Statistical Methods in AI and ML

Nicholas Ruozzi
University of Texas at Dallas

The Course

One of the **most exciting** advances in AI/ML in the last decade

Judea Pearl won the Turing award for his work on Bayesian networks!

(among other achievements)

Prob. Graphical Models

Exploit **locality** and structural features of a given model in order to gain insight about **global properties**

The Course

- What this course is:
  - Probabilistic graphical models
  - Topics:
    - representing data
    - exact and approximate statistical inference
    - model learning
    - variational methods in ML

The Course

- What you should be able to do at the end:
  - Design statistical models for applications in your domain of interest
  - Apply learning and inference algorithms to solve real problems (exactly or approximately)
  - Understand the complexity issues involved in the modeling decisions and algorithmic choices

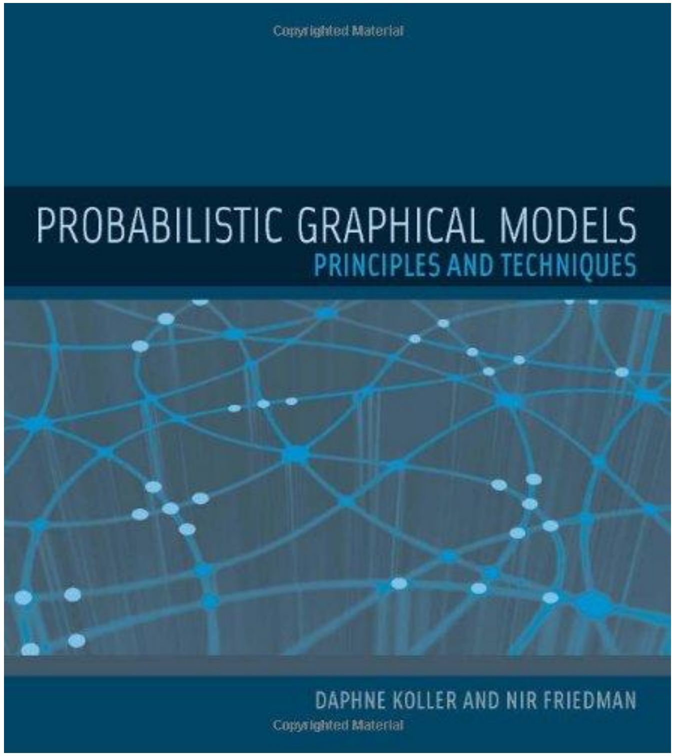
Prerequisites

- CS 5343: Algorithm Analysis and Data Structures
- CS 3341: Probability and Statistics in Computer Science and Software Engineering
- Basically, comfort with probability and algorithms (machine learning is helpful, but not required)

Textbook

Readings will be posted online before each lecture

Check the course website for additional resources and papers



Grading

- 4-6 problem sets (70%)
  - See collaboration policy on the web
- Final project (25%)
- Class participation & extra credit (5%)

-subject to change-

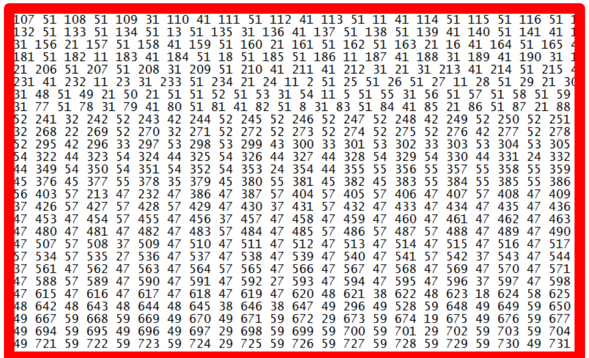
Course Info.

