Chapter-2

Katlyn H. Degamo

2023-03-06

Notation

Some basic definitions

- In this chapter, we'll cover some basic definitions and notation used throughout the book.
- We will try to minimize the amount of mathematics required so that we can focus on the concepts.

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 differential 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. Under sampling assumptions, it estimates a population mean of interest.
- Define the empirical mean as
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
Notice if we subtract the mean from data points, we get data that has mean 0. That is, if we define
$$\tilde{X} = X_i - \bar{X}$$
then the mean of the $\bar{X}_i$ is 0. This process is called centering the random variables.
- Recall from theprevious lecture that 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. We define the empirical variance as:
$$S^n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2)$$

- The empirical standard deviation is defined as $S = \sqrt{S^2}$ Notice that the standard deviation has the same units as the data. The data defined by X_i/s have empirical standard deviation 1. This is called scaling the data.

Normalization

- We can combine centering and scaling of data as follows to get normalized data.
- In particular, the data defined by:
$$Z_i = \frac{X_i - \bar{X}}{s}$$
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.
- Example, a value of 2 from normalized data means that data point was two standard deviations larger than the mean.
- 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 co-vary.
- 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:
$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1}(\sum_{i=1}^{n} X_iY_i - n\bar{X}\bar{Y})$$
- 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:
$$Cor(X,Y) = \frac{Cov(X,Y)}{S_xS_y}$$
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

- First, the order of the arguments is irrelevant $Cor(X,Y) = Cor(Y,X)$
- Secondly, it has to be between -1 and 1, $-1 \leq Cor(X,Y) \leq 1$.
- Thirdly, the correlation is exactly -1 or 1 only when the observations fall perfectly on a negatively or positively sloped, line, respectively.
- Fourthly, $Cor(X,Y)$ measures the strength of the linear relationship between the two variables, with stronger relationships as $Cor(X,Y)$ heads towards -1 or 1.
- Finally, $Cor(X,Y) = 0$ implies no linear relationship.
- Click me
- Book