Machine Learning - Overview

DR DV Ramana
Academic Advisor
And Program Manager

Agenda - Theory

- Introduction to Scikit Learn
- Performing ML using Scikit Learn
- Clustering
  - Hierarchical Clustering
  - K means
- Distance Measure and Data Preparation – Scaling & Weighting
- Evaluation and Profiling of Clusters

Agenda -LAB

1. Explore scikit learn Library
2. Download “mall_customers.csv” dataset from kaggle.
   - (a) Form n no. of clusters according to your observation
   - (b) Find best K value
   - (c) Get wss value for each cluster

Performing ML using Scikit Learn

Labels are the values of the response variables (what's being predicted) that are used by the algorithm along with the feature variables (predictors)

Performing ML using Scikit Learn

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python

Scikit-learn provides a selection of efficient tools for statistical modeling

- Machine learning and
- Including Classification,
- Regression
- Clustering and
- Dimensionality reduction via a consistence interface in Python

Scikit-learn provides many useful functions to create synthetic datasets which are very helpful for practicing machine learning algorithms

https://scikit-learn.org/stable/ https://numpy.org/ https://numpy.org/

https://pandas.pydata.org/docs/getting_started/intro_tutorials/01_table_oriented.html#min-tut-01-tableoriented

https://www.python.org/

Performing ML using Scikit Learn

Following command can be used to install scikit-learn via pip

Using pip

```
pip install -U scikit-learn
```

Using conda

```
conda install scikit-learn
```

Following command can be used to install scikit-learn via pip

Performing ML using Scikit Learn

Specifying the Location of a File

```
#Current Working Directory
import os;
cwd = os.getcwd()
print(cwd)
```

```
import os;
nwd=os.chdir('E://KT/Invited Talk/2018/OU STAT/modules')
nwd= os.getcwd()
print(nwd)
```

Machine Learning
How Unsupervised Learning Works

Unsupervised Learning

No.	Size	Color	Shape	Fruit Name
1	Big	Red	Rounded shape with a depression at the Top	Apple
2	Small	Red	Heart-shaped to nearly globular	Cherry
3	Big	Green	Long curviing cylinder	Banana
4	Small	Green	Round to oval. Bunch shape cylindrical	Grape

Unsupervised Learning

RED COLOR GROUP: apples & cherry fruits.

GREEN COLOR GROUP: bananas & grapes.

so now you will take another physical character such as **size** .

RED COLOR AND BIG SIZE: apple.

RED COLOR AND SMALL SIZE: cherry fruits.

GREEN COLOR AND BIG SIZE: bananas.

GREEN COLOR AND SMALL SIZE: grapes.

job done happy ending.

job done happy ending.

Here you didn't know learn any thing before ,means no train data and no response variable

This type of learning is know unsupervised learning.

clustering comes under unsupervised learning.

Major Techniques of Machine Learning

Clustering : Clustering unsupervised algorithms are mainly used to categorize input data into different clusters or groups based on the pattern of the data

Since there are no previously known groups, the algorithm has to first segment data according to the similarities and dissimilarities and then divide the data into different categories

For instance, clustering can be used in manufacturing to detect any anomalies in production equipment and find the root cause behind the malfunctions

Clustering can be understood with the help of clusters.

So a CLUSTER is a group of objects which are similar in some ways.

It is used to find a pattern in Machine Learning. It is a part of unsupervised learning so deals with the unlabelled dataset.

In layman it is method of grouping objects (similar in some way) called clusters.

A cluster refers to a collection of data points aggregated together because of certain similarities

Major Techniques of Machine Learning

Clustering : Clustering does it by finding some similar patterns in the unlabelled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns.

EXAMPLE

Netflix recommendation systems- With the help of clustering, Netflix data scientists find people who like the series "Lost", "Black Mirror" and "Groundhog Day".

Then it is used to refine its knowledge of the tastes of viewers and thus make better decisions in the creation of new original series.

