MACHINE LEARNING

July 20, 2022

Practical File

BACHELORS OF TECHNOLOGY

Information Technology

SUBMITTED BY

AMANPREET KAUR

University Roll Number : 2104359

College Roll Number : 2121139

GURU NANAK DEV ENGINEERING COLLEGE

LUDHIANA -141006 , INDIA

Contents

1 Aim: Use of Python libraries

INTRODUCTION TO PYTHON LIBRARIES

A Python library is a collection of related modules. It contains bundles of code that can be used repeatedly in different programs. It makes Python Programming simpler and convenient for the programmer. As we don't need to write the same code again and again for different programs.

TYPES OF PYTHON LIBRARIES

NumPy
Pandas
Matplotlib
Scikit

NumPy : NumPy stands for Numerical Python.NumPy was created in 2005 by Travis Oliphant.NumPy is a Python library used for working with arrays.It is an open source project and you can use it freely.It also has functions for working in domain of linear algebra, fourier transform, and matrices.

Implemention :

```
import numpy as np
arr = np.array([1, 2, 3, 4, 5])
print(arr)
print(type(arr))
```

Output :

```
[1 2 3 4 5]
<class 'numpy.ndarray'>
```

2 Aim: Use of Pandas library

Pandas : Pandas is a Python library for data analysis. Started by Wes McKinney in 2008 out of a need for a powerful and flexible quantitative analysis tool, pandas has grown into one of the most popular Python libraries. Pandas acts as a wrapper over these libraries, allowing you to access many of matplotlib's and NumPy's methods with less code. Before pandas, most analysts used Python for data munging and preparation, and then switched to a more domain specific language like R for the rest of their workflow. Pandas introduced two new types of objects for storing data that make analytical tasks easier and eliminate the need to switch tools: Series, which have a list-like structure, and DataFrames, which have a tabular structure.

Python Pandas Data Structure The Pandas provides two data structures for processing the data, i.e., Series and DataFrame, which are discussed below:

1) **Series :** It is defined as a one-dimensional array that is capable of storing various data types. The row labels of series are called the index. We can easily convert the list, tuple, and dictionary into series using "series' method. A Series cannot contain multiple columns. It has one parameter : Data: It can be any list, dictionary, or scalar value.

Implemention :

```
import pandas as pd
import numpy as np
info = np.array(['P','a','n','d','a','s'])
a = pd.Series(info)
print(a)
```

Output :

```
0   P
1   a
2   n
3   d
4   a
5   s
dtype: object
```

Python Pandas DataFrame : It is a widely used data structure of pandas and works with a two-dimensional array with labeled axes (rows and columns). DataFrame is defined as a standard way to store data and has two different indexes, i.e., row index and column index. It consists of the following properties:

The columns can be heterogeneous types like int, bool, and so on.

It can be seen as a dictionary of Series structure where both the rows and columns are indexed. It is denoted as "columns" in case of columns and "index" in case of rows.

PROGRAM

Implemention :

```
import pandas as
x = ['Python', 'Pandas']
df = pd.DataFrame(x)
print(df)
```

Output :

```
      0
0   Python
1   Pandas
```

3 Aim: Use of Matplotlib Library

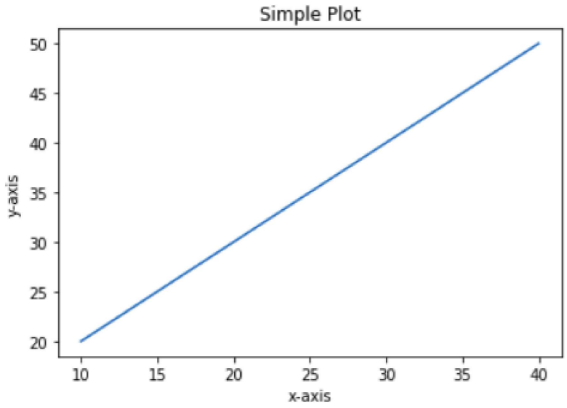
Matplotlib : Matplotlib is a Python library that helps in visualizing and analyzing the data and helps in better understanding of the data with the help of graphical, pictorial visualizations that can be simulated using the matplotlib library. Matplotlib is a comprehensive library for static, animated and interactive visualizations.

PROGRAM

Implemention :

```
import matplotlib.pyplot as plt
# initializing the data
x = [10, 20, 30, 40]
y = [20, 30, 40, 50]
# plotting the data
plt.plot(x, y)
# Adding the title
plt.title("Simple Plot")
# Adding the labels
plt.ylabel("y-axis")
plt.xlabel("x-axis")
plt.show()
```

Output :



4 Aim: Use of Scikit Library

Scikit : Scikit-learn is a key library for the Python programming language that is typically used in machine learning projects. Scikit-learn is focused on machine learning tools including mathematical, statistical and general purpose algorithms that form the basis for many machine learning technologies.Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

5 Aim: Introduction To Data-Science

Data Science is kinda blended with various tools, algorithms, and machine learning principles. Most simply, it involves obtaining meaningful information or insights from structured or unstructured data through a process of analyzing, programming and business skills. It is a field containing many elements like mathematics, statistics, computer science, etc. Those who are good at these respective fields with enough knowledge of the domain in which you are willing to work can call themselves as Data Scientist. It's not an easy thing to do but not impossible too. You need to start from data, it's visualization, programming, formulation, development, and deployment of your model. In the future, there will be great hype for data scientist jobs. Taking in that mind, be ready to prepare yourself to fit in this world.

