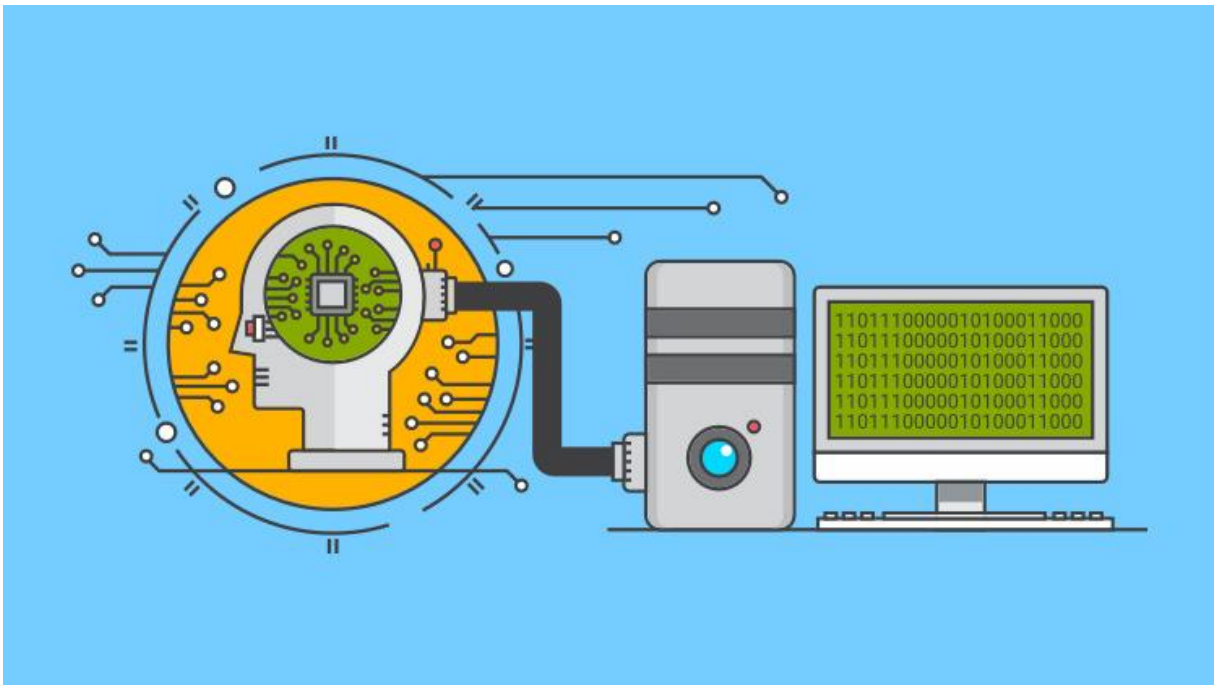


Machine Learning and Data Science

Iheb BOUARICHE | MACHINE LEARNING AND DATA SCIENCE | 16.04.2023

What is DATA?

Data refers to the raw facts, figures, and statistics collected or stored in a structured or unstructured format. In today's world, data is generated at an unprecedented rate, and it is considered to be the backbone of modern-day businesses and organizations.

Data can take various forms, such as text, numbers, images, videos, and audio. And it can be generated from various sources, such as sensors, social media, web sites, customer feedback and others.

The importance of data lies in its ability to provide insights, patterns, and trends that can be used to make informed decisions, solve problems, and optimize processes. However, to derive meaningful insights from data, it needs to be processed, analyzed, and visualized using various tools and techniques. The field of data science and analytics deals with extracting insights from data to drive business decisions and solve real-world problems.

Advances in computing technologies have led to the advent of big data, which usually refers to very large quantities of data, usually at the petabyte scale. Using traditional data analysis methods and computing, working with such large and growing datasets is difficult, even impossible. The relatively new field of data science uses machine learning (and other artificial intelligence (AI)) methods that allow for efficient applications of analytic methods to big data.

Types of data:

1 - Structured data: This refers to data that is organized in a structured manner, usually in the form of tables with clearly defined columns and rows. Examples of structured data include data stored in databases or spreadsheets.

2 - Unstructured data: This refers to data that does not have a predefined structure or format. Examples of unstructured data include text documents, images, videos, and social media posts.

Categories of data:

1 - Qualitative data: This type of data is descriptive and subjective, and it cannot be measured numerically. Examples of qualitative data include opinions, preferences, and emotions.

2 - Quantitative data: This type of data is numerical, and it can be measured and analyzed using mathematical and statistical methods. Examples of quantitative data include numerical measurements such as height, weight, and temperature.

What is statistic?

Statistics is the field of study that deals with the collection, analysis, interpretation, presentation, and organization of data. It involves using mathematical and computational tools to draw meaningful insights and conclusions from data. Statistics is widely used in a variety of fields, including scientific research, social sciences, business, healthcare, and machine learning.

1- What's a population on statistics: a population is the entire group of individuals or objects that a researcher is interested in studying. For example, if a researcher wants to study the average height of all adult males in a country, then the population would be all adult males in that country.
2- What's sample on statistics: sample is a smaller group of individuals or objects that are selected from the population. The goal of selecting a sample is to gather enough information about the population to draw conclusions, without having to study every individual in the

population. For example, instead of measuring the height of every adult male in the country, the researcher might select a sample of 1,000 adult males to study. The process of selecting a sample is called sampling technique.

Sampling techniques:

1- Random Sampling: each member of the population has equal chance of being selected in the sample.
2- Systematic Sampling: every nth record is chosen from the population to be a part of the sample.
3- Stratified Sampling: is a subset of the population that shares at least one common characteristic. Then random sampling used to select from each stratum.