- Instructor: Nicholas Ruozzi
  - Office: ECSS 3.409
  - Office hours: Tues. 11am - 12pm and by appointment
- TA: TBD
  - Office hours and location TBD
- Course website: http://www.utdallas.edu/~nrr150130/cs6347/2016sp/

Main Ideas

- Model the world (or at least the problem) as a collection of random variables related through some joint probability distribution
  - Compactly represent the distribution
  - Undirected graphical models
  - Directed graphical models
- Learn the distribution from observed data
  - Maximum likelihood, SVMs, etc.
- Make predictions (statistical inference)

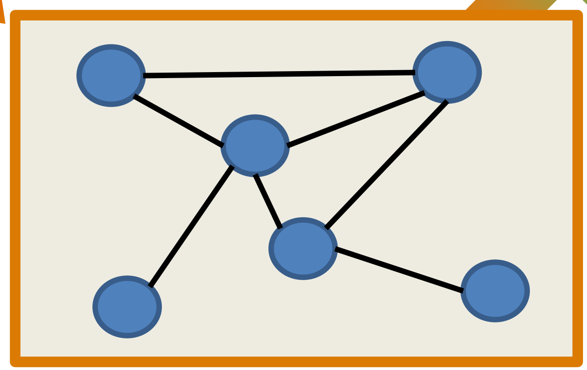
Inference and Learning



Collect Data

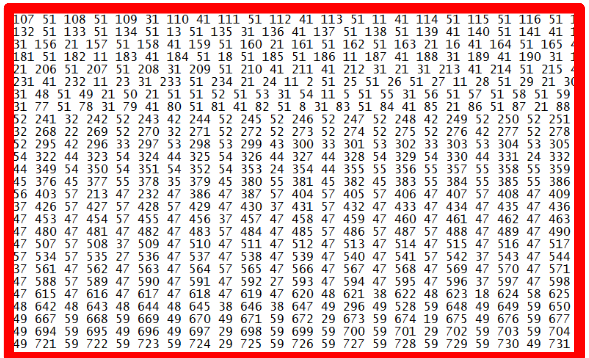
$$Z(\theta) = \sum_x p(x;\theta)$$

Use the model to do inference / make predictions

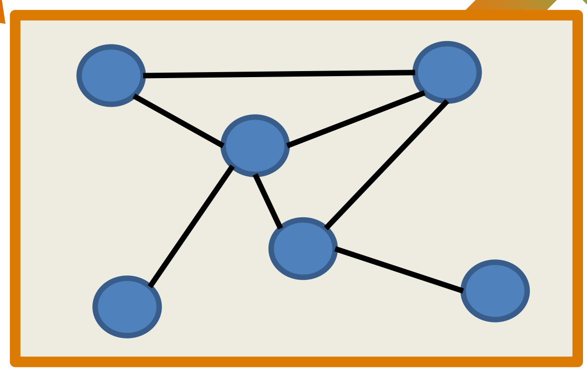


“Learn” a model that represents the observed data

Inference and Learning



Data sets can be large



Data must be compactly modeled

$$Z(\theta) = \sum_{x} p(x; \theta)$$

Inference needs to be fast

Applications

- Computer vision
- Natural language processing
- Robotics
- Computational biology
- Computational neuroscience
- Text translation
- Text-to-speech
- Many more…

Graphical Models

- A graphical model is a graph together with "local interactions"
- The graph and interactions model a global optimization or learning problem
- The study of graphical models is concerned with how to exploit local structure to solve these problems either exactly or approximately

Probability Review

Discrete Probability

- **Sample space** specifies the set of possible outcomes
  - For example, $\Omega = \{\mathrm{H}, \mathrm{T}\}$ would be the set of possible outcomes of a coin flip
- Each element $\omega \in \Omega$ is associated with a number $\mathrm{p}(\omega) \in [0,1]$ called a **probability**

$$\sum_{\omega \in \Omega} p(\omega) = 1$$

  - For example, a biased coin might have $p(H) = .6$ and $p(T) = .4$

Discrete Probability

- An **event** is a subset of the sample space
  - Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ be the 6 possible outcomes of a dice role
  - $A = \{1, 5, 6\} \subseteq \Omega$ would be the event that the dice roll comes up as a one, five, or six
- The probability of an event is just the sum of all of the outcomes that it contains
  - $p(A) = p(1) + p(5) + p(6)$

