Machine Learning — Probability & Statistics

Essential Probability & Statistics for Machine Learning

Javaid Nabi
Jan 7, 2019 · 13 min read

Machine Learning is an interdisciplinary field that uses statistics, probability, algorithms to learn from data and provide insights which can be used to build intelligent applications. In this article, we will discuss some of the key concepts widely used in machine learning.

Probability and statistics are related areas of mathematics which concern themselves with analyzing the relative frequency of events.

> Probability deals with predicting the likelihood of future events, while statistics involves the analysis of the frequency of past events.

Probability

Most people have an intuitive understanding of degrees of probability, which is why we use words like “probably” and “unlikely” in our daily conversation, but we will talk about how to make quantitative claims about those degrees [1].

In probability theory, an **event** is a set of outcomes of an experiment to which a probability is assigned. If `E` represents an event, then `P(E)` represents the probability that `E` will occur. A situation where `E` might happen (*success*) or might not happen (*failure*) is called a ***trial***.

This event can be anything like *tossing a coin, rolling a die* or *pulling a colored ball out of a bag*. In these examples the outcome of the event is random, so the variable that represents the outcome of these events is called a **random variable.**

Let us consider a basic example of tossing a coin. If the coin is fair, then it is just as likely to come up heads as it is to come up tails. In other words, if we were to repeatedly toss the coin many times, we would expect about about half of the tosses to be heads and and half to be tails. In this case, we say that the probability of getting a head is 1/2 or 0.5 .

The **empirical probability** of an event is given by number of times the event occurs divided by the total number of incidents observed. If for `n` trials and we observe `s` successes, the probability of success is s/n. In the above example. any sequence of coin tosses may have more or less than exactly 50% heads.

Theoretical probability on the other hand is given by the number of ways the particular event can occur divided by the total number of possible outcomes. So a head can occur once and possible outcomes are two (head, tail). The true (theoretical) probability of a head is 1/2.

Joint Probability

Probability of events A and B denoted by `P(A and B) or P(A ∩ B)` is the probability that events A and B both occur. `P(A ∩ B) = P(A). P(B)` . This only applies if `A` and `B` are independent, which means that if `A` occurred, that doesn't change the probability of `B` , and vice versa.

Conditional Probability

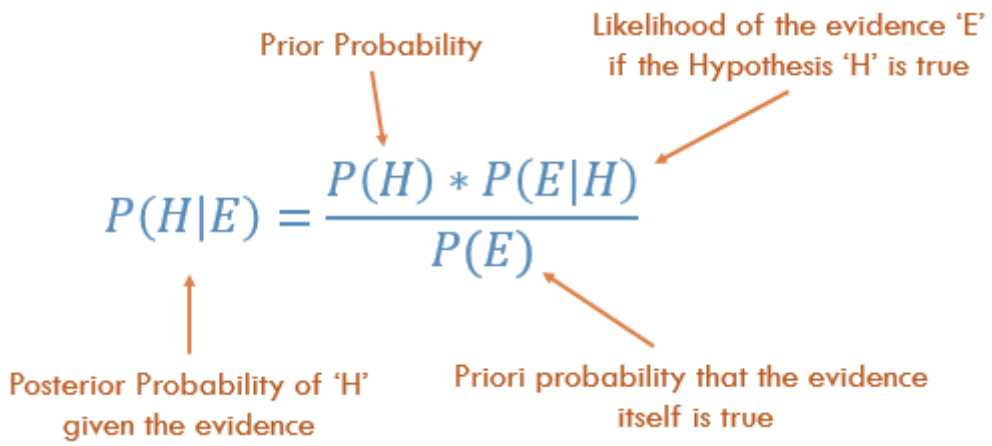
Let us consider A and B are not independent, because if A occurred, the probability of B is higher. When A and B are not independent, it is often useful to compute the conditional probability, P (A|B), which is the probability of A given that B occurred: `P(A|B) = P(A ∩ B)/ P(B)` .