Types of Statistics:

1- **Descriptive statistics:** involves collecting, analyzing, and presenting data in a way that summarizes and describes the main features of the data, such as measures of central tendency (mean, median, mode), measures of variability (range, standard deviation), and graphical representations (histograms, box plots, etc.). Descriptive statistics are used to gain insights and understand the characteristics of a dataset.
   I. Measures of Central tendency:
      a) Mean is the average of a set of numerical data.
      b) Median is the middle value in a dataset when the data is arranged in order.
      c) Mode is the value that appears most frequently in a dataset.
   II. Measures of Variability:
      a) Range: is the difference between the maximum and minimum values in the dataset
      b) Inter Quartile Range: is a measure of the spread of the middle 50% of a dataset, which is the difference between the 75th percentile (Q3) and the 25th percentile (Q1).
      c) Variance: The average of the squared differences from the Mean.
      d) Standard deviation: The Standard Deviation is a measure of how spread-out numbers are. it is the square root of the Variance.

   Concepts used on machine learning:

   - Entropy: measures the impurity or uncertainty present in the data. The entropy equation can be defined as follows:

$$\boldsymbol{H(X) = -\sum p(x)\ \log_2 p(x)}$$

   - Information gain: IG indicates how much "information" a particular feature or variable gives us about the final outcome.

$$\boldsymbol{IG(D,F) = H(D) - H(D|F)}$$

   Where ***IG(D,F)*** is the information gain of a feature **F** with respect to a dataset **D**, **H(D)** is the entropy of the dataset **D**, and ***H(D|F)*** is the conditional entropy of **D** given **F**.

   - Confusion matrix: A confusion matrix is a table used to evaluate the performance of a machine learning algorithm for a classification problem. It compares the predicted output of the model with the true output and displays the number of true positives, true negatives, false positives, and false negatives in a table or a matrix.

A confusion matrix example:

- There are two possible predicted classes: "yes" and "no"
- The classifier made a total of 165 predictions
- Out of those 165 cases, the classifier predicted "yes" 110 times, and "no" 55 times
- In reality, 105 patients in the sample have the disease, and 60 patients do not

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

In this case, we can define the accuracy by summing the number of true positive and true negative predictions and dividing it by the total number of cases.

Check [Machine Learning Full Course - Learn Machine Learning 10 Hours | Machine Learning Tutorial | Edureka] provided by Edureka to understand these concepts and with examples.

What is probability?

Probability is a measure of the likelihood of an event occurring. It's the ratio of desired outcomes to total outcomes.

Random Experiment: A random experiment is a process or an activity whose outcome cannot be predicted with certainty. It is a process that produces one or more outcomes, and the outcome cannot be determined in advance. Examples of random experiments include tossing a coin.

Sample Space: the sample space is the set of all possible outcomes of a random experiment. It is denoted by the symbol S, and it represents the total set of all possible results of an experiment. For example, consider the experiment of tossing a fair coin. The sample space for this experiment is S = {heads, tails}.

Event: an event is a subset of the sample space of a random experiment. It represents a set of possible outcomes that can occur in the experiment. For example, if we toss a coin, the sample space is {heads, tails}. An event could be "getting heads,"

Probability distribution: a probability density function (PDF) is a mathematical function that describes the relative likelihood of different outcomes in a continuous random variable. Unlike discrete random variables, which have a finite number of possible outcomes, continuous random variables can take on any value within a certain range. For example, the height of a person is a continuous random variable that can take on any value between a certain minimum and maximum height. The PDF of a continuous random variable gives the probability of the variable taking on a value within a certain range.

Normal distribution (Gaussian distribution): is a specific type of PDF and it is a bell-shaped curve that is symmetric around the mean value, with most values clustered around the mean and fewer values further away from the mean. the mean is denoted by (μ) and the standard deviation denoted by (σ). The mean represents the center of the distribution, while the standard deviation represents the spread or variability of the distribution.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Graphically, if the standard deviation is large, the curve will be short and wide, and if the standard deviation is small, the curve will be tall and narrow.

Central Limit Theorem: is the distribution of sample from a population and it must have the same mean as the population's mean, the sampling distribution of the mean of any independent random variable will be normal or nearly normal, if the sample size is large enough.

Types of probabilities:

a) Marginal probability: is the probability of occurrence of a single event.
b) Joint probability: is the measure of two event happening at the same time.

More specifically, if we have two events A and B, the joint probability of A and B, denoted as P(A and B), is the probability that both A and B occur together. On the other hand, the marginal probability of A, denoted as P(A), is the probability that A occurs independently, regardless of whether or not B occurs.

c) Conditional probability: is a probability of an event or outcome based on the occurrence of a previous event or outcome. Conditional probability of an event A is the probability that the event will occur given that an event B has already occurred. It is denoted as P(A|B), which is read as "the probability of A given B."

For example, consider the roll of two dice. The probability of rolling a 6 on the first die is 1/6. If we know that the sum of the two dice is 8, what is the probability that the first die is a 6? This can be expressed as P(6 on first die | sum is 8) and can be calculated using conditional probability. Conditional probability is calculated using the formula:

P(A|B) = P(A and B) / P(B) = (1/36) / (5/36) = 1/5

P(B) = 5/36: because there are five ways to get a sum of 8.

Note: that if A and B are independent events then the expression for conditional probability is given by: **P(A|B) = P(B).**

Bayes Theorem: shows the relation between one conditional probability and its inverse. It's derived from the conditional probability equation.

$$P(A|B) = P(B|A) * P(A) / P(B)$$

Check this YouTube video to understand the concept of Bayes Theorem with an example:

https://www.youtube.com/watch?v=XQoLVl31ZfQ

2- **Inferential statistics:** involves making inferences or generalizations about a population based on a sample of data. This can be achieved by Point estimation or interval estimation methods.