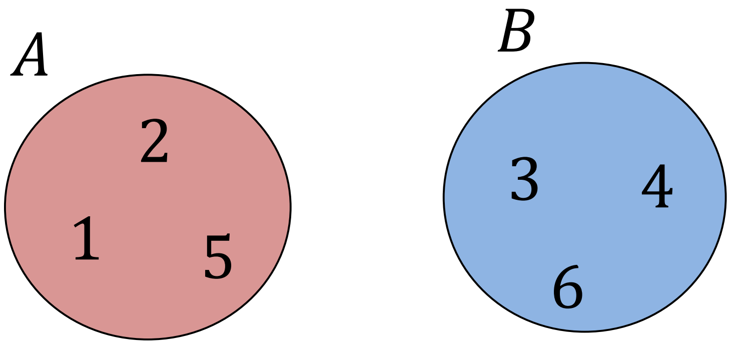
Independence

- Two events A and B are **independent** if

$$p(A \cap B) = p(A)P(B)$$

Let's suppose that we have a fair die: $p(1) = \ \dots = p(6) = 1/6$

If $A = \{1, 2, 5\}$ and $B = \{3, 4, 6\}$ are A and B indpendent?



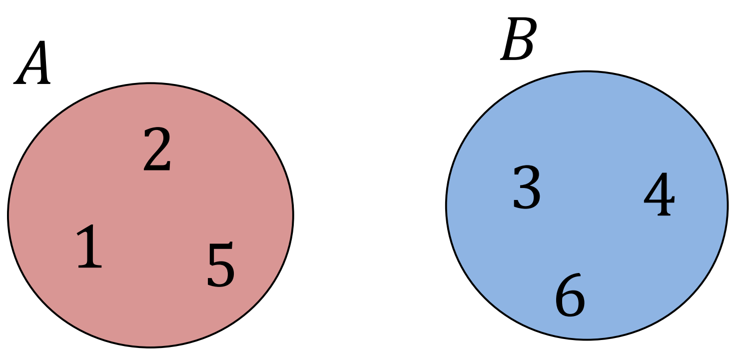
Independence

- Two events A and B are **independent** if

$$p(A \cap B) = p(A)P(B)$$

Let's suppose that we have a fair die: $p(1) = \ ... = p(6) = 1/6$

If $A = \{1, 2, 5\}$ and $B = \{3, 4, 6\}$ are A and B indpendent?



No!

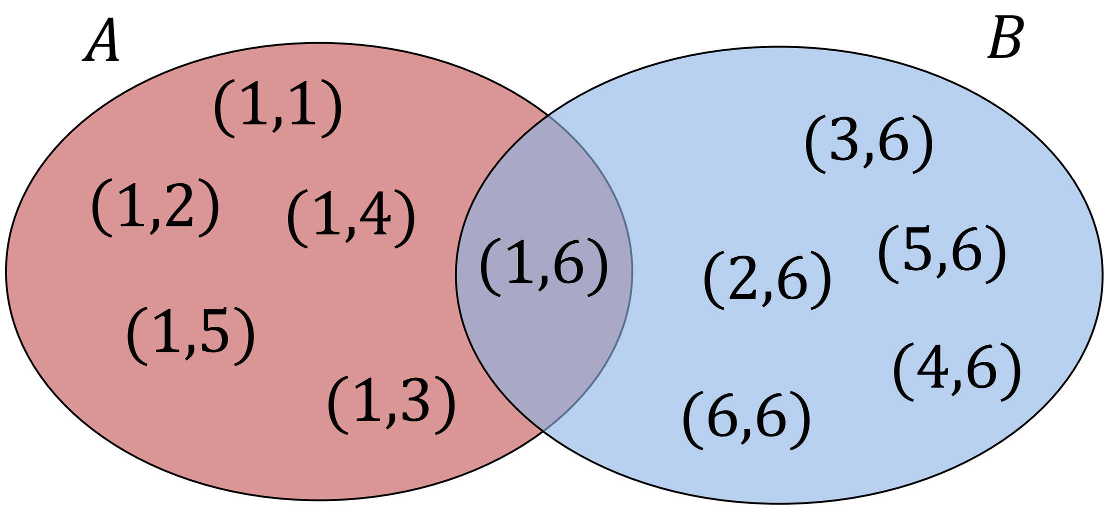
$$p(A \cap B) = 0 \neq \frac{1}{4}$$

Independence

- Now, suppose that $\Omega = \{(1,1), (1,2), \dots, (6,6)\}$ is the set of all possible rolls of two **unbiased** dice