News Recommendation system- Like news is shown in different categories

Like your searched sports news on google, so clustering method will make cluster /group of sports news

Then it is shown in the search results

Clustering

Clustering mainly deals with finding a structure or pattern in a collection of uncategorized data

Clustering is a technique that groups similar objects such that objects in the same group are identical to each other than the objects in the other groups

Clustering mainly deals with finding a structure or pattern in a collection of uncategorized data

The group of similar objects is called a Cluster

four categories:
Dog
Cat
Shark
Goldfish

four categories:
Dog
Cat
Shark
Goldfish

UNSUPERVISED ALGORITHM - Clustering

Clustering algorithms process data to split data points into clusters.

The idea is that data points with similar features should be assigned to the same cluster and that the points in different clusters should have different features.

Some of the different clustering types include:

- K-means clustering
- Hierarchical clustering

Unsupervised learning algorithms are: k-means for clustering problems.

Clustering automatically split the dataset into groups base on their similarities.

Machine Learning

UNSUPERVISED ALGORITHM - Clustering

Clustering algorithms process data to split data points into clusters.

The idea is that data points with similar features should be assigned to the same cluster and that the points in different clusters should have different features.

sample

Cluster/group

Some of the different clustering types include:

K-means clustering

Hierarchical clustering

Unsupervised learning algorithms are: k-means for clustering problems.

Clustering automatically split the dataset into groups base on their similarities.

Major Techniques of Machine Learning

Unsupervised Learning - Clustering

Finding the structure of data; summarization
Clustering groups of similar cases

E.g. Can find similar patients, or can be used for customer segmentation in the banking field.

•It is considered to be one of the most popular unsupervised machine learning techniques used for grouping data points or objects that are somehow similar.

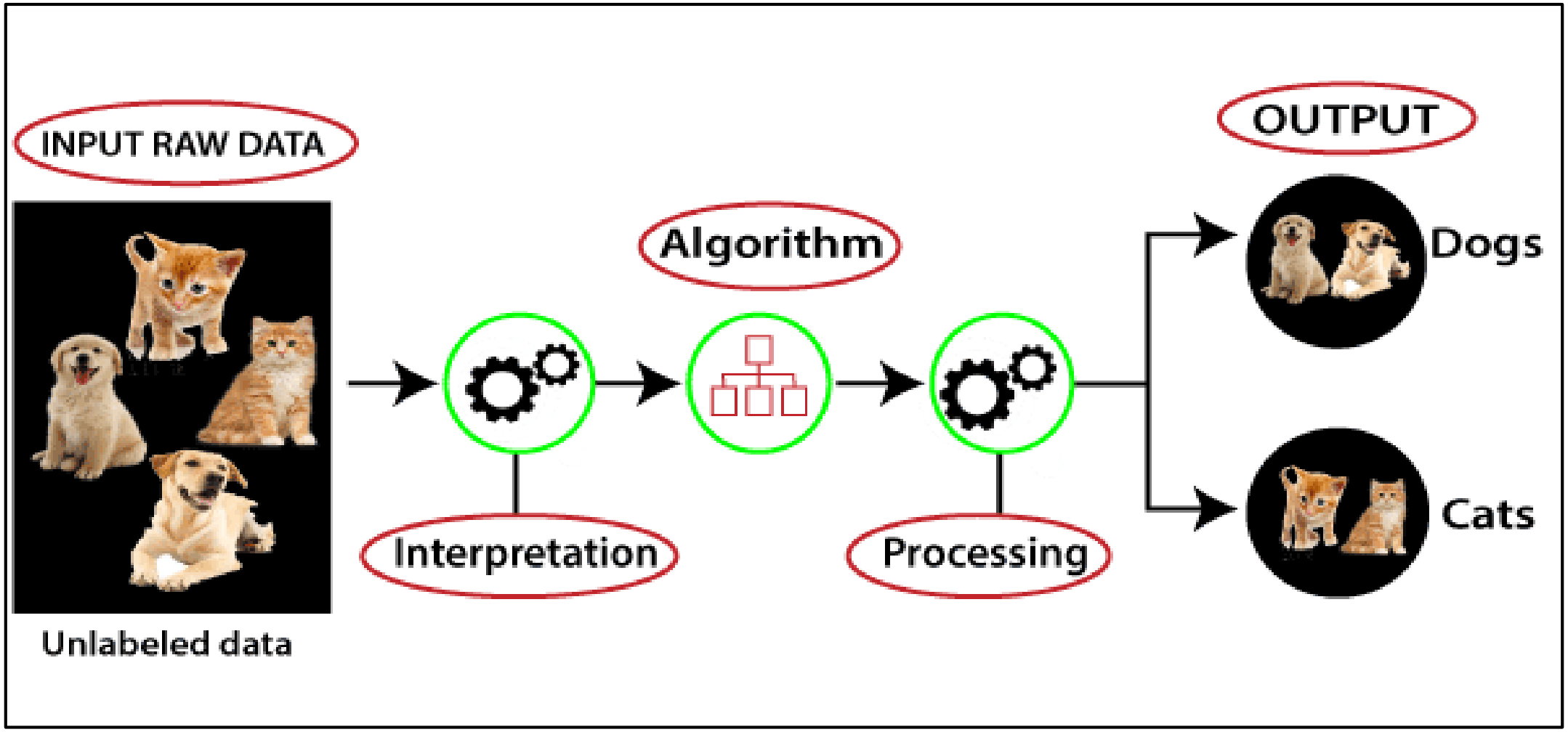
It is a grouping of data points or objects that are somehow similar by