Applications of Data Science

In Search Engines : The most useful application of Data Science is Search Engines. As we know when we want to search for something on the internet, we mostly used Search engines like Google, Yahoo, Safari, Firefox, etc. So Data Science is used to get Searches faster.

In Transport : Data Science also entered into the Transport field like Driverless Cars. With the help of Driverless Cars, it is easy to reduce the number of Accidents.

In Finance : Data Science plays a key role in Financial Industries. Financial Industries always have an issue of fraud and risk of losses. Thus, Financial Industries needs to automate risk of loss analysis in order to carry out strategic

decisions for the company. Also, Financial Industries uses Data Science Analytics tools in order to predict the future. It allows the companies to predict customer lifetime value and their stock market moves.

In E-Commerce : E-Commerce Websites like Amazon, Flipkart, etc. uses data Science to make a better user experience with personalized recommendations.

In Health Care : In the Healthcare Industry data science act as a boon. Data Science is used for:

Detecting Tumor

Drug discoveries

Medical Image Analysis

Virtual Medical Bots

Genetics and Genomics

Predictive Modeling for Diagnosis etc.

6 Aim: Artificial Intelligence

Artificial intelligence (AI) is intelligence demonstrated by machines, as opposed to the natural intelligence displayed by animals including humans. AI research has been defined as the field of study of intelligent agents, which refers to any system that perceives its environment and takes actions that maximize its chance of achieving its goals.[a]

The term "artificial intelligence" had previously been used to describe machines that mimic and display "human" cognitive skills that are associated with the human mind, such as "learning" and "problem-solving". This definition has since been rejected by major AI researchers who now describe AI in terms of rationality and acting rationally, which does not limit how intelligence can be articulated.

Types of Artificial Intelligence

Artificial intelligence can be divided into two different categories: weak and strong.

Weak artificial intelligence : It embodies a system designed to carry out one particular job. Weak AI systems include video games such as the chess example from above and personal assistants such as Amazon's Alexa and Apple's Siri. You ask the assistant a question, and it answers it for you.

Strong artificial intelligence : These systems are systems that carry on the tasks considered to be human-like. These tend to be more complex and complicated systems. They are programmed to handle situations in which they may be required to problem solve without having a person intervene. These kinds of systems can be found in applications like self-driving cars or in hospital operating rooms.

Applications of Artificial Intelligence

The applications for artificial intelligence are endless. The technology can be applied to many different sectors and industries. AI is being tested and used in the healthcare industry for dosing drugs and doling out different treatments tailored to specific patients, and for aiding in surgical procedures in the operating room.

Other examples of machines with artificial intelligence include computers that play chess and self-driving cars. Each of these machines must weigh the consequences of any action they take, as each action will impact the end result. In chess, the end result is winning the game. For self-driving cars, the computer system must account for all external data and compute it to act in a way that prevents a collision.Artificial intelligence also has applications in the financial industry, where it is used to detect and flag activity in banking and finance such as unusual debit card usage and large account deposits—all of which help a bank's fraud department. Applications for AI are also being used to help streamline and make trading easier. This is done by making supply, demand, and pricing of securities easier to estimate.

7 Aim: Machine Learning

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: The ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect.

Applications of Machine learning

Image Recognition: Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc. The popular use case of image recognition and face detection is, Automatic friend tagging suggestion:Facebook provides us a feature of auto friend tagging suggestion. Whenever we upload a photo with our Facebook friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning's face detection and recognition algorithm.

Speech Recognition : While using Google, we get an option of "Search by voice," it comes under speech recognition, and it's a popular application of machine learning.Google assistant, Siri, Cortana, and Alexa are using speech recognition technology to follow the voice instructions.

Traffic prediction: If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions.

Product recommendations: Machine learning is widely used by various e-commerce and entertainment companies such as Amazon, Netflix, etc., for product recommendation to the user. Whenever we search for some product on Amazon, then we started getting an advertisement for the same product while internet surfing on the same browser and this is because of machine learning.