- Let $A = \{(1,1), (1,2), (1,3), \dots, (1,6)\}$ be the event that the first die is a one and let $B = \{(1,6), (2,6), \dots, (6,6)\}$ be the event that the second die is a six

- Are A and B independent?

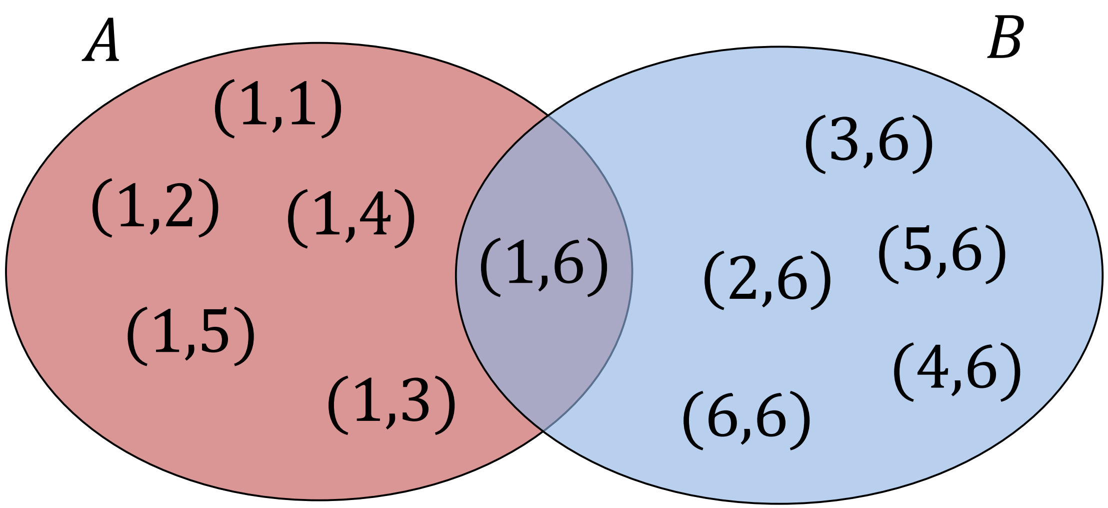


Independence

- Now, suppose that $\Omega = \{(1,1), (1,2), \ldots, (6,6)\}$ is the set of all possible rolls of two **unbiased** dice

- Let $A = \{(1,1), (1,2), (1,3), \ldots, (1,6)\}$ be the event that the first die is a one and let $B = \{(1,6), (2,6), \ldots, (6,6)\}$ be the event that the second die is a six

- Are A and B independent?



Yes!

$$p(A \cap B) = \frac{1}{36} = \frac{1}{6} * \frac{1}{6}$$

Conditional Probability

- The **conditional probability** of an event A given an event B with $p(B) > 0$ is defined to be

$$p(A|B) = \frac{p(A \cap B)}{P(B)}$$

- This is the probability of the event $A \cap B$ over the sample space $\Omega' = B$

- Some properties:
  - $\sum_{\omega \in B} p(\omega|B) = 1$
  - If A and B are independent, then $p(A|B) = p(A)$

UTD

Simple Example

Cheated	Grade	Probability
Yes	A	.15
Yes	F	.05
No	A	.5
No	F	.3

Chain Rule

$$p(A \cap B) = p(A)p(B|A)$$

$$p(A \cap B \cap C) = p(A \cap B)p(C|A \cap B)$$
$$= p(A)p(B|A)p(C|A \cap B)$$

.

