Basic Statistics for Machine Learning

We are going to cover the basics of statistics that are required for machine learning procedures. Why are statistics required for machine learning is primarily as we need, measurable techniques in the planning of training and testing for our machine learning algorithms, meaning that we need to be able to mathematically validate if we are able to get an algorithm up and running and ready for its implementation in the real world.

We are going to cover the following components (Not necessarily in the same order) in our endeavor to understand statistics specifically for their implementation in machine learning.

- Data scaling
- Data sampling
- Missing value imputation
- Outlier detection
- Central Tendency
- Data Distributions (primarily Normal Distributions)

Now, how important are Statistics?

Here are some of the excerpts from a news article related to the recent Cornavirus pandemic:

On **December 30, 2019**, BlueDot, a Toronto-based startup that uses a platform built around artificial intelligence, machine learning, and big data to track and predict the outbreak and spread of infectious diseases, alerted its private sector and government clients about a cluster of “unusual pneumonia” cases happening around a market in Wuhan, China.

That was the first recognition of the novel coronavirus that has come to be known asCOVID-19. It would be another nine days before the World Health Organization released its statement alerting people to the emergence of a novel coronavirus.

So how were they able to do this

The simpler answer will be using statistics, and specifically using one of the above-mentioned techniques that we are going to cover in detail, **Outlier detection.**

Let’s start with the formal definitions

Instead of the usual hard-hitting mathematical definitions, we are going to cover more basic understandings instead.

Mean

The "mean" is the "average" you're used to, where you add up all the numbers and then divide by the number of numbers. For example

In the list of values: **13, 18, 13, 14, 13, 16, 14, 21, 13**

The mean is the usual average, so I'll add and then divide:

(13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13) ÷ 9 = 15

Median

The "median" is the "middle" value in the list of numbers. To find the median, your numbers have to be listed in numerical order from smallest to largest, so you may have to rewrite your list before you can find the median. For example:

Sorting the above list in the numerical order

13, 13, 13, 13, 14, 14, 16, 18, 21

There are nine numbers in the list, so the middle one will be the $(9 + 1) \div 2 = 10 \div 2 = 5$th number:

13, 13, 13, 13, ***14***, 14, 16, 18, 21

Mode

The "mode" is the value that occurs most often. If no number in the list is repeated, then there is no mode for the list. For example

In the above list, the number that is most repeated is 13, so the mode of this data is **13.**

References:

1. https://www.purplemath.com/modules/meanmode.htm
2. https://diginomica.com/how-canadian-ai-start-bluedot-spotted-coronavirus-anyone-else-had-clue
3. https://www.kaggle.com/benhamner/python-data-visualizations