> The probability of an event A conditioned on an event B is denoted and defined `P(A|B) = P(A∩B)/P(B)`

Similarly, `P(B|A) = P(A ∩ B)/ P(A)` . We can write the joint probability of as A and B as `P(A ∩ B)= p(A).P(B|A)` , which means : *"The chance of both things happening is the chance that the first one happens, and then the second one given the first happened."*

Bayes' Theorem

Bayes's theorem is a relationship between the conditional probabilities of two events. For example, if we want to find the probability of selling ice cream on a hot and sunny day, Bayes' theorem gives us the tools to use prior knowledge about the likelihood of selling ice cream on any other type of day (rainy, windy, snowy etc.).



where H and E are events, $P(H|E)$ is the conditional probability that event H occurs given that event E has already occurred. The probability $P(H)$ in the equation is basically frequency analysis; given our ***prior*** *data* what is the probability of the event occurring. The $P(E|H)$ in the equation is called the ***likelihood*** and is essentially the probability that the evidence is correct, given the information from the frequency analysis. $P(E)$ is the probability that the actual **evidence** is true.

Let H represent the event that we sell ice cream and E be the event of the weather. Then we might ask *what is the probability of selling ice cream on any given day given the type of weather?* Mathematically this is written as `P(H=ice cream sale | E= type of weather)` which is equivalent to the left hand side of the equation. $P(H)$ on the right hand side is the expression that is known as the **prior** because we might already know the marginal probability of the sale of ice cream**.** In our example this is $P(H = ice\ cream\ sale)$, i.e. the probability of selling ice cream regardless of the type of weather outside. For example, I could look at data that said 30 people out of a potential 100 actually bought ice cream at some shop somewhere. So my $P(H = ice\ cream\ sale) = 30/100 = 0.3$, *prior to me knowing anything about the weather*. This is how Bayes' Theorem allows us to incorporate prior information [2].

A classic use of Bayes's theorem is in the interpretation of clinical tests. Suppose that during a routine medical examination, your doctor informs you that you have tested positive for a rare disease. You are also aware that there is some uncertainty in the results of these tests. Assuming we have a **Sensitivity** (also called the **true positive rate)** result for 95% of the patients with the disease, and a **Specificity** (also called the **true negative rate**) result for 95% of the healthy patients.

If we let "+" and "−" denote a positive and negative test result, respectively, then the test accuracies are the conditional probabilities : $P(+|disease) = 0.95$, $P(-|healthy) = 0.95$,

In Bayesian terms, we want to compute the probability of disease given a positive test, $P(disease|+)$.

$$P(disease|+) = P(+|disease) * P(disease) / P(+)$$

How to evaluate `P(+)` **, all positive cases ?** We have to consider two possibilities, `P(+|disease)` and `P(+|healthy)` . The probability of a false positive, `P(+|healthy)` , is the complement of the `P(-|healthy)`. Thus `P(+|healthy) = 0.05` .

$$P(\text{disease}|+) = \frac{P(+|\text{disease})P(\text{disease})}{P(+|\text{disease})P(\text{disease}) + P(+|\text{healthy})P(\text{healthy})}$$

Importantly, Bayes' theorem reveals that in order to compute the conditional probability that you have the disease given the test was positive, you need to know the "prior" probability you have the disease `P(disease)` , given no information at all. That is, you need to know the overall incidence of the disease in the population to which you belong. Assuming these tests are applied to a population where the actual disease is found to be 0.5%, `P(disease)= 0.005` which means `P(healthy) = 0.995`.

So, `P(disease|+) = 0.95 * 0.005 /(0.95 * 0.005 + 0.05 * 0.995) = 0.088`

In other words, despite the apparent reliability of the test, the probability that you actually have the disease is still less than 9%. Getting a positive result increases the probability you have the disease. But it is incorrect to interpret the 95 % test accuracy as the probability you have the disease.

Descriptive Statistics

Descriptive statistics refers to methods for summarizing and organizing the information in a data set. We will use below table to describe some of the statistical concepts [4].

Characteristics of 10 loan applicants

Applicant	Marital Status	Mortgage	Income ($)	Income Rank	Year	Risk
1	Single	y	38,000	2	2009	Good
2	Married	y	32,000	7	2010	Good
3	Other	n	25,000	9	2011	Good
4	Other	n	36,000	3	2009	Good
5	Other	y	33,000	4	2010	Good
6	Other	n	24,000	10	2008	Bad
7	Married	y	25,100	8	2010	Good
8	Married	y	48,000	1	2007	Good
9	Married	y	32,100	6	2009	Bad
10	Married	y	32,200	5	2010	Good

Elements: The entities for which information is collected are called the elements. In the above table, the elements are the 10 applicants. Elements are also called cases or subjects.

Variables: The characteristic of an element is called a variable. It can take different values for different elements.e.g., marital status, mortgage, income, rank, year, and risk. Variables are also called attributes.

Variables can be either **qualitative** or **quantitative**.