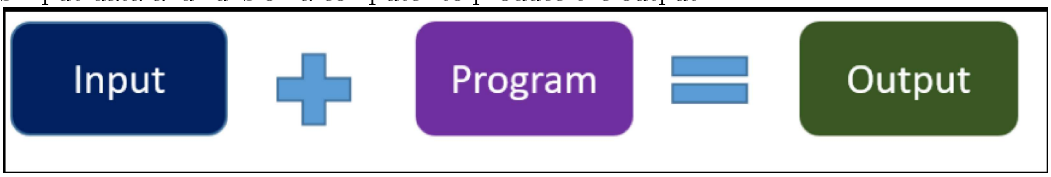
Google understands the user interest using various machine learning algorithms and suggests the product as per customer interest.

As similar, when we use Netflix, we find some recommendations for entertainment series, movies, etc., and this is also done with the help of machine learning.

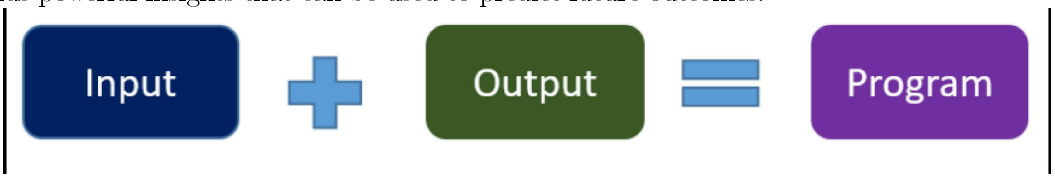
Self-driving cars: One of the most exciting applications of machine learning is self-driving cars. Machine learning plays a significant role in self-driving cars. Tesla, the most popular car manufacturing company is working on self-driving car. It is using unsupervised learning method to train the car models to detect people and objects while driving.

8 Aim : Traditional Programming vs Machine learning

Traditional Programming: It refers to any manually created program that uses input data and runs on a computer to produce the output.



Machine Learning: Machine Learning, also known as augmented analytics, the input data and output are fed to an algorithm to create a program. This yields powerful insights that can be used to predict future outcomes.



9 Aim: Types of Machine Learning

```
Supervised learning
Semi-supervised Learning
Unsupervised Learning
```

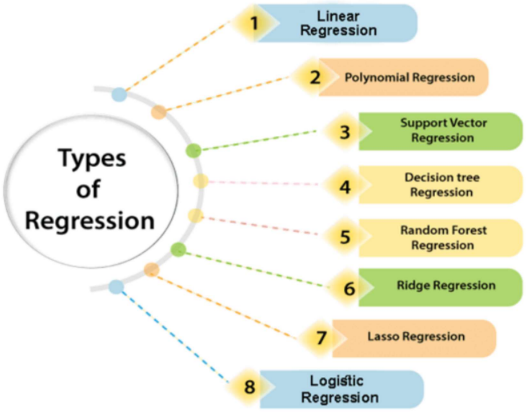
Supervised Learning: It is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output.

Semi-supervised learning: It is a learning problem that involves a small number of labeled examples and a large number of unlabeled examples.Learning problems of this type are challenging as neither supervised nor unsupervised learning algorithms are able to make effective use of the mixtures of labeled and untellable data.

Unsupervised learning: It is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.

10 Aim: Regression

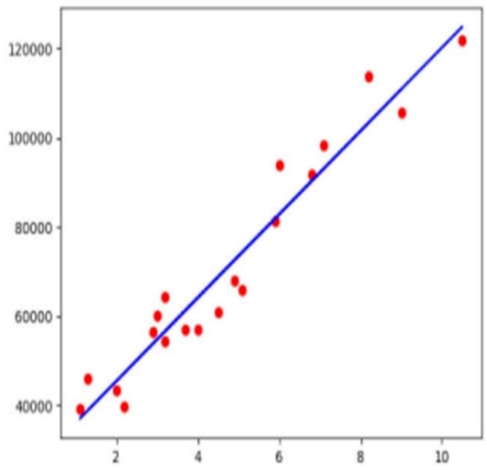
Regression is a method for understanding the relationship between independent variables or features and a dependent variable or outcome.Machine learning regression generally involves plotting a line of best fit through the data points. The distance between each point and the line is minimised to achieve the best fit line.



Simple Linear Regression: Linear regression is the most basic form of regression algorithms in machine learning. The model consists of a single parameter and a dependent variable has a linear relationship. When the number of independent variables increases, it is called the multiple linear regression models.We denote simple linear regression by the following equation given below:

$$y = mx + c + e$$

where m is the slope of the line, c is an intercept, and e represents the error in the model.

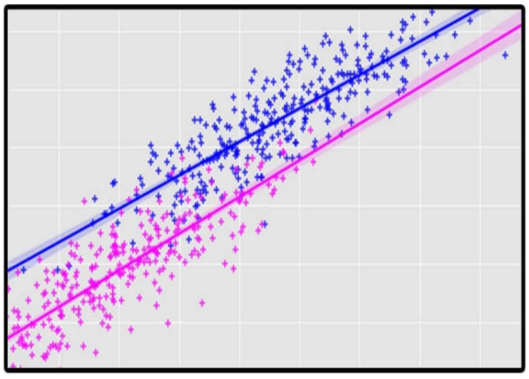


Simple Linear Regression

Multiple Linear Regression: Simple linear regression allows a data scientist or data analyst to make predictions about only one variable by training the model and predicting another variable. In a similar way, a multiple regression model extends to several more than one variable.

