Data 100, Discussion 2

Suraj Rampure

Wednesday, September 4th, 2019

Hi, I'm **Suraj!**

I'm a senior EECS major from Windsor, Ontario, Canada . This is my third time TAing for Data 100, but I've TA'd for Data 8, CS 61A, CS 70, and CS 375 before as well.

- **Email**: suraj.rampure@berkeley.edu – feel free to email me about **anything**!
- **Lab**: Mondays, 3-4PM, Evans B6
- **Discussion**: Wednesdays, 3-4PM, Etcheverry 3107
- **Office Hours**: Tuesdays, 1-3PM, Evans 426 (room change soon!)

All discussion/lab slides, notes, and other resources will be posted at

surajrampure.com/teaching/ds100.html

tinyurl.com/firstdiscdata

Before leaving, please fill out this form. It contains some introductory questions, just for me to learn who is in the class.

Today, we'll review...

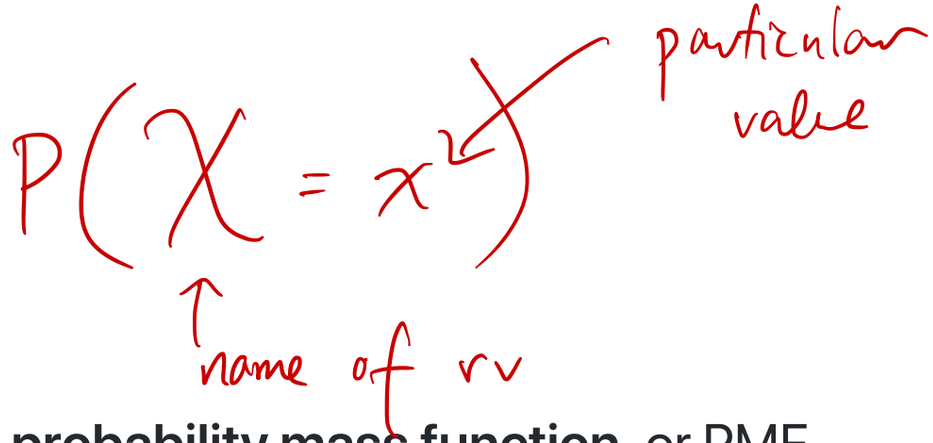
- Random Variables
- Modelling and Loss
- Simple Random Sampling
- Sampling Terminology

Random Variables

A **random variable** is a variable whose values are determined by probabilities. In other words, a random variable is a function that takes outcomes of a random process to real numbers.

- Usually, we use X or Y to denote a random variable.
- In this course, we will only consider discrete random variables (i.e., random variables whose set of outcomes are countable)

Describing Random Variables



Each discrete random variable has associated with it a **probability mass function**, or PMF.

- $P(X = x)$ represents the probability of random variable X taking on the value of x. (x is normally numerical, but it doesn't necessarily have to be.)
- $\mathbb{X}$, then, represents the set of possible outcomes (e.g. dice rolls, cards, etc.)

Properties that $P(X = x)$ must satisfy:

$$0 \leq P(X = x) \leq 1, \forall x \in \mathbb{X}$$

$$\sum_{x \in \mathbb{X}} P(X = x) = 1$$

We can also use a *distribution table*, as seen in lecture, to describe a random variable.

Example: Dice rolls

x	$P(X=x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

→ distribution table

Expectation

Often, we want to determine the expected, or average, outcome of a probabilistic event. For that, we define the **expectation** of a random variable:

$$\mathbb{E}[X] = \sum_{x \in \mathbb{X}} x \cdot P(X = x) = \sum \text{outcome} \cdot P(\text{outcome})$$

This is a weighted average of all possible outcomes. **Note, the expectation is just a constant!** Sometimes denoted with μ.

Properties that the expectation satisfies:

$$\mathbb{E}[c] = c, \forall c \in \mathbb{R}$$

$$\mathbb{E}[aX] = a\mathbb{E}[X]$$

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

} linearity of expectation

Example: Dice rolls

$$E[X] = \sum_x x \cdot P(X = x)$$

$$= 1 \cdot 1/6 + 2 \cdot 1/6 + \cdots + 6 \cdot 1/6$$

$$= \boxed{7/2}$$

Sampling Context

So far, we've studied the definition of expectation with regards to random variables.

- When we collect a sample, we can treat our data points $\{x_1, x_2, ..., x_n\}$ as random variables $\{X_1, X_2, ..., X_n\}$.
- As such, there are parallels between the definitions of means / variances for random variables and sample points.

Sample mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Similarly, the sample variance is defined as

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

From this, can you determine the random variable definition of variance?

Modelling and Loss

Suppose we have a collection of data points $\{x_1, x_2, \ldots, x_n\}$, and we want to come up with a **summary statistic** c for this data, that is the "best", in some sense.

- Prediction error: $x_i - c$
- To determine the "best" c, we need a function in terms of our true value x_i and prediction c, that increases as our error increases

L2 loss for a single point: $(x_i - c)^2$

(actual - pred)^2

Average L2 loss for entire dataset:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$$

In the worksheet, we'll figure out how to minimize this quantity (and its parallel in terms of random variables).

Simple random samples (SRS)

Sampling uniformly at random from the population

- For example, if we have five students, A, B, C, D, and E, and want to select two of them, our sample will look like AB, AC, AD, AE, BC, BD, BE, CD, CE, or DE.
- There are $\binom{5}{2}$ total possible samples, and each is equally likely (with probability $\frac{1}{\binom{5}{2}} = \frac{1}{10}$).
- Each student appears in exactly 4 of the samples, so the probability that any one specific student appears in our sample is $\frac{5-1}{\binom{5}{2}} = \frac{4}{10} = \frac{2}{5}$.
    - (Extra): In general, if we have n students and want to choose k of them, the probability that one specific student is chosen in our sample is $\frac{\binom{n-1}{k-1}}{\binom{n}{k}}$.

Question: Why use SRS, and probability samples in general? When do we need to?

→ to ensure our sample resembles our population (try to remove confounding factors)

Sampling Terminology

Let's discuss and define the following terms.

Population of Interest

The group we want to answer some question about

Sampling Frame

The group that we have access to to sample from

Sample

The group that we actually study (subset of sampling frame)

Confounding Factors

factors that we didn't account for that may influence our results