- Discovery Structure
- Summarization
- Anomaly detection

Cluster analysis has many applications in different domains, whether it be a bank's desire to segment its customers based on certain characteristics, or helping an individual to organize and group his/her favorite types of books!

Major Techniques of Machine Learning

Unsupervised Learning - Clustering



Supervised vs. Unsupervised Machine Learning

Parameters	Supervised machine learning technique	Unsupervised machine learning technique
Input Data	Algorithms are trained using labeled data.	Algorithms are used against data which is not labelled
Computational Complexity	Supervised learning is a simpler method.	Unsupervised learning is computationally complex
Accuracy	Highly accurate and trustworthy method.	Less accurate and trustworthy method.
Process	input and output variables will be given.	learning model, only input data will be given
Algorithms Used	Support vector machine, Neural network, Linear and logistics regression, random forest, and Classification trees.	Unsupervised algorithms can be divided into different categories: like Cluster algorithms, K-means, Hierarchical clustering, etc.
Computational Complexity	simpler method.	computationally complex
Use of Data	uses training data to learn a link between the input and the outputs.	does not use output data.
Real Time Learning	Learning method takes place offline.	Learning method takes place in real time.

Supervised vs. Unsupervised Machine Learning

Parameters	Supervised machine learning technique	Unsupervised machine learning technique
Number of Classes	Number of classes is known.	Number of classes is not known
Computational Complexity	Supervised learning is a simpler method.	Unsupervised learning is computationally complex
Main Drawback	Classifying big data can be a real challenge in Supervised Learning.	You cannot get precise information regarding data sorting, and the output as data used in unsupervised learning is labeled and not known.
Data available	Both input and output data is available	Only unlabeled input data is available
Goal	Understand the relationship between input and output data, and predict future data accordingly.	Identify the underlying structure and hidden pattern present in the input data.
Feedback	It takes direct feedback from the programmer to check if the predictions are correct or not.	It does not take any feedback.
Complexity and accuracy	While it is comparatively less complex, it provides a higher accuracy rate	More complex than supervised learning and the accuracy levels are also relatively less

Supervised vs. Unsupervised Machine Learning

Parameters	Supervised machine learning technique	Unsupervised machine learning technique
Use cases	Used for speech recognition, image recognition, financial analysis, forecasting, and training neural networks	Mainly used to pre-process data or to pre-train supervised learning algorithms.