Qualitative: A qualitative variable enables the elements to be classified or categorized according to some characteristic. The qualitative variables are `marital status`, `mortgage`, `rank`, and `risk`. Qualitative variables are also called **categorical** variables.

Quantitative: A quantitative variable takes numeric values and allows arithmetic to be meaningfully performed on it. The quantitative variables are `income` and `year`. Quantitative variables are also called **numerical** variables.

Discrete Variable: A numerical variable that can take either a finite or a countable number of values is a discrete variable, for which each value can be graphed as a separate point, with space between each point. `'year'` is an example of a discrete variable..

Continuous Variable: A numerical variable that can take infinitely many values is a continuous variable, whose possible values form an interval on the number line, with no space between the points. `'income'` is an example of a continuous variable.

Population: A population is the set of all elements of interest for a particular problem. A parameter is a characteristic of a population.

Sample: A sample consists of a subset of the population. A characteristic of a sample is called a statistic.

Random sample: When we take a sample for which each element has an equal chance of being selected.

Measures of Center: Mean, Median, Mode, Mid-range

> *Indicate where on the number line the central part of the data is located.*

Mean

The mean is the arithmetic average of a data set. To calculate the mean, add up the values and divide by the number of values.The sample mean is the arithmetic average of a sample, and is denoted $\bar{x}$ ("x-bar"). The population mean is the arithmetic average of a population, and is denoted μ ("myu", the Greek letter for m).

Median

The median is the middle data value, when there is an odd number of data values and the data have been sorted into ascending order. If there is an even number, the median is the mean of the two middle data values. When the income data are sorted into ascending order, the two middle values are $32,100 and $32,200, the mean of which is the median income, $32,150.

Mode

The mode is the data value that occurs with the greatest frequency. Both quantitative and categorical variables can have modes, but only quantitative variables can have means or medians. Each income value occurs only once, so there is no mode. The mode for year is 2010, with a frequency of 4.

Mid-range

The mid-range is the average of the maximum and minimum values in a data set. The mid-range income is:

```
mid-range(income) = (max(income) + min(income))/2 = (48000 +
24000)/2 = $36000
```

Measures of Variability: Range, Variance, Standard Deviation

> *Quantify the amount of variation, spread or dispersion present in the data.*

Range

The range of a variable equals the difference between the maximum and minimum values. The range of income is:

```
range(income) = max (income) - min (income) = 48,000 - 24,000
=$24000
```

Range only reflects the difference between largest and smallest observation, but it fails to reflect how data is centralized.

Variance

Population variance is defined as the average of the squared differences from the Mean, denoted as σ^2 (“sigma-squared”):

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

Larger Variance means the data are more spread out.

The sample variance s^2 is approximately the mean of the squared deviations, with N replaced by $n-1$. This difference occurs because the sample mean is used as an approximation of the true population mean.

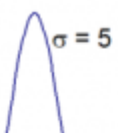
$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

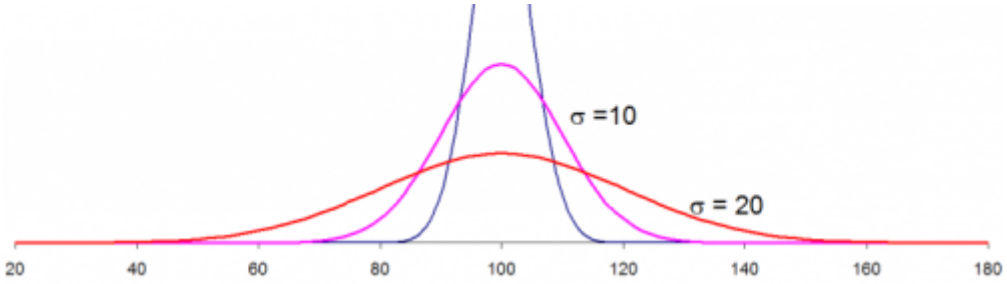
Standard Deviation

The *standard deviation* or sd of a bunch of numbers tells you how much the individual numbers tend to differ from the mean.

The sample standard deviation is the square root of the sample variance: $sd = \sqrt{s^2}$. For example, incomes deviate from their mean by $7201.

The population standard deviation is the square root of the population variance: $sd = \sqrt{\sigma^2}$.





Three different data distributions with same mean (100) and different standard deviation (5,10,20)

The smaller the standard deviation, narrower the peak, the data points are closer to the mean. The further the data points are from the mean, the greater the standard deviation.

