Why probability and statistics in CS?

Course overview: Probability and statistics - key concepts & applications

Course prerequisites:
calculus, programming

Course work and evaluation:
ten homeworks, midterm exam, final exam

Why you shouldn’t take this class

Waiting list, registration, etc.
Why Probability?

Most advanced computer applications today involve randomization.

- Secured web connections are **probabilistically secured**
- Web search engines apply **statistical inference** to determine which pages are most relevant to ambiguous queries
- Computer games would be boring without **randomization**
- Spam filters, recommendation systems, and web advertising are constructed via **(statistical) machine learning**
- Efficient data structures are often **randomized** (e.g., hashing)
- Wireless communications: split bandwidth via **random codes**
- Computational finance: Must model market **uncertainty**
- Computational biology: DNA sequencing, clinical trials, …
- Robotics: navigating new environments, human interaction, …
Why Rigorous Mathematical Approach?

Probabilistic reasoning is often counterintuitive

Daniel Kahneman was awarded the 2002 Nobel Prize in Economic Sciences for his empirical findings challenge the assumption of human rationality prevailing in decision-making under uncertainty.
Example: The Monty Hall Problem

- A prize is equally likely to be behind one of three doors.
- You choose a door and before that door is opened another door without the price is opened.
- You are offered to switch your choice to the other closed door.
- Should you do it?
Why probability and statistics in CS?

Course overview: Probability and statistics - key concepts & applications

Course prerequisites:
calculus, programming

Course evaluation:
ten homeworks, midterm exam, final exam

Why you shouldn’t take this class?

Waiting list, registration, etc.
Sample spaces: When a random event happens, what is the set of all possible outcomes? May be discrete or continuous.

Conditioning: Suppose I observe some data. How does my probability model change?

Independence: Is there any relationship between pairs of variables in my model? Would data provide knowledge?

Weather example:
- Raining or not
- Sunny or not
- Hot or not
Suppose I toss a coin 10 times. The number of tosses that come up heads, rather than tails, is an example of a *discrete random variable*. A *probability mass function* gives the (non-negative) probability of each possible outcome. These probabilities sum to one.
- **Joint Distribution:** Probability of each possible outcome.
- **Marginal Distribution:** If some variables are not observed and not relevant, how do I remove them from the model?
- **Conditional Distribution:** What if I observe some data?
Continuous Random Variables

CDF: cumulative distribution function

\[ F(q) \triangleq p(X \leq q) \]

\[ p(a < X \leq b) = F(b) - F(a) \]

PDF: probability density function

\[ f(x) = \frac{d}{dx} F(x) \]

\[ P(a < X \leq b) = \int_{a}^{b} f(x) \, dx \]

\[ P(x \leq X \leq x + dx) \approx p(x) \, dx \]

Model processes or data which are encoded as real numbers: temperature, commodity price, DNA expression level, light on camera sensor, …
Gaussian (Normal) Distributions

Summaries: Mean, median, mode, variance, standard deviation, …
Central Limit Theorem

- In a large population, how likely is a person to be much taller than average?
- How likely is a request on my web server to be much larger than average?

Gaussian Body Shape Modeling

http://bodyvisualizer.com/
Monte Carlo Methods

Hurricane Sandy made landfall in New Jersey on October 29, 2012.

Weather Wisdom, Boston.com
Markov Chains

Markov Property: Conditioned on the present, past & future are independent

- Building block for modeling random processes that evolve and change over time.
- What is the long-term behavior of some process?
  What is the probability of reaching a good state? A bad state?
- Allows agents to reason about future consequences of actions.

$$p(x) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2)p(x_4 \mid x_3) \cdots = p(x_1) \prod_{t=2}^{T} p(x_t \mid x_{t-1})$$
Markov Chains for Robot Navigation