- **Point estimation:** is concerned with the use of the sample data to measure a single value which serves as an approximate value or the best estimate of an unknown population parameter. For example, if we want to estimate the mean of the population, we use only a sample from a data to estimate it. The question now is how to find the estimate value? The answer is that there are multiple methods to solve this problem, some of these methods are:

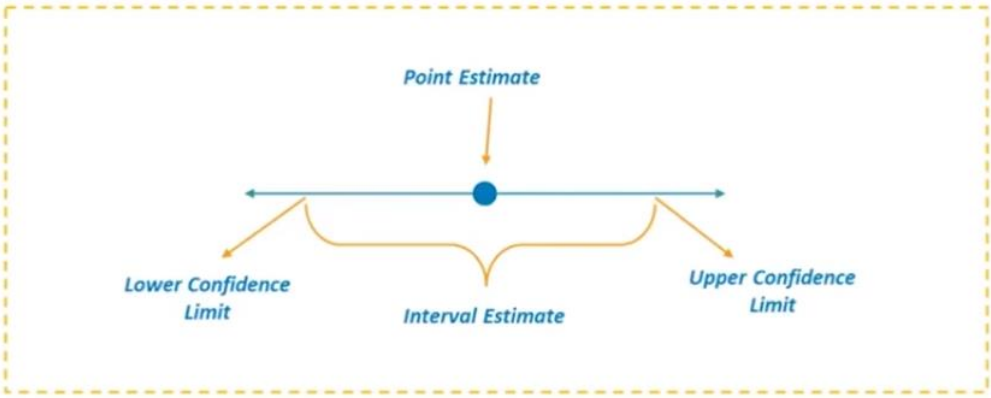
1- Method of Moment: estimates are found out by equating the first k sample moments (for example: mean, variance, etc.…) to the corresponding k population moments.

2- Maximum of Likelihood: uses a model and the values in the model to maximize a likelihood function. This results in the most likely parameter for the inputs selected (this method on probabilities)

3- Bayes' Estimators: minimizes the average risk (an expectation of random variables and this method is based on the Bayes' theorem).

4- Best Unbiased Estimator: several unbiased estimators can be used to approximate a parameter (which one is "best" depends on what parameter you are trying to find)

- **Interval estimate:** is an interval, or range of values, used to estimate a population parameter.



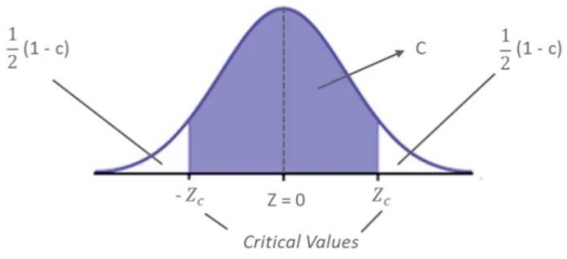
For example, let assume that I want to estimate the mean value of a population, so we need to build a range that contain this mean value.

Note: that the main difference between interval and point estimate is that a point estimate is a single value that is used to estimate an unknown population parameter, whereas an interval estimate is a range of values that is used to estimate the parameter with a certain degree of confidence.

Confidence interval: is the measure of your confidence that the interval estimate contains the population parameter (for example the mean of the population)

Level of confidence: is the probability that the interval estimate contains the population parameter.

The difference between the point estimate and the actual population parameter value is called sampling error.



C is the area beneath the normal curve between the critical values
Corresponding Z score can be calculated using the standard normal table

Margin of error: E for a given level of confidence is the greatest possible distance between the point estimate and the value of the parameter it is estimating.

Hypothesis testing: statisticians use hypothesis testing to formally check whether the hypothesis is accepted or rejected with the following methods:

- ✓ State the hypotheses: this stage involves stating the null and alternative hypothesis. For example, if we found that after some numbers of samples that an events is don't occur and the probability of that event is very high. It maybe means that there is an error.
- ✓ Formulate an Analysis Plan: this stage involves the construction of an analysis plan.
- ✓ Analyze sample Data: this stage involves the calculation and interpretation of the test statistic as described in the analysis plan.
- ✓ Interpret results: this stage involves the application of the decision rule described in the analysis plan.

Check this YouTube video for explained examples: https://www.youtube.com/watch?v=XQoLVl31ZfQ

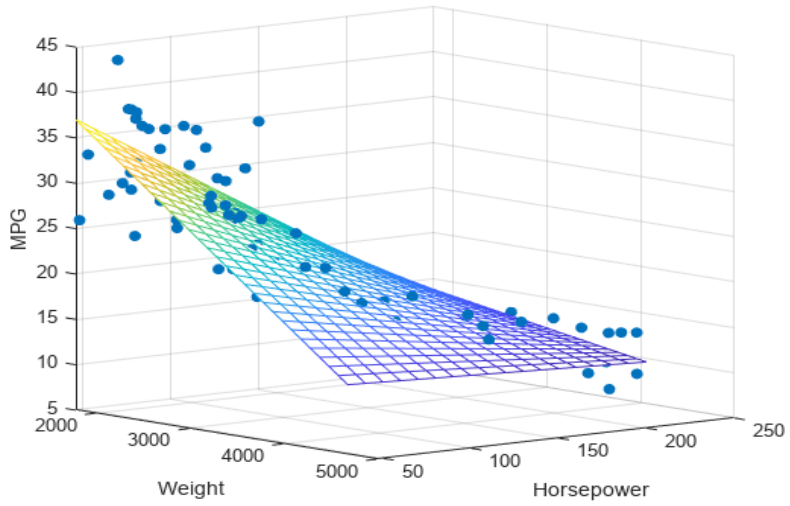
Machine Learning:

- ❖ **Supervised Learning:** is a type of machine learning in which an algorithm is trained on a labeled dataset. The input data and the desired output data are both provided to the algorithm during training. Supervised learning algorithms learns to map the input to the output by iteratively adjusting their parameters to minimize the difference between the predicted output and the true output. There are multiple used algorithms for machine learning and supervised learning, we will cover most of them in the next topics.