Simple linear regression uses the following linear function to predict the value of a target variable y, with independent variable x?:

$$y = b0 + b1x1$$

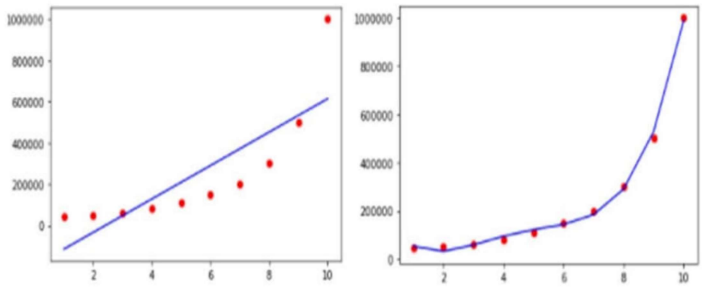


Multiple Linear Regression

Polynomial Regression: In a polynomial regression, the power of the independent variable is more than 1. The equation below represents a polynomial equation:

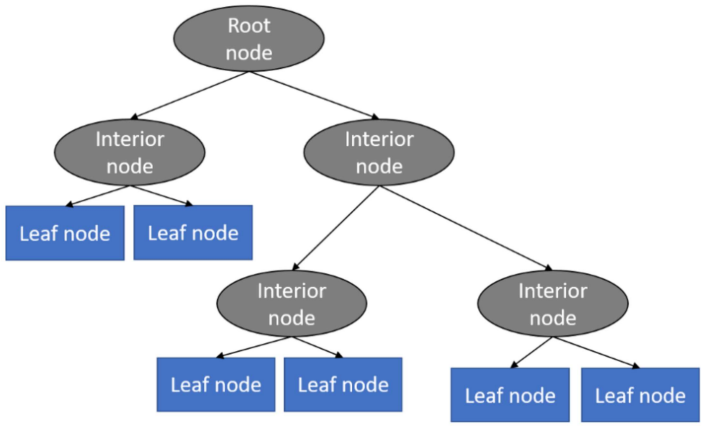
$$y = a + bx2$$

In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.



Polynomial Regression

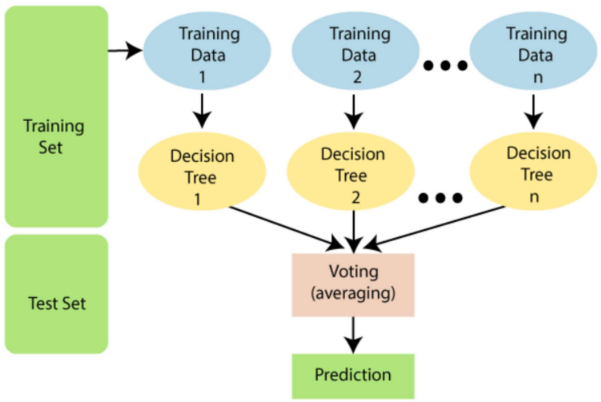
Decision-Tree Regression: It can be used to solve both Regression and Classification tasks with the latter being put more into practical application.It is a tree-structured classifier with three types of nodes. The Root Node is the initial node which represents the entire sample and may get split further into further nodes. The Interior Nodes represent the features of a data set and the branches represent the decision rules. Finally, the Leaf Nodes represent the outcome. This algorithm is very useful for solving decision-related problems.



Decison-Tree Regression

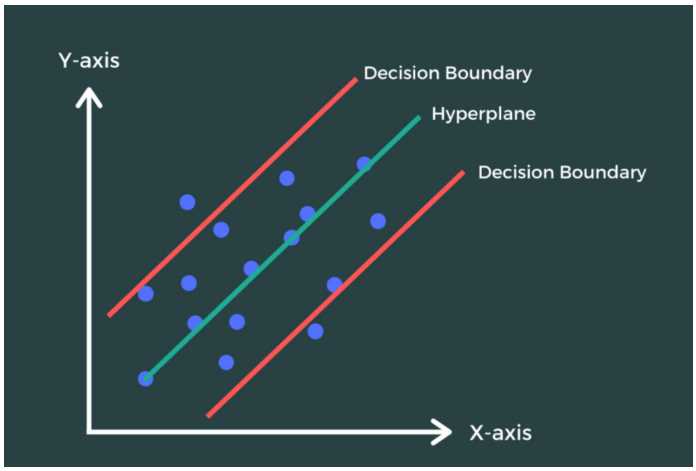
Random Forest Regression: Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the

average to improve the predictive accuracy of that dataset.The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.