Applications of unsupervised machine learning

Clustering automatically split the dataset into groups base on their similarities

Anomaly detection can discover unusual data points in your dataset. It is useful for finding fraudulent transactions

Association mining identifies sets of items which often occur together in your dataset

Latent variable models are widely used for data preprocessing. Like reducing the number of features in a dataset or decomposing the dataset into multiple components

Disadvantages of Unsupervised Learning

Cannot get precise information regarding data sorting, and the output as data used in unsupervised learning is labeled and not known

Less accuracy of the results is because the input data is not known and not labeled by people in advance. This means that the machine requires to do this itself.

The spectral classes do not always correspond to informational classes.

The user needs to spend time interpreting and label the classes which follow that classification

Spectral properties of classes can also change over time so you can't have the same class information while moving from one image to another.

Unsupervised Learning

Unsupervised learning is a machine learning technique, where you do not need to supervise the model.

Unsupervised machine learning helps you to finds all kind of unknown patterns in data.

Clustering and Association are two types of Unsupervised learning.

1) Exclusive

2) Agglomerative

3) Overlapping

4) Probabilistic

Unsupervised Learning

Important clustering types are:

1)Hierarchical clustering

2) K-means clustering

Supervised Learning vs Unsupervised Learning

Classification produces discrete values and dataset to strict categories, while regression gives you continuous results that allow you to better distinguish differences between individual points.

You would use classification over regression if you wanted your results to reflect the belongingness of data points in your dataset to certain explicit categories

Example:

If you wanted to know whether a name was male or female rather than just how correlated they were with male and female names.