Simultaneous Localization & Mapping (FastSLAM, Montemerlo 2003)

$$\mathbf{p}(x_t, m | z_{1:t}, u_{1:t})$$

- $x_t$ = State of the robot at time $t$
- $m$ = Map of the environment
- $z_{1:t}$ = Sensor inputs from time 1 to $t$
- $u_{1:t}$ = Control inputs from time 1 to $t$

Raw odometry (controls)
True trajectory (GPS)
Inferred trajectory & landmarks
Markov Chains for Web Search: PageRank

[Image of network and smiley faces]
In *probability theory* we compute the probability that 20 independent flips of a fair (unbiased) coin give the sequence

\[ \text{HTTHTHTHHTTHTHTHHTTT} \]

In *statistics* we ask: Given that we observed the sequence

\[ \text{HTTHTHTHHTTHTHTHHTTT} \]

what is the likelihood that the coin is fair (unbiased)?
The Frequentist Model: The probability of an outcome in a trial is the frequency of that outcome in a long sequence of identical and independent such trials (limiting frequency).
The Frequentist Model: The probability of an outcome in a trial is the frequency of that outcome in a long sequence of identical and independent such trials (limiting frequency).

The Bayesian Belief Model: Based on all the information we have seen so far, the probability is our best estimate for the chance of a particular outcome.

I offer a bet on the super bowl outcome at q to 1 odds: you lose $1 if you’re wrong, and earn $q if you’re right. A rational person would take this bet only if they believe the probability of winning is at least 1/(1+q).

Probabilities and odds express the same tradeoff.
Spam Filtering

- Binary classification: Is this e-mail useful (ham) or spam?
- Noisy training data: Messages previously marked as spam
- Estimate: Probability that certain words are used in spam and non-spam emails
- Estimate: Probability that certain servers send spam

**Spam Filter Express:** [http://www.spam-filter-express.com/]
1.2. Supervised learning

Face detection and recognition

A harder problem is to find objects within an image; this is called object detection or object localization. An important special case of this is face detection. One approach to this problem is to divide the image into many small overlapping patches at different locations, scales, and orientations, and to classify each such patch based on whether it contains face-like texture or not. This is called a sliding window detector. The system then returns those locations where the probability of face is sufficiently high. See Figure 1.6 for an example. Such face detection systems are built-in to most modern digital cameras; the locations of the detected faces are used to determine the center of the auto-focus. Another application is automatically blurring out faces in Google's StreetView system.

Having found the faces, one can then proceed to perform face recognition, which means estimating the identity of the person (see Figure 1.10(a)). In this case, the number of class labels might be very large. Also, the features one should use are likely to be different than in the face detection problem: for recognition, subtle differences between faces such as hairstyle may be important for determining identity, but for detection, it is important to be invariant to such details, and to just focus on the differences between faces and non-faces.

K. Murphy & Family

Based on classifiers trained from tens of thousands of example faces (Viola & Jones, 2004)
Digit & Hand Gesture Recognition

Database (60,000 images)

Database (107,328 images)

Athitsos et al., CVPR 2004 & PAMI 2008
IBM’s Watson wins Jeopardy!
Summary of Course Topics

I. Probability Models
II. Discrete Random Variables
III. Continuous Random Variables
IV. Normal Distributions
V. Limit Theorems
VI. Markov Chains
VII. Monte Carlo Methods
VIII. Bayesian Statistical Inference
IX. Frequentist Statistical Inference
CS145: Lecture 0 Outline

- Probability and statistics: key concepts & applications
- **Course Details: People**
- Course prerequisites:
  - *calculus, programming*
- Course work and evaluation:
  - *ten homeworks, midterm exam, final exam*
- Why you shouldn’t take this class
- Waiting list, registration, etc.
CS145 Course Staff

**Instructor:**  
Prof. Eli Upfal

**Co-Instructor:**  
Lorenzo de Stefani

**Graduate TA:**  
Cyrus Cousins

**Head Undergraduate TA:**  
Zachary Shinkar <zack@shinkar.com>

**Undergraduate TAs:**  
Misha Sohan, Pranav Nagalamadaka, Sachin Sastri, Josh Beck, Sandra Ha, Ruizhao Zhu
Course Prerequisites