Random Forest Regression

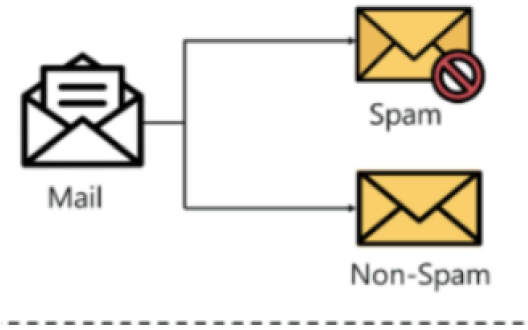
Support Vector Regression: Support Vector Regression is similar to Linear Regression in that the equation of the line is y= wx+b In SVR, this straight line is referred to as hyperplane. The data points on either side of the hyperplane that are closest to the hyperplane are called Support Vectors which is used to plot the boundary line.



Support Vector Regression

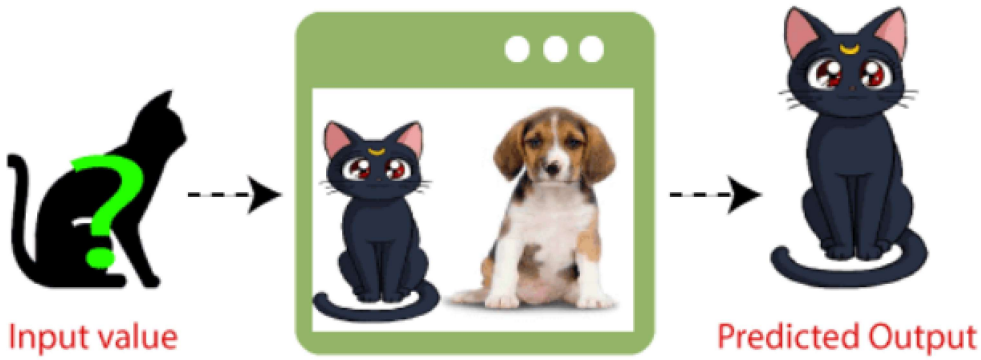
11 Aim: Classification

Classification is a process of categorizing a given set of data into classes, It can be performed on both structured or unstructured data. Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.The process starts with predicting the class of given data points.The classification predictive modeling is the task of approximating the mapping function from input variables to discrete output variables.



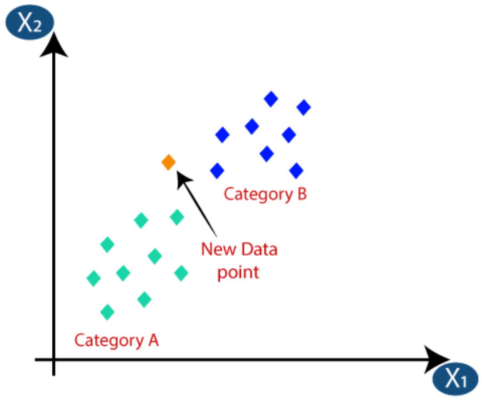
K-Nearest Neighbor(KNN) Algorithm for Machine Learning: K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.



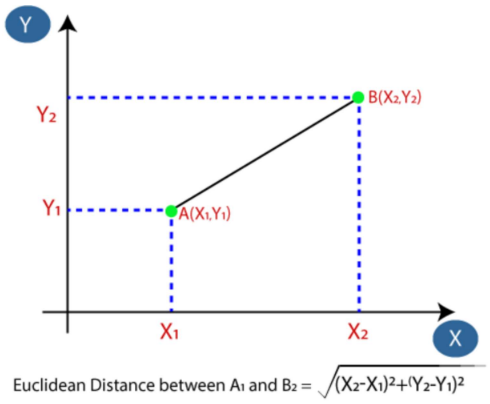
How does K-NN work?

Suppose we have a new data point and we need to put it in the required category. Consider the below image:

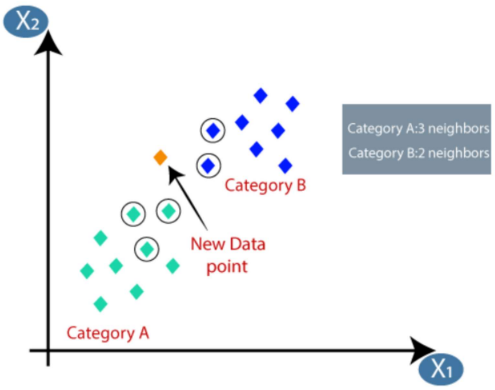


Firstly, we will choose the number of neighbors, so we will choose the k=5.

