Chapter-2

Laika Jean Paculba

March 11, 2023

Notation

Notation for data

- We write $X_1, X_2, ..., X_n$ to describe n data points.
- As an example, consider the data set {1, 2, 5}

then $X_1 = 1, X_2 = 2, X_3 = 5$ and n = 3.

Of course, there's nothing in particular about the variable X. We often use a different letter, such as Y_1, . . . , Y_n to describe a data set. We will typically use Greek letters for things we don't know. Such as, μ being a population mean that we'd like to estimate.

The empirical mean

- The empirical mean is a measure of center of our data, it estimates a population mean of interest. Define the empirical mean as:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- if we subtract the mean from data points, we get data that has mean 0 if we define:

$$\tilde{x} = X_i - \overline{X}.$$

- then the mean of the $\tilde{x}_i$ is 0. This process is called centering the random variables. Since the empirical mean is the least squares solution for minimizing

$$\sum_{i=1}^{n}(X_i - \mu)^2$$

The empirical standard deviation and variance

- The variance and standard deviation are measures of how spread out our data is. Under sampling assumptions, they estimate variability in the population.

  The empirical variance is define as:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_1 = \overline{X})^2 = \frac{1}{n-1}(\sum_{i=1}^{n}(X_1^2 = n\overline{X}^2)$$

- The empirical standard deviation is defined as

$$S = \sqrt{S^2}$$

- The data defined by X_i/s have empirical standard deviation 1. This is called **scaling** the data

Normalization

- The data defined by

$$Z_i = \frac{X_i - \overline{X}}{s}$$

has empirical mean zero and empirical standard deviation 1.

- The process of centering then scaling the data is called normalizing the data.
- Normalized data are centered at 0 and have units equal to standard deviations of the original data.
- Normalized data are centered at 0 and have units equal to standard deviations of the original data.

Normalization is very useful for creating data that comparable across experiments by getting rid of any shifting or scaling effects.

The empirical covariance

This class is largely considering how variables covary. This is estimated by the empirical covariance.

- Consider now when we have pairs of data, (X_i, Y_i). Their empirical covariance is defined as:

$$Cov = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n-1}(\sum_{i=1}^{n} X_i Y_i - n\overline{XY})$$

- This measure is of limited utility, since its units are the product of the units of the two variables. A more useful definition normalizes the two variables first. The correlation is defined as:

$$Cor(X, Y) = \frac{Cov(X, Y)}{S_x S_y}$$

where S_x and S_y are the estimates of standard deviations for the X observations and Y observations, respectively.

- The correlation is simply the covariance of the separately normalized X and Y data. Because the data have been normalized, the correlation is a unit free quantity and thus has more of a hope of being interpretable across settings.

Some facts about correlation

- Cor(X, Y) = Cor(Y, X)
- -1 $\leq Cor(X, Y) \leq 1$
- Cor(X, Y) = 1 and Cor(X, Y) = -1 only when the X and Y observations fall perfectly on the positive or negative slope line, respectively.
- Cor(X, Y) measures the strength of the linear relationship between the X and Y data, with stringer relationships as Cor(X, Y) heads towards -1 or 1.
- Cor(X, Y) = 0 implies no linear relationships.