.

.

$$p\left(\bigcap_{i=1}^{n} A_i\right) = p(A_1)p(A_2|A_1) \ldots p(A_n|A_1 \cap \cdots \cap A_{n-1})$$

Conditional Independence

- Two events A and B are independent if learning something about B tells you nothing about A (and vice versa)

- Two events A and B are **conditionally independent** given C if

$$p(A \cap B|C) = p(A|C)p(B|C)$$

- This is equivalent to

$$p(A|B \cap C) = p(A|C)$$

  – That is, given C, information about B tells you nothing about A (and vice versa)

Conditional Independence

- Let $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$ be the outcomes resulting from tossing two different fair coins

- Let A be the event that the first coin is heads
- Let B be the event that the second coin is heads
- Let C be the even that both coins are heads or both are tails

- A and B are independent, but A and B are not independent given C

Discrete Random Variables

- A discrete **random variable**, X, is a function from the state space Ω into a discrete space D
  - For each $x \in D$,

$$p(X = x) \equiv p(\{\omega \in \Omega : X(\omega) = x\})$$

    is the probability that X takes the **value** x

  - $p(X)$ defines a probability distribution
    - $\sum_{x \in D} p(X = x) = 1$
- Random variables partition the state space into disjoint events

Example: Pair of Dice

- Let Ω be the set of all possible outcomes of rolling a pair of dice
- Let p be the uniform probability distribution over all possible outcomes in Ω
- Let $X(\omega)$ be equal to the sum of the value showing on the pair of dice in the outcome ω
  - $p(X = 2) = ?$
  - $p(X = 8) = ?$

Example: Pair of Dice

- Let Ω be the set of all possible outcomes of rolling a pair of dice
- Let p be the uniform probability distribution over all possible outcomes in Ω
- Let $X(\omega)$ be equal to the sum of the value showing on the pair of dice in the outcome ω
  - $p(X = 2) = \frac{\mathbf{1}}{\mathbf{36}}$
  - $p(X = 8) =$?

Example: Pair of Dice

- Let Ω be the set of all possible outcomes of rolling a pair of dice
- Let p be the uniform probability distribution over all possible outcomes in Ω
- Let $X(\omega)$ be equal to the sum of the value showing on the pair of dice in the outcome ω
  - $p(X = 2) = \frac{\mathbf{1}}{\mathbf{36}}$
  - $p(X = 8) = \frac{\mathbf{5}}{\mathbf{36}}$

Discrete Random Variables

- We can have vectors of random variables as well

$$X(\omega) = [X_1(\omega), \dots, X_n(\omega)]$$

- The **joint distribution** is $p(X_1 = x_1, \dots, X_n = x_n)$ is

$$p(X_1 = x_1 \cap \cdots \cap X_n = x_n)$$

  typically written as

$$p(x_1, \dots, x_n)$$

- Because $X_i = x_i$ is an event, all of the same rules - independence, conditioning, chain rule, etc. - still apply

