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

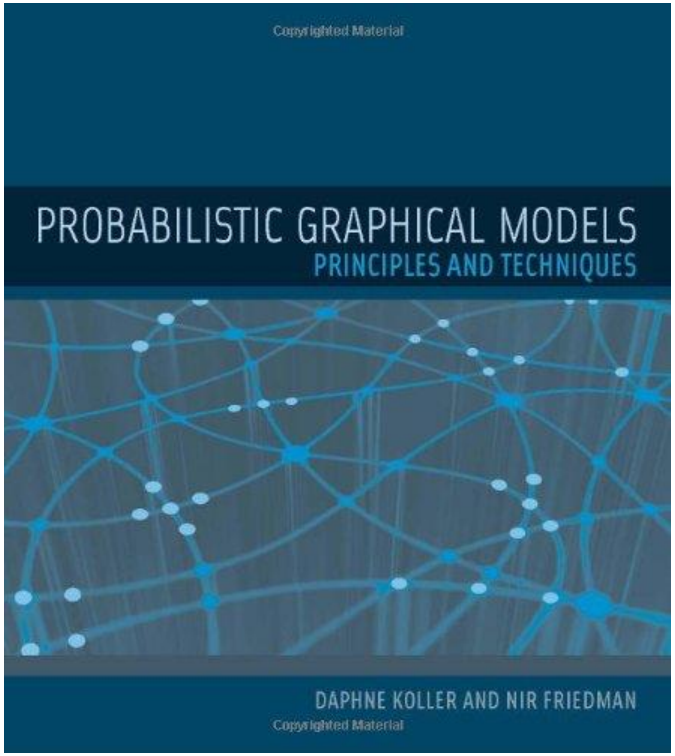
Prerequisites

- 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)

Suggested Textbook

Readings will be posted online before each lecture

Check the course website for additional resources and papers



Textbook

- In addition, lecture notes, in book format, will (hopefully) be made available for each topic
- The idea is to build a set of notes that aligns well with the presentation of course material
- Comments, suggestions, corrections are welcome/encouraged

Grading

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

-subject to change-

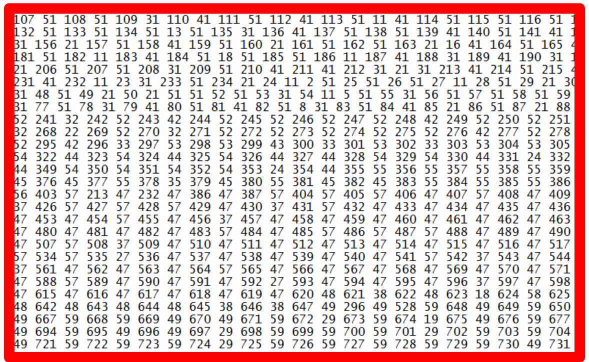
Course Info.

- Instructor: Nicholas Ruozzi
  - Office: ECSS 3.409
  - Office hours: Tues. 10am – 11am and by appointment
- TA: TBD
  - Office hours and location TBD
- Course website: http://www.utdallas.edu/~nrr150130/cs6347/2017sp/