Next, we will calculate the Euclidean distance between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:



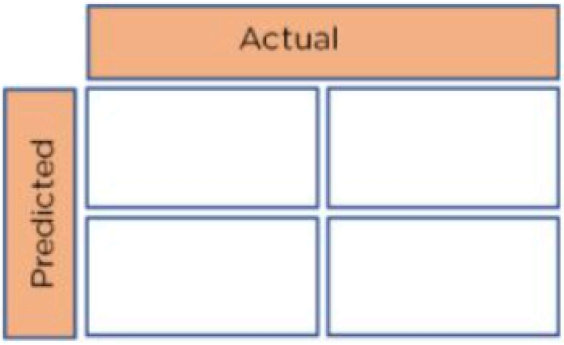
By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the below image:



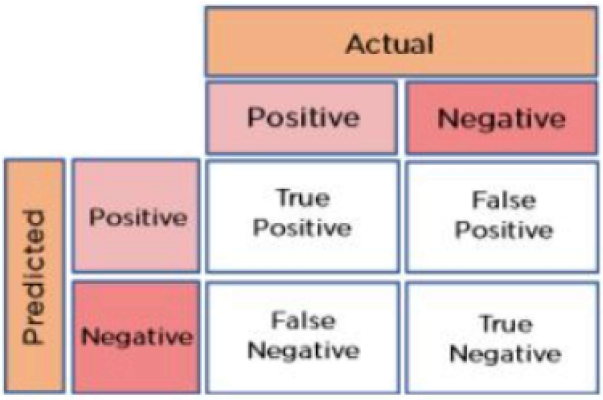
As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.

Confusion Matrix: A confusion matrix presents a table layout of the different outcomes of the prediction and results of a classification problem and helps visualize its outcomes.

It plots a table of all the predicted and actual values of a classifier.



We can obtain four different combinations from the predicted and actual values of a classifier:



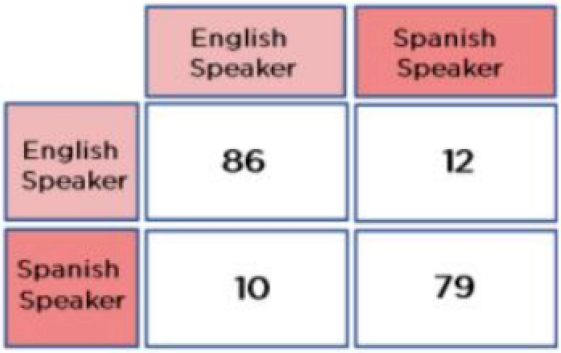
True Positive: The number of times our actual positive values are equal to the predicted positive. You predicted a positive value, and it is correct.

False Positive: The number of times our model wrongly predicts negative values as positives. You predicted a negative value, and it is actually positive.

True Negative: The number of times our actual negative values are equal to predicted negative values. You predicted a negative value, and it is actually negative.

False Negative: The number of times our model wrongly predicts negative values as positives. You predicted a negative value, and it is actually positive.

Confusion Matrix Metrics



Consider a confusion matrix made for a classifier that classifies people based on whether they speak English or Spanish.

From the above diagram, we can see that:

True Positives (TP) = 86

True Negatives (TN) = 79

False Positives (FP) = 12

False Negatives (FN) = 10

Accuracy: The accuracy is used to find the portion of correctly classified values. It tells us how often our classifier is right. It is the sum of all true values divided by total values.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Introduction to K-Means Algorithm: K-means clustering algorithm computes the centroids and iterates until we it finds optimal centroid. It assumes that the number of clusters are already known. It is also called flat clustering algorithm. The number of clusters identified from data by algorithm is represented by 'K' in K-means.

Intuition: By following the procedure for initialization, we pick up centroids that are far away from one another. This increases the chances of initially picking up centroids that lie in different clusters. Also, since centroids are picked up from the data points, each centroid has some data points associated with it at the end.

How did it do that ?

STEP 1: Choose the number K of clusters

STEP 2: Select at random K points, the centroids (not necessarily from your dataset)

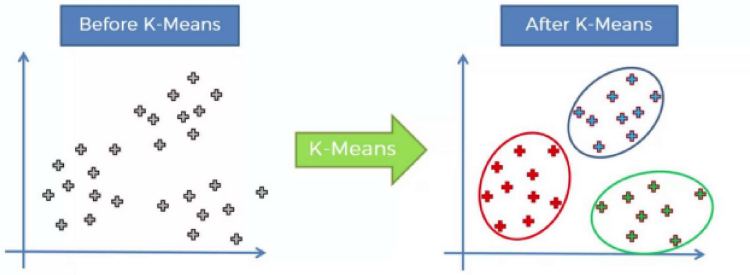
STEP 3: Assign each data point to the closest centroid ➡ That forms K clusters

STEP 4: Compute and place the new centroid of each cluster

STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.