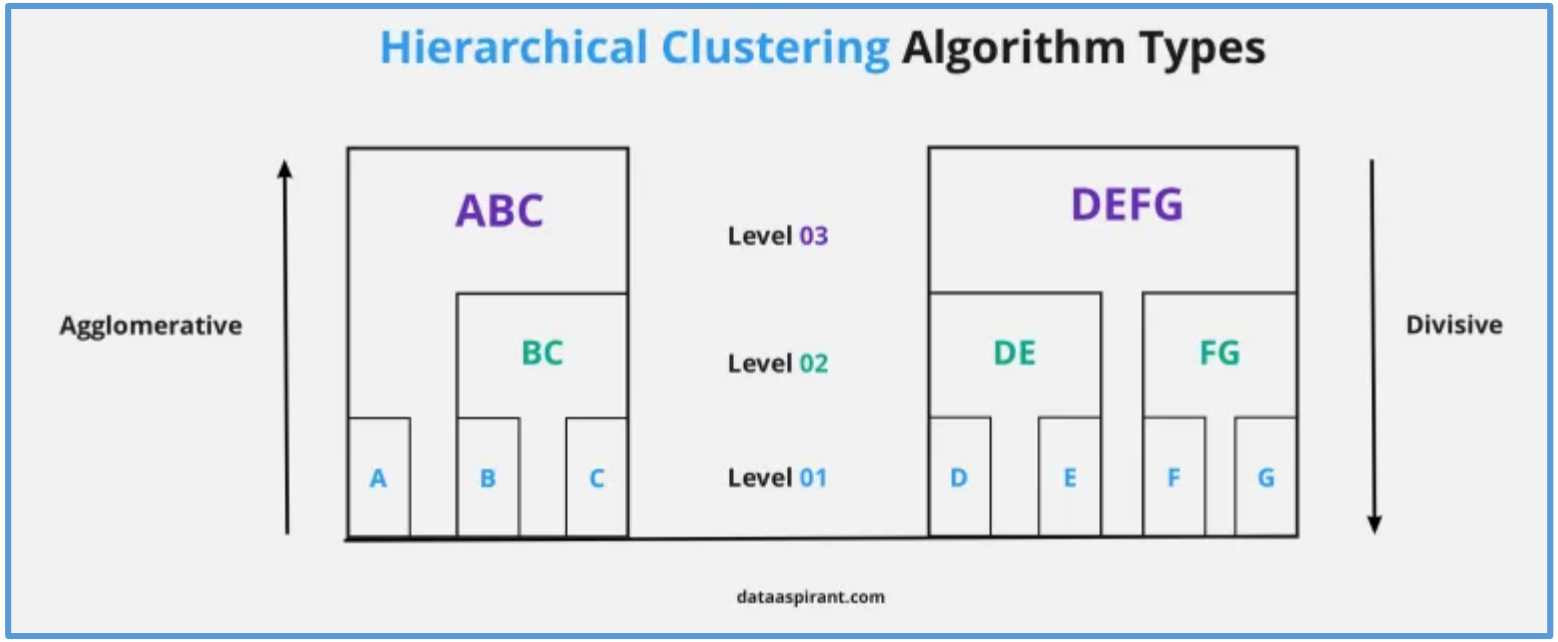
Hierarchical Clustering

Hierarchical clustering is an unsupervised learning method for clustering data points

Unsupervised learning means that a model does not have to be trained, and we do not need a "target" variable

Hierarchical clustering is an unsupervised learning method for clustering data points

Hierarchical clustering also known as Hierarchical Clustering Analysis (HCA)



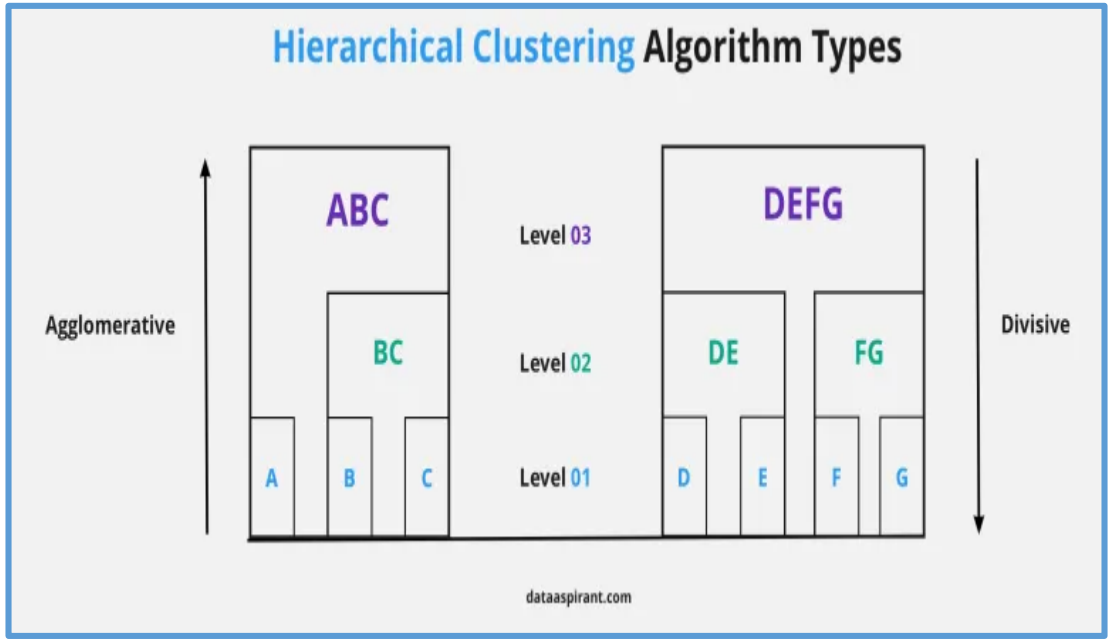
Hierarchical Clustering

Hierarchical clustering is an unsupervised learning method for clustering data points

Unsupervised learning means that a model does not have to be trained, and we do not need a "target" variable

Hierarchical clustering is an unsupervised learning method for clustering data points

Hierarchical clustering also known as Hierarchical Clustering Analysis (HCA)



Agglomerative:
Hierarchy created from bottom to top.

Divisive:
Hierarchy created from top to bottom.