Not formally enforced, but we will assume comfort with:

Calculus

- Two semesters of college-level calculus
- AP Calculus BC exam, or Brown MATH 0100/0170
- Topics: limits, basic derivatives & chain rule, basic integrals & fundamental theorem of calculus, sequences & series, …

Programming
Example: Buffon’s Needle

A surface is ruled with parallel lines, which are at distance \( d \) from each other. Suppose that we throw a needle of length \( l < d \) on the surface at random. What is the probability that the needle will intersect one of the lines?

\[
f_{X, \Theta}(x, \theta) = \begin{cases} 
\frac{4}{\pi d} & \text{if } x \in [0, d/2] \text{ and } \theta \in [0, \pi/2], \\
0 & \text{otherwise.}
\end{cases}
\]

\[
\mathbf{P}(X \leq (l/2) \sin \Theta) = \int_{0}^{(l/2) \sin \Theta} \int_{0}^{\pi/2} f_{X, \Theta}(x, \theta) \, dx \, d\theta
\]

\[
= \frac{4}{\pi d} \int_{0}^{\pi/2} \int_{0}^{(l/2) \sin \theta} dx \, d\theta
\]

\[
= \frac{4}{\pi d} \int_{0}^{\pi/2} \frac{l}{2} \sin \theta \, d\theta
\]

\[
= \frac{2l}{\pi d} (-\cos \theta) \Big|_{0}^{\pi/2}
\]

\[
= \frac{2l}{\pi d}.
\]
Course Prerequisites

Not formally enforced, but we will assume comfort with:

Calculus

- Two semesters of college-level calculus
- AP Calculus BC exam, or Brown MATH 0100/0170
- Topics: limits, basic derivatives & chain rule, basic integrals & fundamental theorem of calculus, sequences & series, ...

Programming

- Any single-semester programming course: CS4, CS15, CS17, CS19, etc.
- Or, other experience that gives comfort with writing simple functions
- All support code will be provided in Matlab (easy to learn if you’ve seen Java, Python, etc.), but you can use any other language.
- Topics: functions, loops & flow control, arrays, ...
Supplemental readings (online) for a few advanced topics
CS145: Lecture 0 Outline

- Probability and statistics: key concepts & applications
- Course Details: People
- Course prerequisites: calculus, programming
- Course work and evaluation: ten homeworks, midterm exam, final exam
- Why you shouldn’t take this class
- Waiting list, registration, etc.
Course Work and Evaluation

Homework 50%, Midterm 20%, Final 30% of course grade

- Ten equally weighted homework assignments
- Probabilistic derivations, calculations, and reasoning (i.e., math)
- Usually one implementation problem: Monte Carlo, statistical data analysis, etc.
- Submitted electronically by Tuesday class, out for one week
- Read the syllabus and grading policies document on the website
- In class midterm and final

- HW0 is on the website now. Will not be graded, solutions posted on Tuesday.
- HW1 out on Tuesday for a week.

- Waiting list, registration:
  - If you get > 70% grade on HW 1 you’ll have a place in the class.
  - Pls. drop the course asap if you registered but decided not to take it.
Recitations

- An *optional* review and problem-solving session led by graduate TA.
- Materials will be posted online for those with conflicts.
- Wednesdays 6pm.
- Begins next week with Gradescope and Matlab tutorial
Why You Should NOT Take This Course

• You can satisfy the requirement with APMA 1650/1655
• APMA 1650/5 is less “mathematical”
• APMA 1650/5 covers no CS/algorithmic applications
• APMA 1650/5 has no computation/implementation assignments
• APMA 1650/5 is EASIER!

• BUT – if you plan to study machine learning, data science, or theory, consider CSCI 1450
Questions?