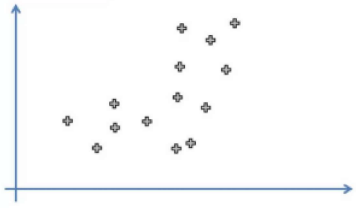
Your Model is Ready

What K-Means does for you



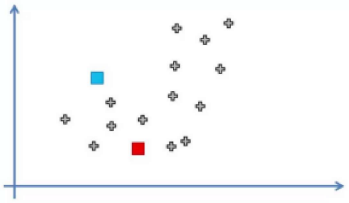
K-Means algorithm

STEP 1: Choose the number K of clusters: K = 2



K-Means algorithm

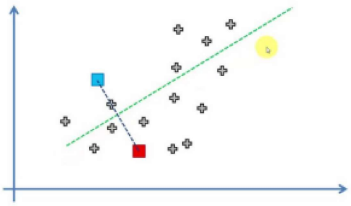
STEP 2: Select at random K points, the centroids (not necessarily from your dataset)



Machine Learning A-Z © SuperDataScience

K-Means algorithm

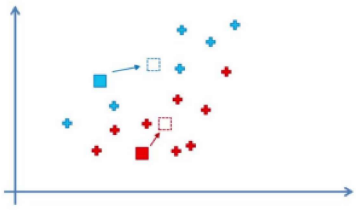
STEP 3: Assign each data point to the closest centroid → That forms K clusters



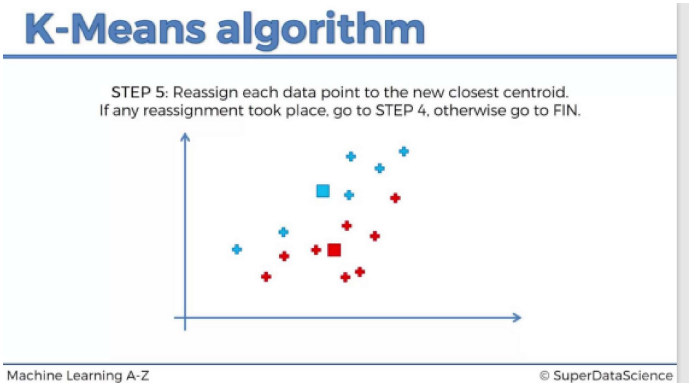
Machine Learning A-Z © SuperDataScience

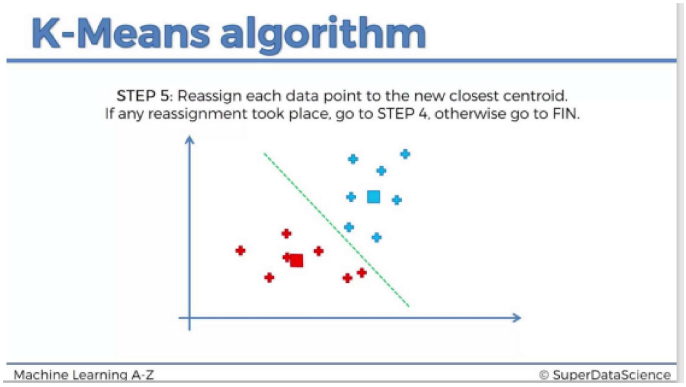
K-Means algorithm

STEP 4: Compute and place the new centroid of each cluster



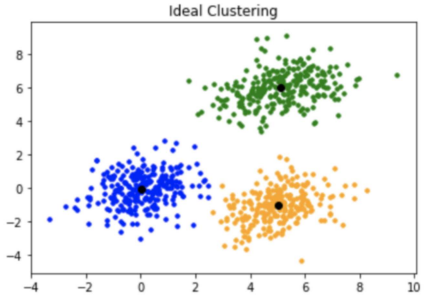
Machine Learning A-Z © SuperDataScience





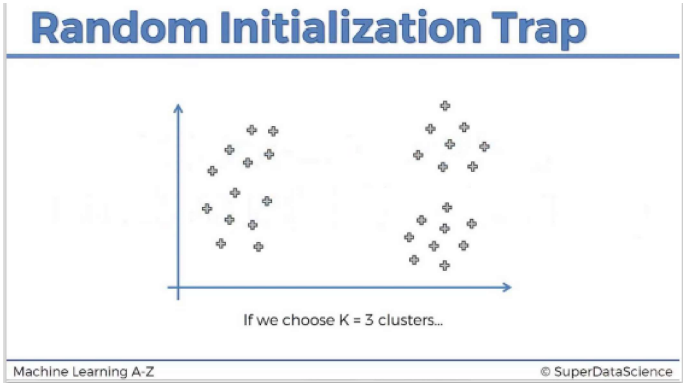
For example, consider the images shown below. A poor initialization of centroids resulted in poor clustering.