Discrete Random Variables

- Two random variables X_1 and X_2 are independent if

$$p(X_1 = x_1, X_2 = x_2) = p(X_1 = x_1)p(X_2 = x_2)$$

for all values of x_1 and x_2

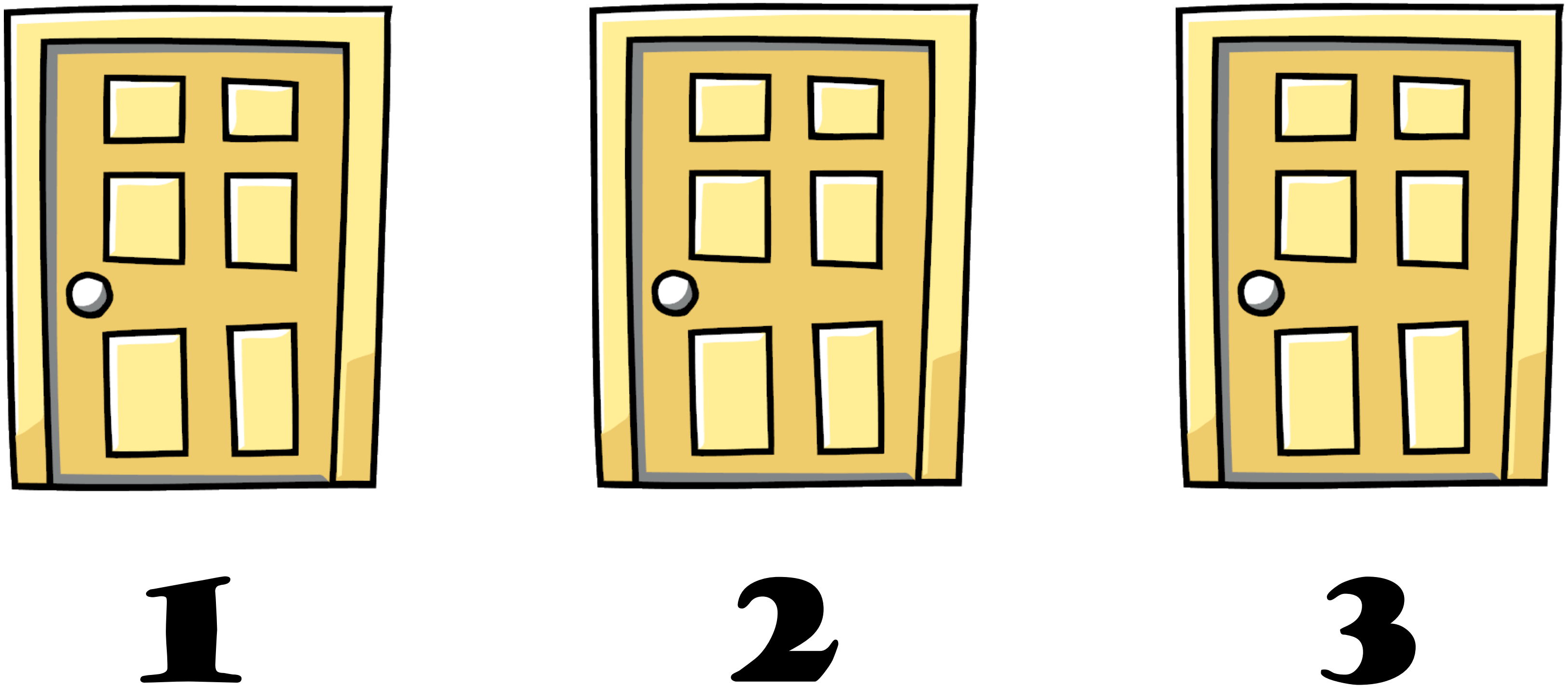
- Similar definition for conditional independence

- The conditional distribution of X_1 given $X_2 = x_2$ is

$$p(X_1 | X_2 = x_2) = \frac{p(X_1, X_2 = x_2)}{p(X_2 = x_2)}$$

this means that this relationship holds for all choices of x_1

The Monty Hall Problem



Expected Value

- The expected value of a real-valued random variable is the weighted sum of its outcomes

$$E[X] = \sum_{x \in \mathrm{D}} p(X = d) \cdot d$$

- Expected value is linear

$$E[X + Y] = E[X] + E[Y]$$

Expected Value: Lotteries

- Powerball Lottery currently has a grand prize of $1.4 billon
- Odds of winning the grand prize are $1/292{,}201{,}338$
- Tickets cost $2 each
- Expected value of the game

$$= \frac{-2 \cdot 292{,}201{,}337}{292{,}201{,}338} + \frac{1{,}400{,}000{,}000 \; - 2}{292{,}201{,}338} \approx \$3$$

