\alpha) > 0, \forall \alpha, \alpha \neq \emptyset$

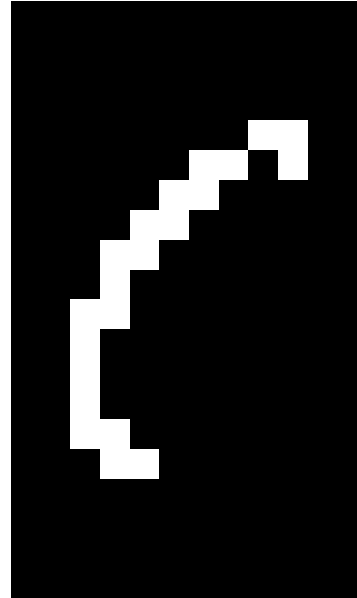
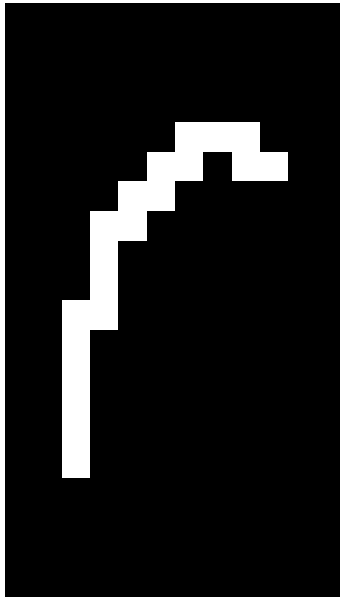
- **Notation:**  $I(P)$  – independence properties entailed by  $P$

# Bayesian networks



- One of the most exciting advancements in statistical AI in the last 10-15 years
- Compact representation for exponentially-large probability distributions
- Fast marginalization too
- Exploit conditional independencies

# Handwriting recognition



# Webpage classification



Company home page

VS

Personal home page

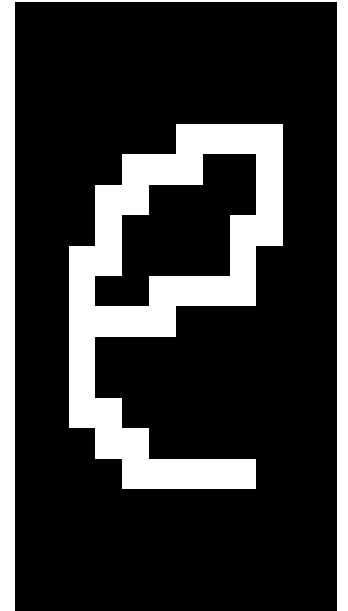
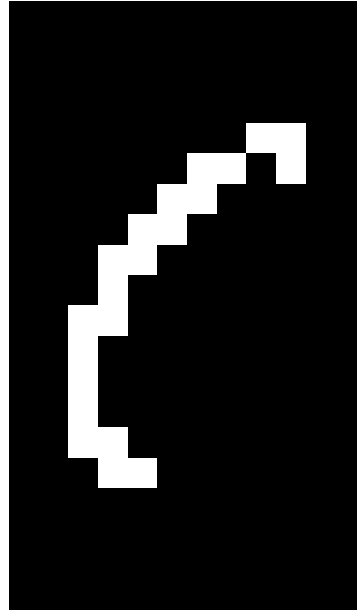
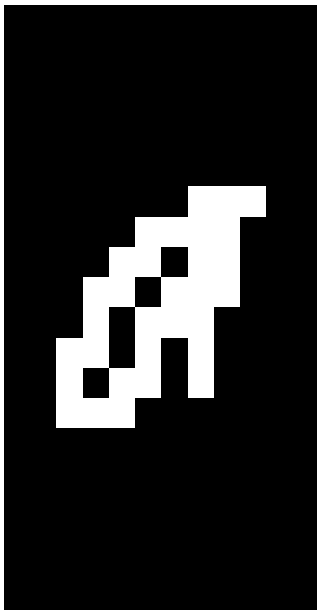
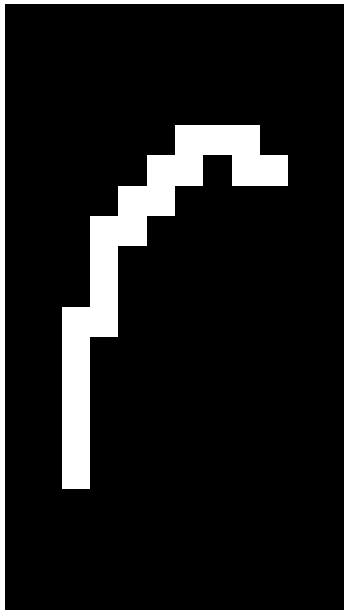
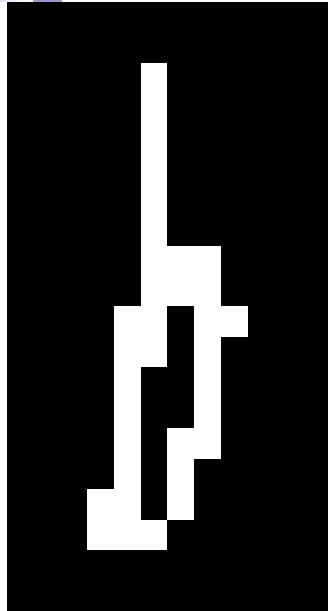
VS

Univeristy home page

VS

...

# Handwriting recognition 2



# Webpage classification 2



# 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?

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

# 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