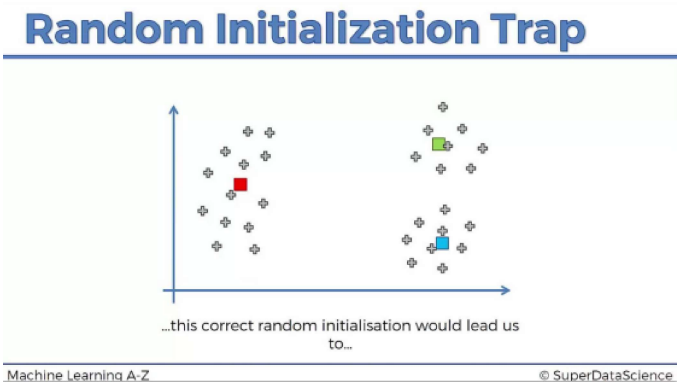
This is how the clustering should have been:

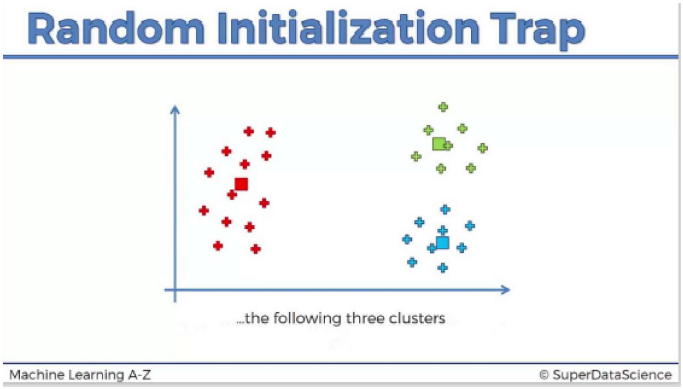


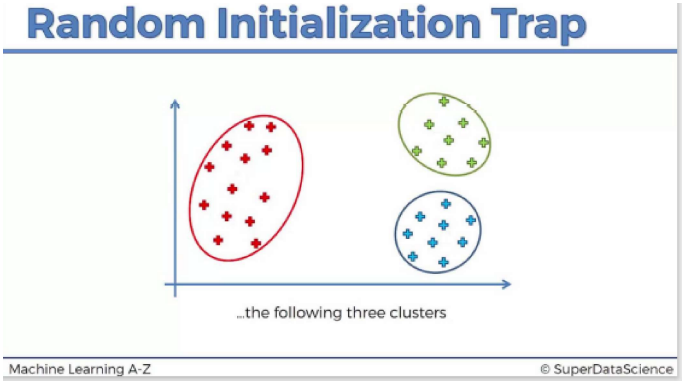
K-mean++: This algorithm ensures a smarter initialization of the centroids and improves the quality of the clustering.

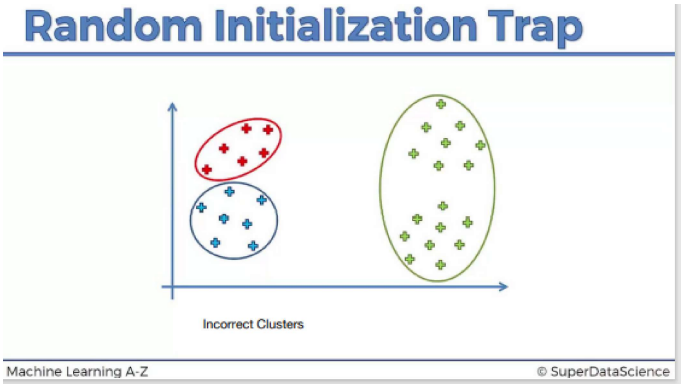
Random Initialization Trap:






Random Initialization Trap
...the following three clusters
Machine Learning A-Z
© SuperDataScience

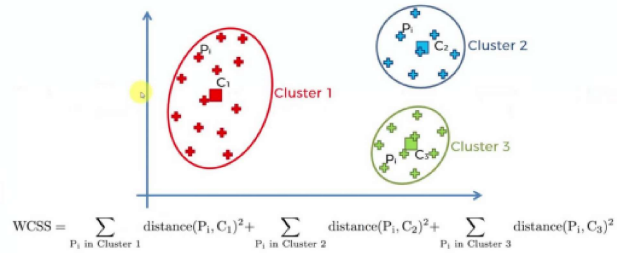

Random Initialization Trap
...the following three clusters
Machine Learning A-Z
© SuperDataScience


Random Initialization Trap
Incorrect Clusters
Machine Learning A-Z
© SuperDataScience

Random Initialization Trap

Machine Learning A-Z © SuperDataScience

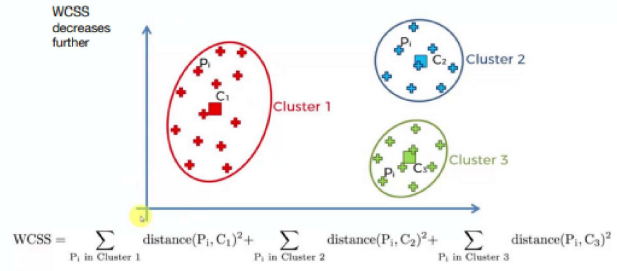
Choosing the right number of clusters



$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

Machine Learning A-Z © SuperDataScience

Choosing the right number of clusters



$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

Machine Learning A-Z © SuperDataScience

Clusters can increase until clusters = number of points
But then WCSS = 0.
Every point would have it's own cluster & centroid.
Distance will be zero.
(over fitting, not optimal)