K means Clustering

K-means clustering is unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes

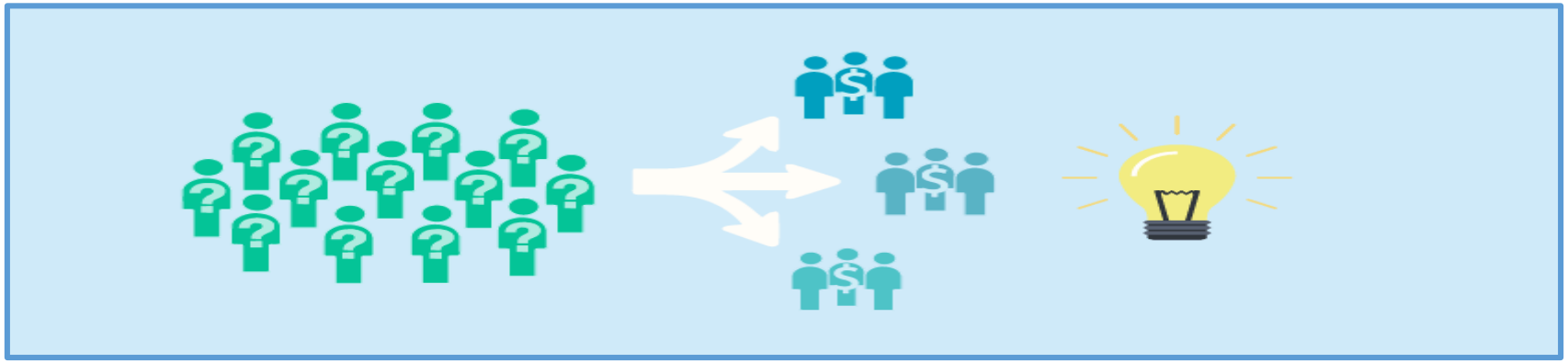
K-means is a centroid-based clustering algorithm, where we calculate the distance between each data point and a centroid to assign it to a cluster

The goal is to identify the K number of groups in the dataset

Use K means clustering to generate groups comprised of observations with similar characteristics

Example

If you have customer data, you might want to create sets of similar customers and then target each group with different types of marketing.



K means Clustering

How the K-means algorithm works

The k-means clustering algorithm mainly performs two tasks

Determines the best value for K center points or centroids by an iterative process

Assigns each data point to its closest k-center

Those data points which are near to the particular k-center, create a cluster

K means Clustering

How the K-means algorithm works

The k-means clustering algorithm mainly performs two tasks

Determines the best value for K center points or centroids by an iterative process

Assigns each data point to its closest k-center

Those data points which are near to the particular k-center, create a cluster

K-means algorithm is not capable of determining the number of clusters.

We need to define it when creating the KMeans object which may be a challenging task

Distance Measure and Data Preparation – Scaling & Weighting

Distance measures play an important role in machine learning.

A distance measure is an objective score that summarizes the relative difference between two objects in a problem domain

Most commonly, the two objects are rows of data that describe a subject (such as a person, car, or house), or an event (such as a purchase, a claim, or a diagnosis).

Distance Measure and Data Preparation – Scaling & Weighting

Hamming Distance

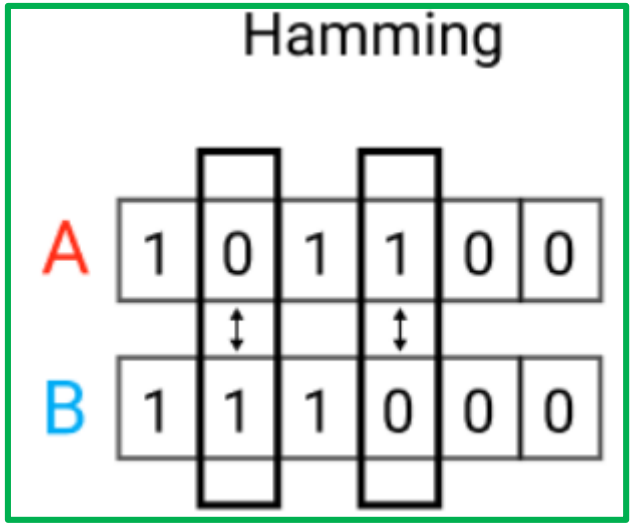
Hamming distance calculates the distance between two binary vectors, also referred to as binary strings or bitstrings for short