"It called supervised learning because the process of an algorithm learning from the training dataset can be thought as a teacher supervising the learning process."

1. **Linear Regression algorithm:**

Linear regression algorithm shows a linear relationship between a dependent (Y) and one or more independent (X) variable(s), hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable(s).



The linear equation for the predicted dependent value is defined as:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

Where:

x_n **:** are the independent variables.

$\hat{y}$ **:** is the predicted dependent variable.

b_n **:** are the coefficients of regression equation.

Cost function: For Linear Regression, we use the Mean Squared Error (MSE) cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

$$\mathbf{MSE} = \frac{1}{\mathbf{N}} \sum_{\mathbf{i=1}}^{\mathbf{N}} (y - \hat{y})^2$$

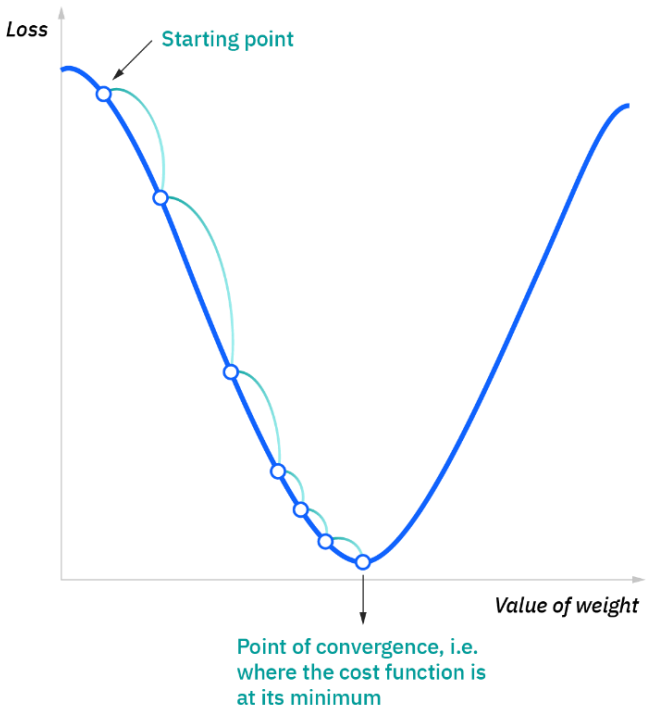
Where:

$\mathbf{N}$: is the number of data points.

$\mathbf{y_i}$: is the true value.

$\hat{\mathbf{y}}_\mathbf{i}$: is the predicted value.

Gradient descent: is an optimization algorithm used to minimize the MSE. It involves iteratively updating the values of the regression coefficients in the direction of the steepest descent of the cost function. This is done by computing the gradient of the cost function with respect to each coefficient and updating the coefficients in the direction that minimizes the MSE. The update rule involves multiplying the gradient by a learning rate and subtracting it from the current value of the coefficient. The process is repeated until convergence.



Gradient descent optimization algorithm:

$$b_i = b_{i-1} - \alpha \frac{d}{db_i} MSE$$

R-Squared method:

In linear regression, the R-squared value is a number between 0 and 1 that indicates how well the regression line fits the observed data. An R-squared value of 1 indicates a perfect fit, where all the variance in the dependent variable can be explained by the independent variable(s). An R-squared value of 0 indicates that none of the variance in the dependent variable can be explained by the independent variable(s). The R-Squared equation can be defined as follows:

$$R\_squared = \frac{\sum_{i=1}^{N}(y - \hat{y})^2}{\sum_{i=1}^{N}(y - \acute{y})^2}$$

Where:

$\mathbf{y_i}$: is the true value.

$\hat{\mathbf{y}}_\mathbf{i}$: is the predicted value.

$\acute{y}$: is the mean of true values

The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.

2. **Logistic Regression:**

Logistic regression is a statistical method used in machine learning for binary classification problems. It estimates the probability of a binary outcome based on the values of input variables. The logistic function, also known as the sigmoid function, maps the output of linear regression to a value between 0 and 1, representing the probability of the outcome being in a particular class. Therefore, logistic regression is useful for predicting the probability of a binary outcome based on one or more input variables.

Gradient descent is used to optimize the weights of the logistic regression model by minimizing the cost function, thereby improving the accuracy of the predicted probabilities for binary classification problems.