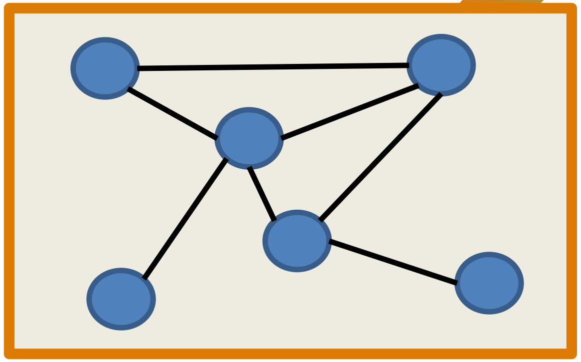
Main Ideas

- 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

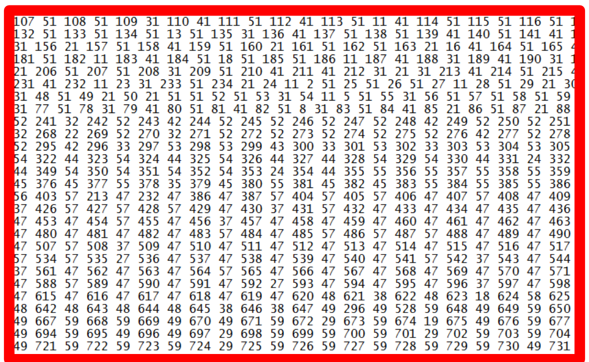


"Learn" a model that represents the observed data

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

Use the model to do inference / make predictions

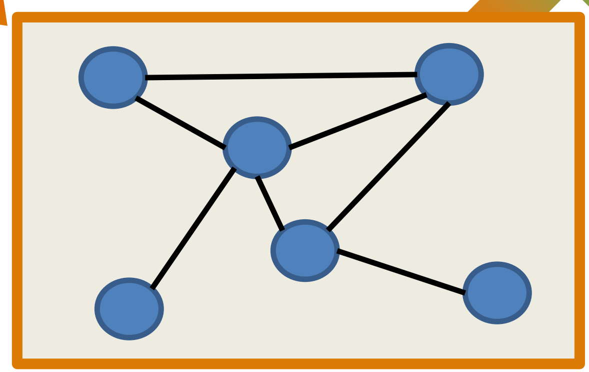
Inference and Learning



Data sets can be large

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

Inference needs to be fast



Data must be compactly modeled

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
  - $p(A) = p(1) + p(5) + p(6)$

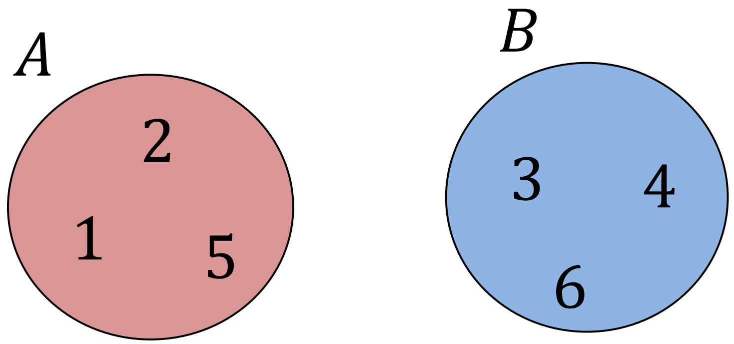
Independence

- 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: $p(1) = \ \ldots = p(6) = 1/6$

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



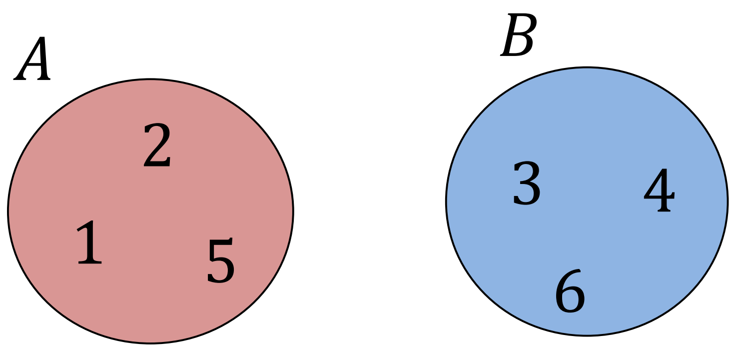
Independence

- 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?