Most likely going to encounter bitstrings when you one-hot encode categorical columns of data

Hamming distance is the number of values that are different between two vectors

Hamming distance is typically used to compare two binary strings of equal length

Hamming distance can also be used for strings to compare how similar they are to each other by calculating the number of characters that are different from each other



Distance Measure and Data Preparation – Scaling & Weighting

Euclidean Distance

Euclidean distance calculates the distance between two real-valued vectors

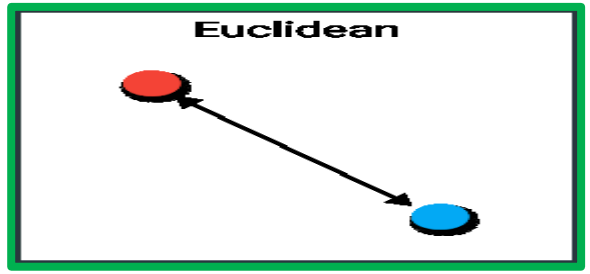
Most likely to use Euclidean distance when calculating the distance between two rows of data that have numerical values, such a floating point or integer values

Euclidean Distance is a distance measure that best can be explained as the length of a segment connecting two points

Euclidean Distance is a distance measure that best can be explained as the length of a segment connecting two points

Formula

$$D(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2)}$$



Distance Measure and Data Preparation – Scaling & Weighting

Manhattan Distance (Taxicab or City Block Distance)

Manhattan distance, also called the Taxicab distance or the City Block distance, calculates the distance between two real-valued vectors

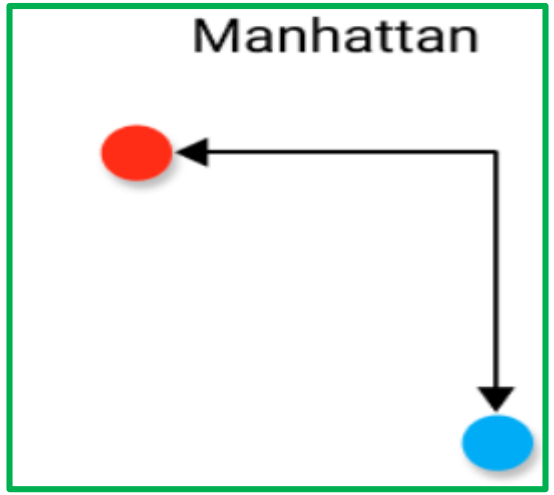
Manhattan distance is perhaps more useful to vectors that describe objects on a uniform grid, like a chessboard or city blocks

The taxicab name for the measure refers to the intuition for what the measure calculates: the shortest path that a taxicab would take between city blocks (coordinates on the grid)

Imagine vectors that describe objects on a uniform grid such as a chessboard

Manhattan distance then refers to the distance between two vectors if they could only move right angles

There is no diagonal movement involved in calculating the distance



$$D(x, y) = \sum_{i=1}^{k} |x_i - y_i|$$

Evaluation and Profiling of Clusters

Three important factors by which clustering can be evaluated are

- Clustering tendency
- Number of clusters, k
- Clustering quality

Within-Cluster Sum of Squares (WSS)

Within-Cluster Sum of Squares (WSS)

Objective:

Find the "Elbow" of the WSS curve in order to determine the smallest number of clusters that captures the most amount of signal in your data

Within-Cluster Sum of Squares (WSS) is a measure of how far away each centroid is from their respective class instances

The larger the WSS, the more dispersed the cluster values are from the centroid

LAB

LAB

1. Explore scikit learn Library

2. Download “mall_customers.csv” dataset from kaggle.

   (a) Form n no. of clusters according to your observation

   (b) Find best K value

   (c) Get wss value for each cluster

Thank you for Listening

Any Questions

DR DV Ramana
Academic Advisor
And Program Manager

Mail Address: pythonpmg@gmail.com

To contact: +91 9959423084