Measures of Position: Percentile, Z-score, Quartiles

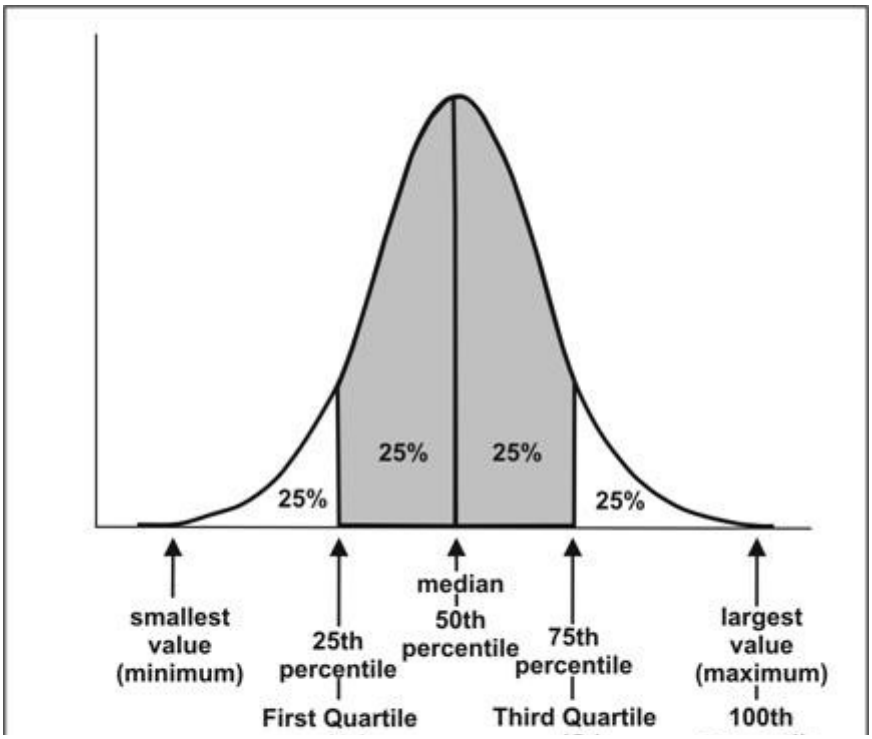
> *Indicate the relative position of a particular data value in the data distribution.*

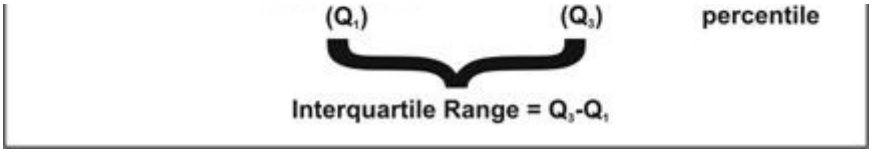
Percentile

The pth percentile of a data set is the data value such that p percent of the values in the data set are at or below this value. The 50th percentile is the median. For example, the median income is $32,150, and 50% of the data values lie at or below this value.

Percentile rank

The percentile rank of a data value equals the percentage of values in the data set that are at or below that value. For example, the percentile rank. of Applicant 1's income of $38,000 is 90%, since that is the percentage of incomes equal to or less than $38,000.





Interquartile Range (IQR)

The first quartile (Q1) is the 25th percentile of a data set; the second quartile (Q2) is the 50th percentile (median); and the third quartile (Q3) is the 75th percentile.

The **IQR** measures the difference between 75th and 25th observation using the formula: IQR = Q3 − Q1.

A data value x is an outlier if either `x ≤ Q1 - 1.5(IQR), or x ≥ Q3 + 1.5(IQR).`