Variance

- The variance of a random variable measures its squared deviation from its mean

$$var(X) = E[(X - E[X])^2]$$

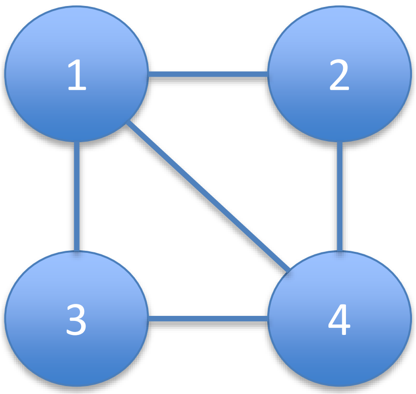
Example: Independent Sets

- Let Ω be the set of all vertex subsets in a graph $G = (V, E)$
- Let p be the uniform probability distribution over all independent sets in Ω
- Define for each $\mathrm{v} \in V$ and each subset of vertices ω

$$X_v(\omega) = 1, \qquad \text{if } v \in \omega \text{ and}$$
$$X_v(\omega) = 0, \qquad \text{otherwise}$$

- $p(X_v = 1)$ is the fraction of all independent sets in G containing v
- $p(x_1, \ldots, x_n) \neq 0$ if and only if the x's define an independent set

Example: Independent Sets



Consider the graph on the left, with the sample space and probabilities from the last slide

- $p(X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 1) = ?$
- $p(X_1 = 0, X_2 = 1, X_3 = 1, X_4 = 0) = ?$
- $p(X_1 = 1) = ?$

Example: Independent Sets

- How large of a table is needed to store the joint distribution $p(X_V)$ for a given graph $G = (V, E)$?

Example: Independent Sets

- How large of a table is needed to store the joint distribution $p(X_V)$ for a given graph $G = (V, E)$?

$$\mathbf{2^{|V|}\text{-}1}$$

Structured Distributions

- Consider a general joint distribution $p(X_1, \dots, X_n)$ over binary valued random variables

- If $X_1, \dots, X_n$ are all independent random variables, then

$$p(x_1, \dots, x_n) = p(x_1) \dots p(x_n)$$

- How much information is needed to store the joint distribution?

Structured Distributions

- Consider a general joint distribution $p(X_1, \dots, X_n)$ over binary valued random variables

- If $X_1, \dots, X_n$ are all independent random variables, then

$$p(x_1, \dots, x_n) = p(x_1) \dots p(x_n)$$

- How much information is needed to store the joint distribution?

n numbers

- This model is boring: knowing the value of any one variable tells you nothing about the others

Structured Distributions

- Consider a general joint distribution $p(X_1, \dots, X_n)$ over binary valued random variables

- If $X_1, \dots, X_n$ are all independent given a different random variable Y, then

$$p(x_1, \dots, x_n | y) = p(x_1 | y) \dots p(x_n | y)$$

and

$$p(y, x_1, \dots, x_n) = p(y) p(x_1 | y) \dots p(x_n | y)$$

- These models turn out to be surprisingly powerful, despite looking nearly identical to the previous case!

Marginal Distributions

- Given a joint distribution $p(X_1, \dots, X_n)$, the marginal distribution over the i^{th} random variable is given by

$$p_i(X_i = x_i) = \sum_{x_1} \sum_{x_2} \dots \sum_{x_{i-1}} \sum_{x_i+1} \dots \sum_{x_n} p(X_1 = x_1, \dots, X_n = x_n)$$

- In general, marginal distributions are obtained by fixing some subset of the variables and summing out over the others
  - This can be an expensive operation!

Inference/Prediction

- Given fixed values of some subset, E, of the random variables, compute the conditional probability over the remaining variables, S

$$p(X_S|X_E = x_E) = \frac{p(X_S, X_E = x_E)}{p(X_E = x_E)}$$

- This involves computing the marginal distribution $p(X_E = x_E)$, so we refer to this as marginal inference

Inference/Prediction

- Given fixed values of some subset, E, of the random variables, compute the most likely assignment of the remaining variables, S

$$\operatorname*{argmax}_{x_S} p(X_S = x_s | X_E = x_E)$$

- This is called maximum a posteriori (MAP) inference
- We don't need to do marginal inference to compute the MAP assignment, why not?