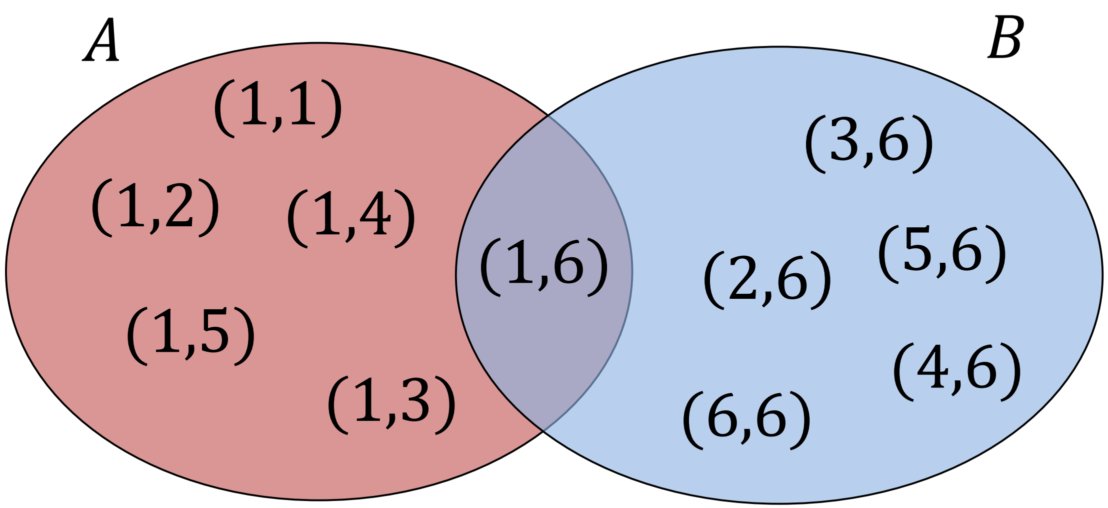


No!

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

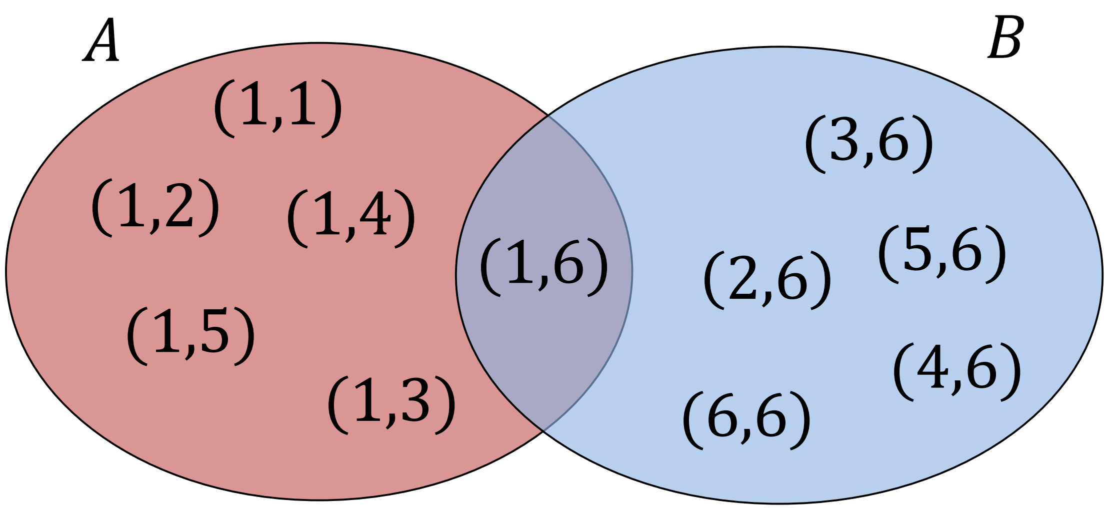
Independence

- Now, suppose that $\Omega = \{(1,1), (1,2), \ldots, (6,6)\}$ 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?



