Data 100, Discussion 3

Suraj Rampure

Wednesday, September 11th, 2019

All discussion will be posted at

surajrampure.com/teaching/ds100.html

Today, we'll review...

- Modelling and loss
- Sampling techniques

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

L_2 loss for a single point: $(x_i - c)^2$

L_1 loss for a single point: $|x_i - c|$

In general, we want to **minimize the average loss (i.e. empirical risk) over our entire dataset:**

$$\frac{1}{n}\sum_{i=1}^{n} L(x_i, c)$$

Last week, we saw that

$$c = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{minimizes} \quad \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$$

i.e., that the sample mean minimizes average squared loss.

In today's worksheet, we'll determine the value of c that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - c|$$

i.e., the value of c that minimizes average absolute loss.

Sampling Techniques

Simple Random Sampling – 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 $\frac{\binom{n-1}{k-1}}{\binom{n}{k}} = \frac{k}{n}$.

Stratified sampling

In stratified sampling, we split our population into **disjoint groups**, called **strata**. We then use SRS to sample *within every strata*.

- For example, if we wanted to survey some number of CS and EECS majors at Berkeley, we could split our population into four strata – 1st years, 2nd years, 3rd years and 4th years
- We could then use SRS within each strata to select students
- The end result – our sample contains representation from all four years