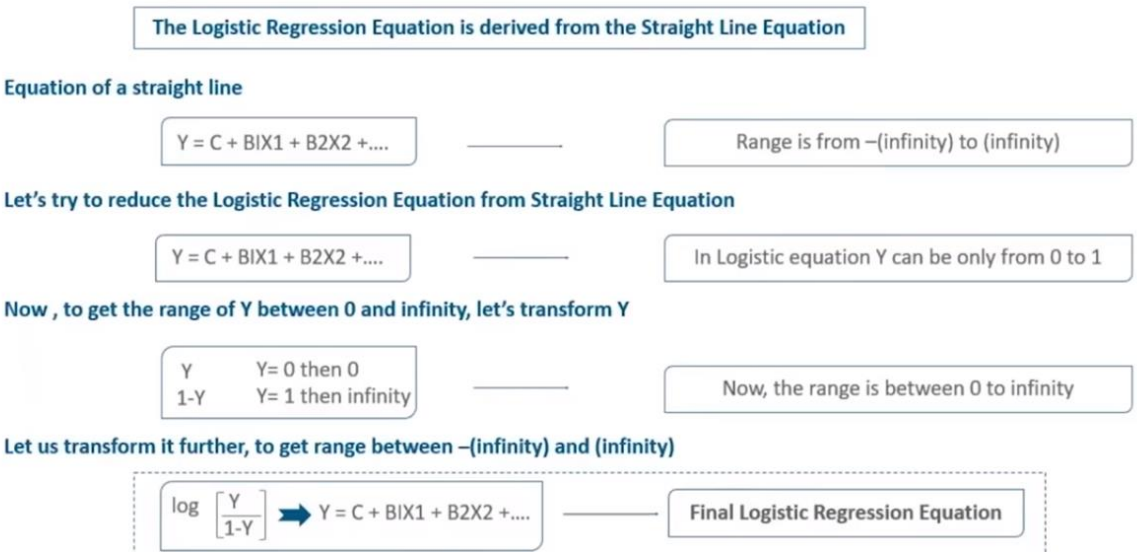


Figure 1: how to derive a logistic regression equation from linear equation.

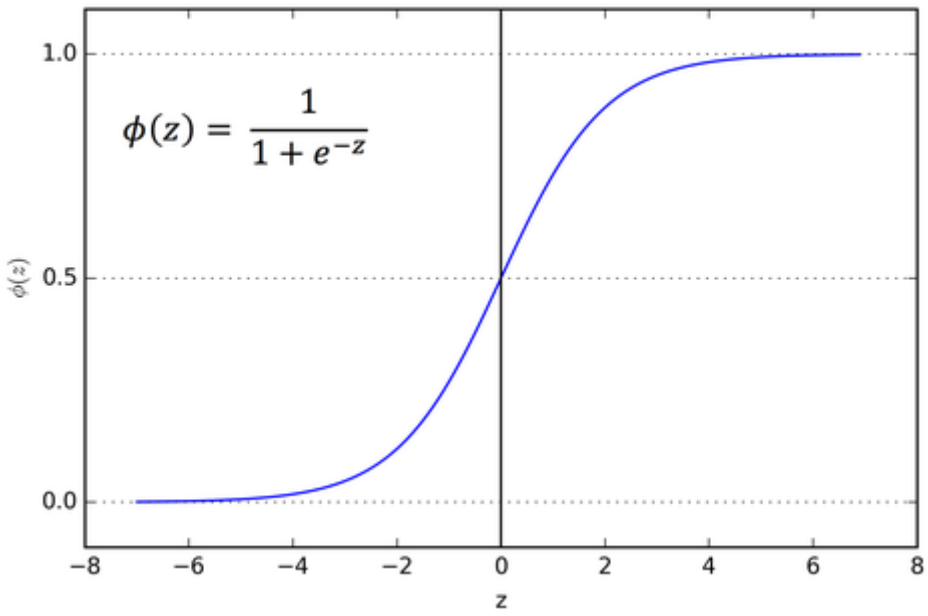


Figure 2: Sigmoid curve

Where: $\boldsymbol{z = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n}$

And: $\boldsymbol{b_n}$ **:** are the weights.

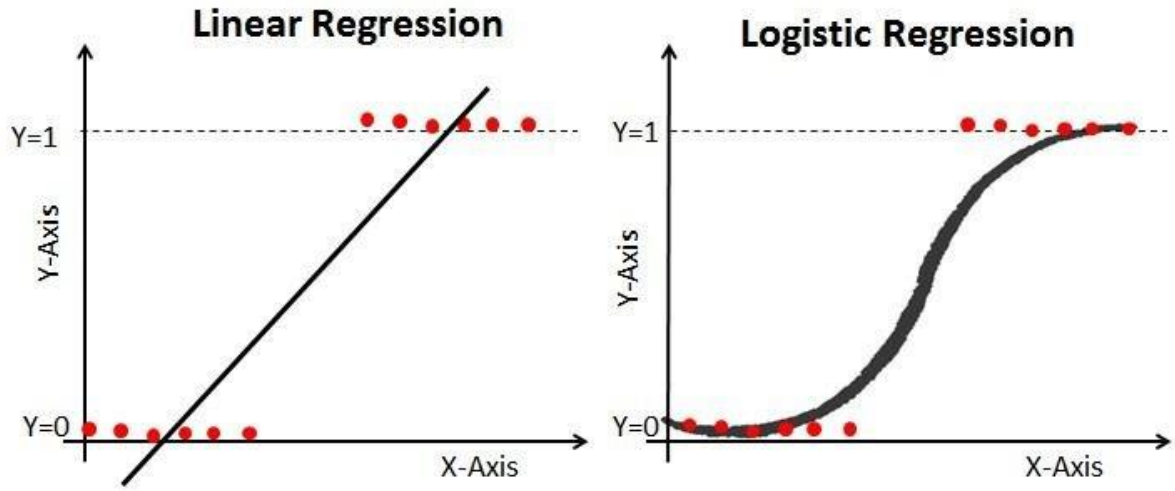


Figure 3: Logistic regression can fit very accurately on classification problems compared to linear regression.

Logistic regression uses a cost function called log loss (also known as cross-entropy). In logistic regression, we cannot use the mean squared error as the cost function because of the non-linearity of the logistic function used for modeling. Using mean squared error can result in a cost function with many local optima, which is not preferred. Therefore, we use the log loss (or cross-entropy) as the cost function for logistic regression. The log loss function penalizes confident and wrong predictions more heavily than less confident ones, which is desirable in classification tasks.