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

Simple Example

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

Simple Example

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

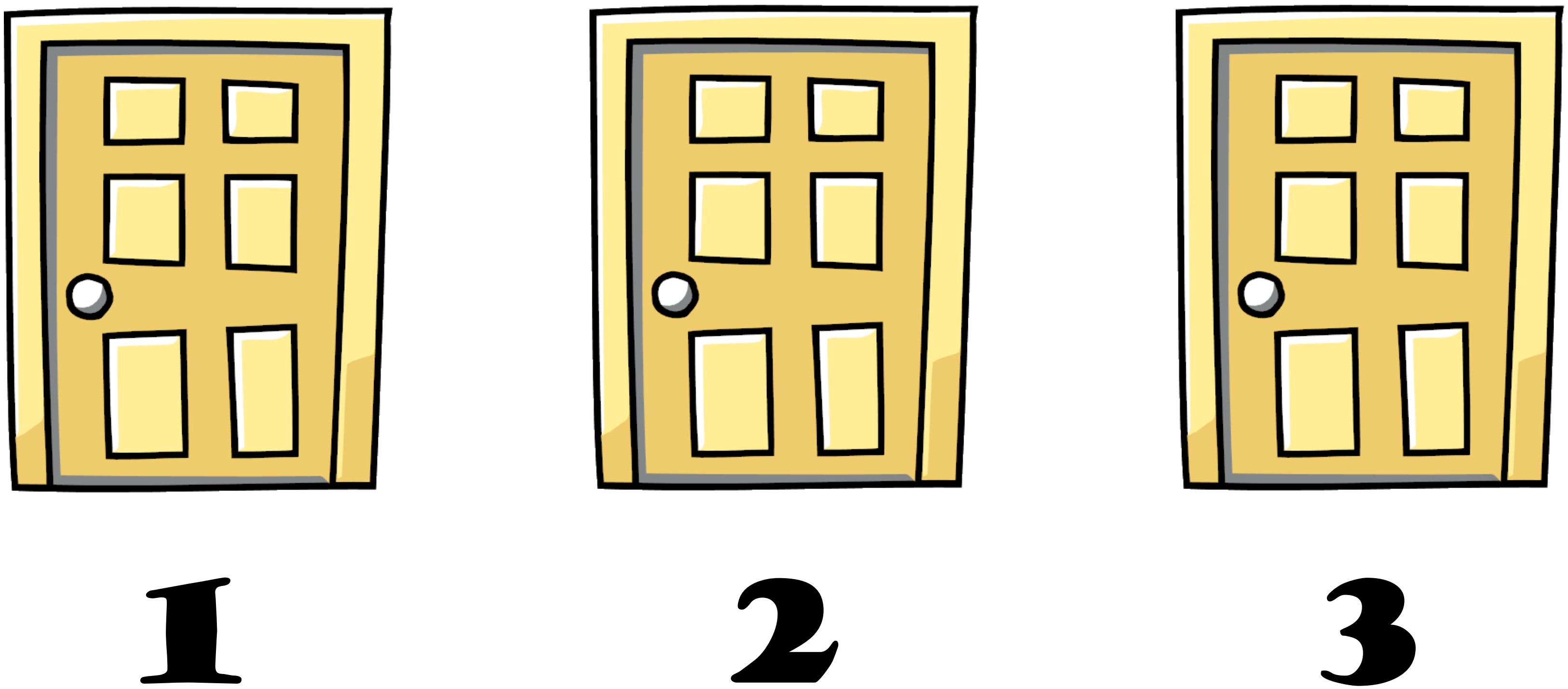
$$p(Cheated = Yes \mid Grade = F) = ?$$

Simple Example

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

$$p(Cheated = Yes\,|Grade = F) = \frac{.05}{.35} \approx .14$$

The Monty Hall Problem



Chain Rule

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

$$\begin{aligned} p(A \cap B \cap C) &= p(A \cap B)p(C|A \cap B) \\ &= p(A)p(B|A)p(C|A \cap B) \end{aligned}$$

.

.

.

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

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
  - $p(X = 2) = \frac{\mathbf{1}}{\mathbf{36}}$
  - $p(X = 8) =$ **?**

UTD

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

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 \$106 million
- Odds of winning the grand prize are $1/$292,201,338
- Tickets cost \$2 each
- Expected value of the game

$$= \frac{-2 \cdot 292{,}201{,}337}{292{,}201{,}338} + \frac{106{,}000{,}000 \; - 2}{292{,}201{,}338} \approx \$ - 1.6$$

Variance

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

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

  – Estimates the square of the expected amount by which a random variable deviates from its expected value