DSE 210: Probability and statistics

Overview

The kinds of questions we'll study

- Design a spam filter.
- What fraction of San Diegans like Donald Trump?
- Categorize New York Times articles by their underlying topics.
- Two new malaria vaccines are under consideration. How can we determine which is better?
- We've obtained user ratings of many movies. Visualize them.
- A dating service asks each user to answer 200 multiple choice questions. Summarize each user's responses by a few numbers.

Intermediate-level questions

- Regression. How do you fit a line to a set of points?
- Clustering. Given a bunch of data points, partition them into groups that are distinct from each other.
- Laws of large numbers. A drunk starts off from a bar and at each time step, takes either a step to the right or a step to the left. Where will he be, approximately, after n time steps?
- Hypothesis testing. You are given two alternatives and wish to test which is better. Design an experiment to do this.
- Dimensionality reduction. Find the primary axes of variation in a data set.

Low-level questions

- If you toss a coin 10 times, what is the chance of getting heads every time?
- Throw 20 balls into 20 bins at random. What is the probability that at least one of the bins remains empty?
- If each cereal box contains one of k action figures, how many boxes do you need to buy, on average, before getting all the figures?
- What fraction of a bell curve lies at least one standard deviation away from the mean?
- Find a concise description of a data matrix.

Course outline

1. Probability basics
2. Fitting distributions to data
3. Regression, classification, embedding, and visualization
4. Sampling and hypothesis testing
5. Advanced probabilistic modeling