The Log loss function can be defined as follow:

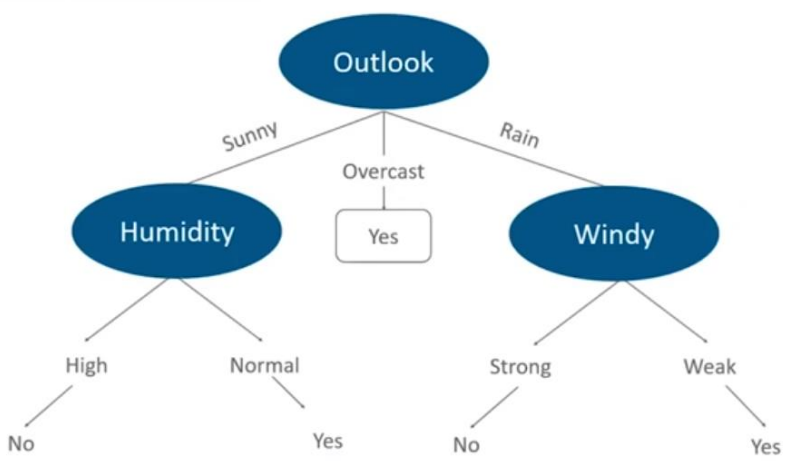
$$-\frac{1}{N}\sum_{i=1}^{n}[y_i \ log(p(\hat{y}_i)) + (1-y_i) \ log(1-p(\hat{y}_i))$$

3. Decision Tree:

A decision tree is a tree-shaped model used to make decisions or predictions about an outcome by mapping out all possible options and their potential consequences. The decision tree starts with a single node, known as the root node, which represents the entire dataset. From there, the tree branches off into multiple nodes, each representing a decision point or attribute. At each node, the algorithm evaluates a certain feature or attribute of the data and splits the dataset into subsets based on that feature. The decision tree continues to split the dataset into smaller subsets until it reaches a leaf node, which represents a final decision or prediction. Each leaf node represents a specific outcome or class, and the path from the root node to the leaf node represents the sequence of decisions that were made to arrive at that outcome. Decision trees can be used for both classification and regression tasks, and they have the advantage of being easy to interpret and visualize.

To select a root in a decision tree, we calculate the information gain between the parent node and its child nodes for each feature, and we select the feature that results in the highest information gain value.

Check the example in this video: https://www.youtube.com/watch?v=GwIo3gDZCVQ&t=14557s



outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Information Gain (IG) calculation was used to evaluate the importance of different features in the dataset, and the Outlook feature had the highest IG value, indicating that it is the most informative feature for predicting the output variable in comparison to the other features.

4. Random Forest:

Random forest is a machine learning algorithm that is used for classification, regression, and other tasks. It is an ensemble method that combines multiple decision trees to improve the accuracy of

predictions. In a random forest, each decision tree is trained on a random subset of the data and a random subset of the features, which helps to reduce overfitting and increase the diversity of the trees. The final prediction is made by aggregating the predictions of all the individual trees.

Check the examples in these URLs:

https://www.youtube.com/watch?v=v6VJ2RO66Ag

https://www.youtube.com/watch?v=GwIo3gDZCVQ&t=17199s

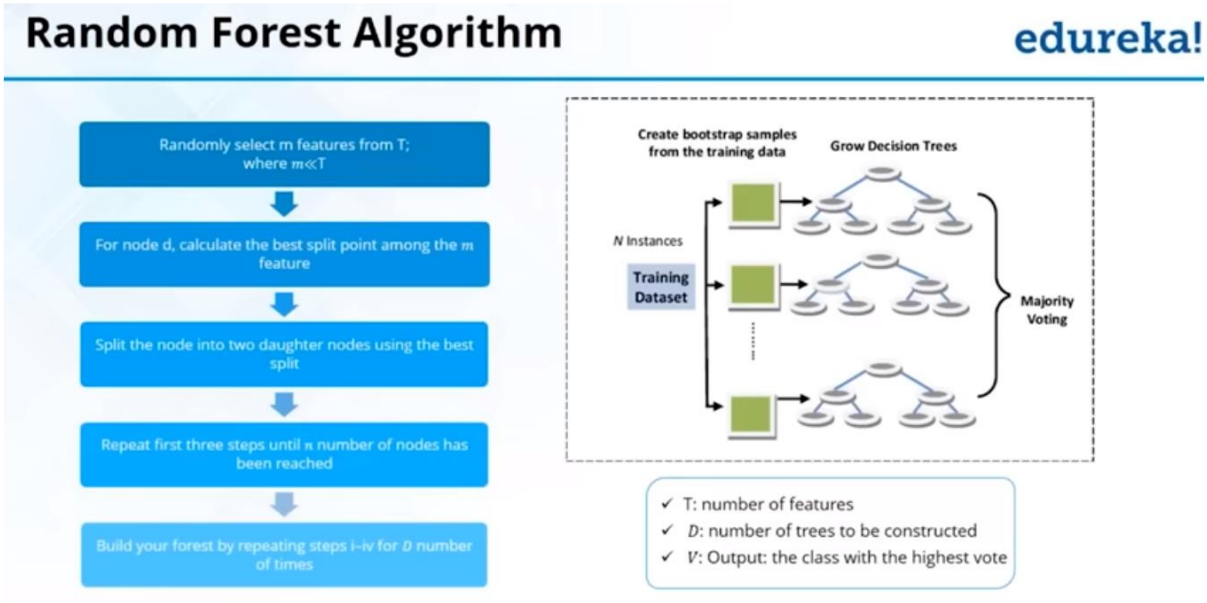


Figure 4: Random Forest algorithm from **Edureka!**

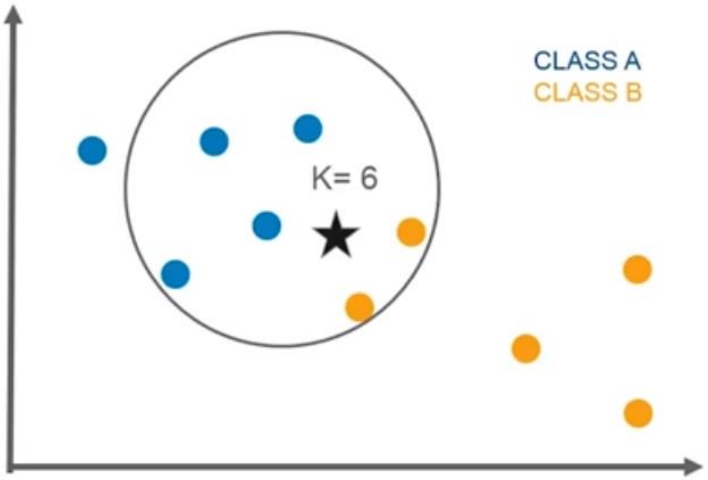
The advantages of Random Forest are: most accurate learning algorithm, work well for both classification and linear regression, runs efficiently on large dataset, requires almost no input preparation, can be easily grown in parallel, balancing error in unbalanced data sets.

5. **KNN Algorithm:**

KNN stands for K-Nearest Neighbors, which is a type of supervised machine learning algorithm used for classification and regression tasks.

The KNN algorithm is based on the principle that similar data points tend to be close to each other. Therefore, the algorithm assigns a new data point to the class of its K nearest neighbors in the feature space. In other words, the algorithm searches for the K nearest training examples to the new data point and assigns the most common class among those K neighbors to the new data point.

The value of K is typically chosen beforehand and determines the number of neighbors to consider for classification.



Check the example in this video: https://www.youtube.com/watch?v=GwIo3gDZCVQ&t=17199s

6. **Naïve Bayes:**

Naive Bayes classifier is a probabilistic algorithm based on Bayes' theorem, which describes the occurrence of an event, given another event. The Naive Bayes classifier assumes that the features used to classify an instance are conditionally independent.

The algorithm starts by calculating the conditional probabilities of each feature given each class label using the labeled training data. Then, we apply Bayes' theorem to calculate the posterior probability of each class label given the observed features of a new instance, and the class with the highest probability is assigned as the predicted class label for that instance.

In other words, the Naive Bayes classifier calculates the conditional probability of each feature given each class label from the training data, and uses these probabilities along with Bayes' theorem to predict the class label of new instances based on their observed feature values.

Naïve Bayes model is defined as follows:

$$\boldsymbol{y = arg}_{\boldsymbol{k\in\{1,\ldots\ldots,K\}}}\boldsymbol{max}\ \boldsymbol{p(C_k)}\prod_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{p(x_i|C_k)}$$

Where: y is the selected class output with the maximum posterior (the probability of an output given the inputs). And $\boldsymbol{p(C_k)}$ is the probability of a class. And $\boldsymbol{p(x_i|C_k)}$ is the probability of having a feature $\boldsymbol{x_i}$ given the class $\boldsymbol{C_k}$.

Gaussian Naive Bayes:

Gaussian Naive Bayes is a variant of the Naive Bayes classifier that handles continuous feature variables. Instead of computing the probability of each feature value given the class label, the algorithm calculates the probability density function of each continuous feature value given the class label using the mean and variance of the feature values of the corresponding class label in the training data. This probability density function is then used to compute the conditional probability of a new feature value given a class label. The algorithm then uses these conditional probabilities along with Bayes' theorem to predict the class label of new instances based on their observed feature values.

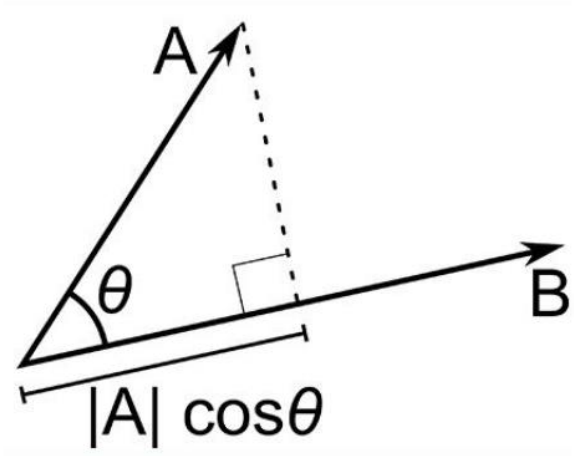
Check this web site for examples about Naïve Bayes:

https://www.geeksforgeeks.org/naive-bayes-classifiers/

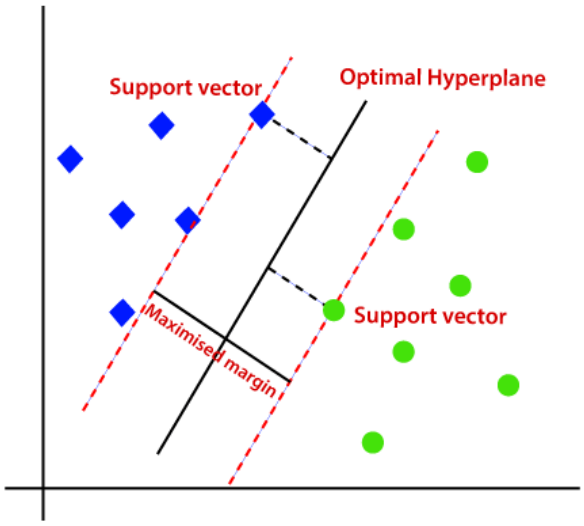
Note: that there exist other variants of the Naive Bayes classifier, such as Multinomial Naive Bayes and Bernoulli Naive Bayes, which are not covered in this paper.