Z-score

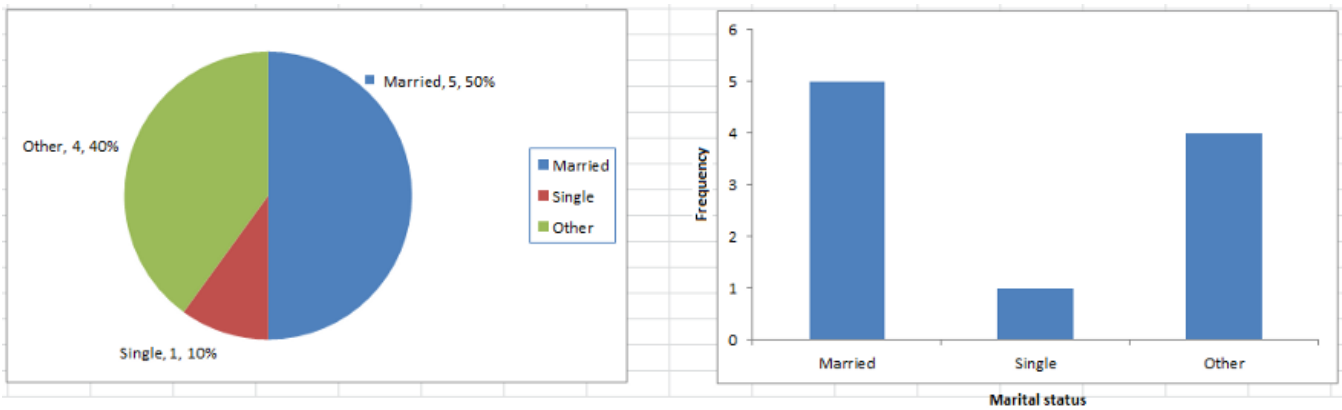
The Z-score for a particular data value represents how many standard deviations the data value lies above or below the mean.

$$Z\text{-score} = \frac{x - \bar{x}}{s}$$

So, If z is positive, it means that the value is above the average. For Applicant 6, the Z-score is (24,000 − 32,540)/ 7201 ≈ −1.2, which means the income of Applicant 6 lies 1.2 standard deviations below the mean.

Uni-variate Descriptive Statistics

Different ways you can describe patterns found in uni-variate data include central tendency : mean, mode and median and dispersion: range, variance, maximum, minimum, quartiles , and standard deviation.



Pie chart [left] & Bar chart [right] of Marital status from loan applicants table.

The various plots used to visualize uni-variate data typically are Bar Charts, Histograms, Pie Charts. etc.

Bi-variate Descriptive Statistics

Bi-variate analysis involves the analysis of two variables for the purpose of determining the empirical relationship between them. The various plots used to visualize bi-variate data typically are scatter-plot, box-plot.

Scatter Plots

The simplest way to visualize the relationship between two quantitative variables , x and y. For two continuous variables, a *scatter-plot* is a common graph. Each (x, y) point is graphed on a Cartesian plane, with the x axis on the horizontal and the y axis on the vertical. Scatter plots are sometimes called correlation plots because they show how two variables are correlated.

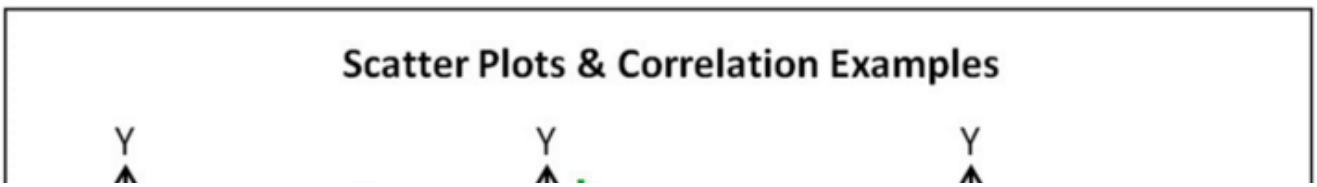
Correlation

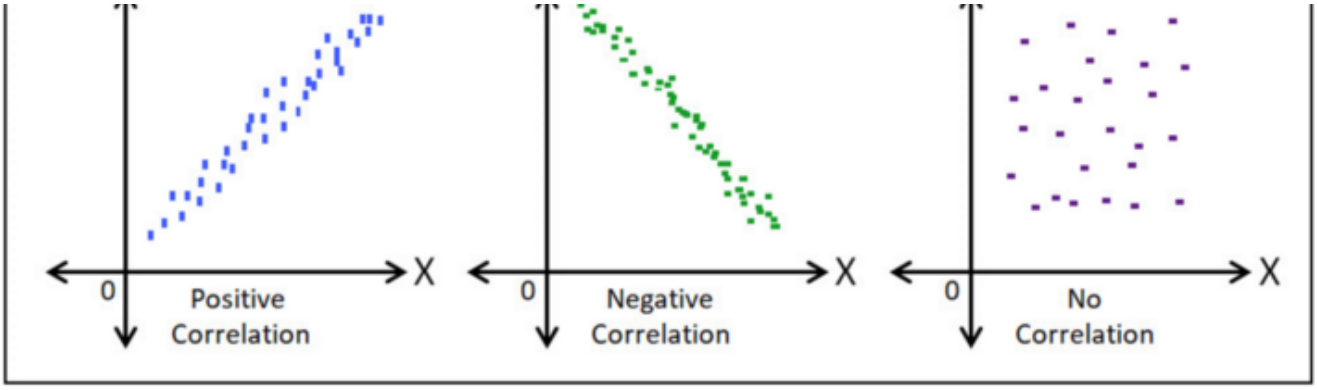
A correlation is a statistic intended to quantify the strength of the relationship between two variables. The **correlation coefficient** `r` quantifies the strength and direction of the linear relationship between two quantitative variables. The correlation coefficient is defined as:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n-1)\, s_x s_y}$$