7. **Support Vector Machine:**

First, we need to understand the Dot-Product. The dot product, also known as scalar product or inner product, is a mathematical operation that takes two vectors and returns a scalar value. The dot product can be expressed as the product of the magnitudes of two vectors and the cosine of the angle between them, so it can be defined as the projection of one vector along with another.



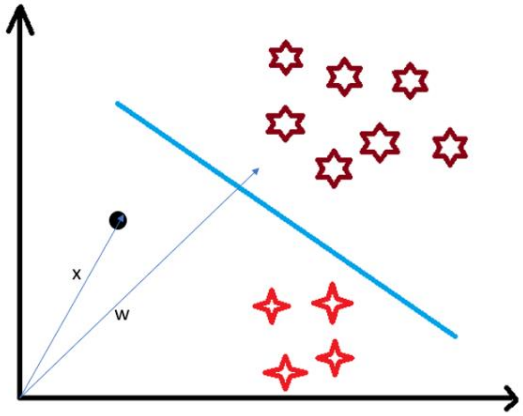
Support Vector Machine is a supervised machine learning algorithm used for classification and regression analysis. It works by finding the hyperplane that maximally separates two classes of data in a high-dimensional space. SVMs are effective for handling both linear and non-linearly separable data, and can handle high-dimensional data with a relatively small number of training samples.



A hyperplane can be defined as follows:

$$\boldsymbol{X}.\boldsymbol{W} + \boldsymbol{b} = \mathbf{0}$$

where: W: is a perpendicular vector to the hyperplane, X: is a vector from the origin to a random point from the data. And b is an offset (distance between the origin and the hyperplane along the w vector). We use the dot product to project X on W.



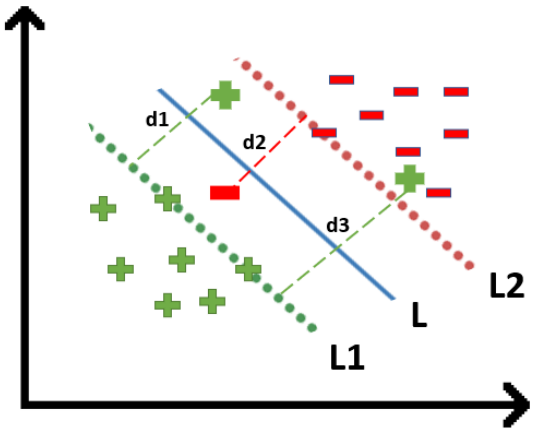
SVM aim to minimize the classification error by finding the hyperplane that has the largest margin, which is the distance between the hyperplane and the closest data points from each class.

the equation which we have to maximize is:

$$\boldsymbol{arrmax(W,b)\frac{2}{\|W\|}\ such\ that\ y_i(W.X+b)\geq 1}$$

We assume that negative classes have y=-1 and positive classes have y=1. This equation derived from maximizing the difference between two projected point on W from different classes.

Soft Margin SVM is based on an optimization problem that allows for a certain amount of misclassification of data points, by adding a parameter that controls the trade-off between maximizing the margin and minimizing the classification error.



In soft margin equation we add 2 more terms to the cost function of standard SVM, which are zeta multiplied by a hyperparameter ‘c’. and with inversing the cost function because it is common to minimize a cost function (by Gradient Descent for example).

$$\boldsymbol{argmin(W,X)\ \frac{\|W\|}{2}+c\sum_{i=1}^{n}\zeta_i}$$

Kernels in Support Vector Machine: In SVM, a kernel function is used to map the input data from the original feature space into a higher-dimensional space where it may become linearly separable. When we use a kernel function to map input data, we don't actually compute the coordinates of the mapped data points in the higher-dimensional space. Instead, we compute the dot product of the mapped feature vectors, which is a scalar value that depends only on the input data points and the chosen kernel function.

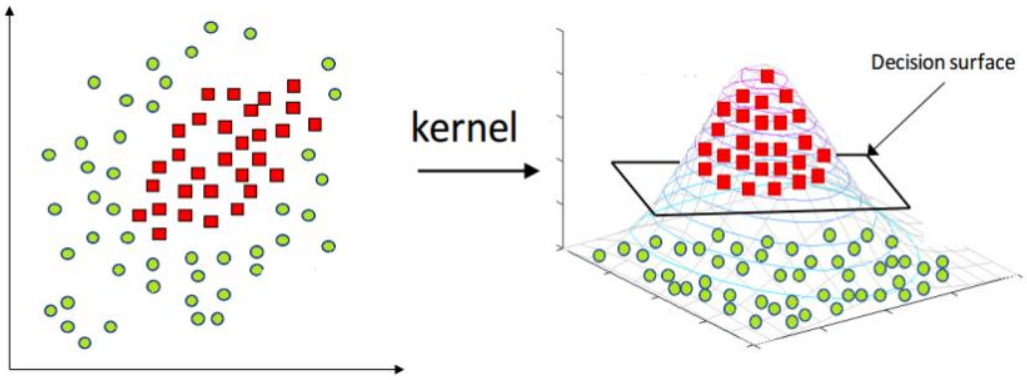


Figure: in the left side we have a non-linear separated data, as we see we can map the features from 2D spaces to higher 3D space using a Kernel.

Note: that there are different kernel functions, such as Polynomial, Sigmoid, RBF, Bessel function, Anova, and others…

The following URL contains a detailed explanation of SVM and Soft Margin SVM and kernels in SVM:

https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

This report is still a work in progress. unsupervised learning theory will be added soon.