where `sx` and `sy` represent the standard deviation of the x-variable and the y-variable, respectively. `-1 ≤ r ≤ 1` .

> *If r is positive and significant, we say that x and y are **positively correlated**. An increase in x is associated with an increase in y.*
>
> *If r is negative and significant, we say that x and y are **negatively correlated**. An increase in x is associated with a decrease in y.*



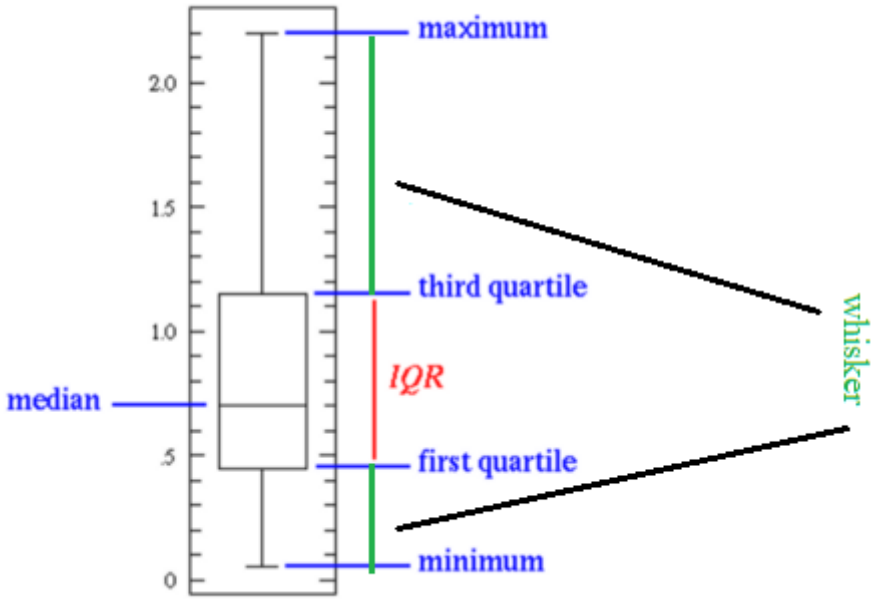


Positive correlation (r > 0), Negative correlation (r < 0), No correlation (r = 0)

Box Plots

A box plot is also called a box and whisker plot and it's used to picture the distribution of values. When one variable is categorical and the other continuous, a box-plot is commonly used. When you use a box plot you divide the data values into four parts called quartiles. You start by finding the median or middle value. The median splits the data values into halves. Finding the median of each half splits the data values into four parts, the quartiles.

Each box on the plot shows the range of values from the median of the lower half of the values at the bottom of the box to the median of the upper half of the values at the top of the box. A line in the middle of the box occurs at the **median** of all the data values. The **whiskers** then point to the largest and smallest values in the data.



The five-number summary of a data set consists of the `minimum`, `Q1`, the `median`, `Q3`, and the `maximum`.

> Box plots are especially useful for indicating whether a distribution is skewed and whether there are potential unusual observations (outliers) in the data set.

The left whisker extends down to the minimum value which is not an outlier. The right whisker extends up to the maximum value that is not an outlier. When the left whisker is longer than the right whisker, then the distribution is left-skewed and vice versa. When the whiskers are about equal in length, the distribution is symmetric.

Conclusion

Basic concepts of probability and statistics are a must have for anyone interested in machine learning. I covered briefly some of the essential concepts that are mostly used in machine learning. I hope you enjoyed this post and learned something new and useful.

Thank you for reading.

References

[1] http://greenteapress.com/thinkstats/thinkstats.pdf

[2] https://seeing-theory.brown.edu/basic-probability/index.html

[3] https://towardsdatascience.com/probability-concepts-explained-bayesian-inference-for-parameter-estimation-90e8930e5348

[4] https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118874059.app1

[5] https://medium.com/analytics-vidhya/descriptive-statistics-for-data-science-2f304a36ac34