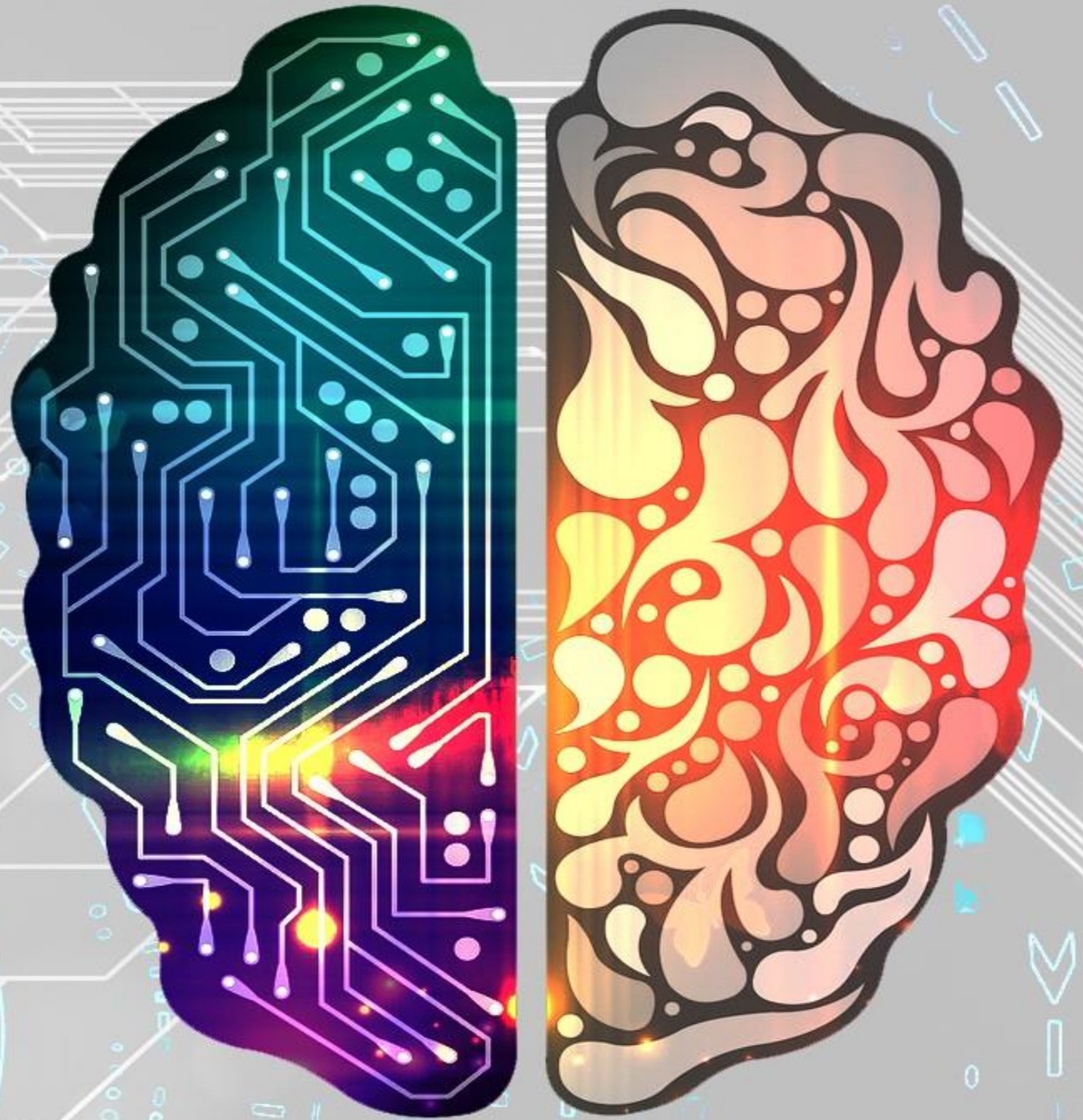


MACHINE LEARNING



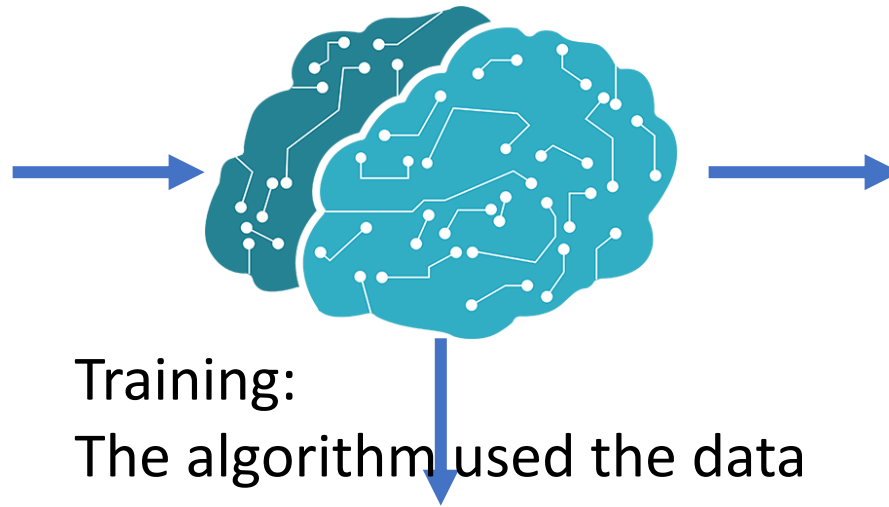
Machine Learning (ML):

- ML is a subfield of AI that uses previous data to teach computers to do what humans can do, without being explicitly programmed.
- Unlike other computer programs, ML algorithms do not follow predefined rules to classify and make predictions.
- ML algorithms solve problems on their own through learning from large sets of examples in the provided data.

Detecting heart problem

- The problem: Determining if the patient suffers from coronary artery disease. Can we use ML to diagnose the problem ?

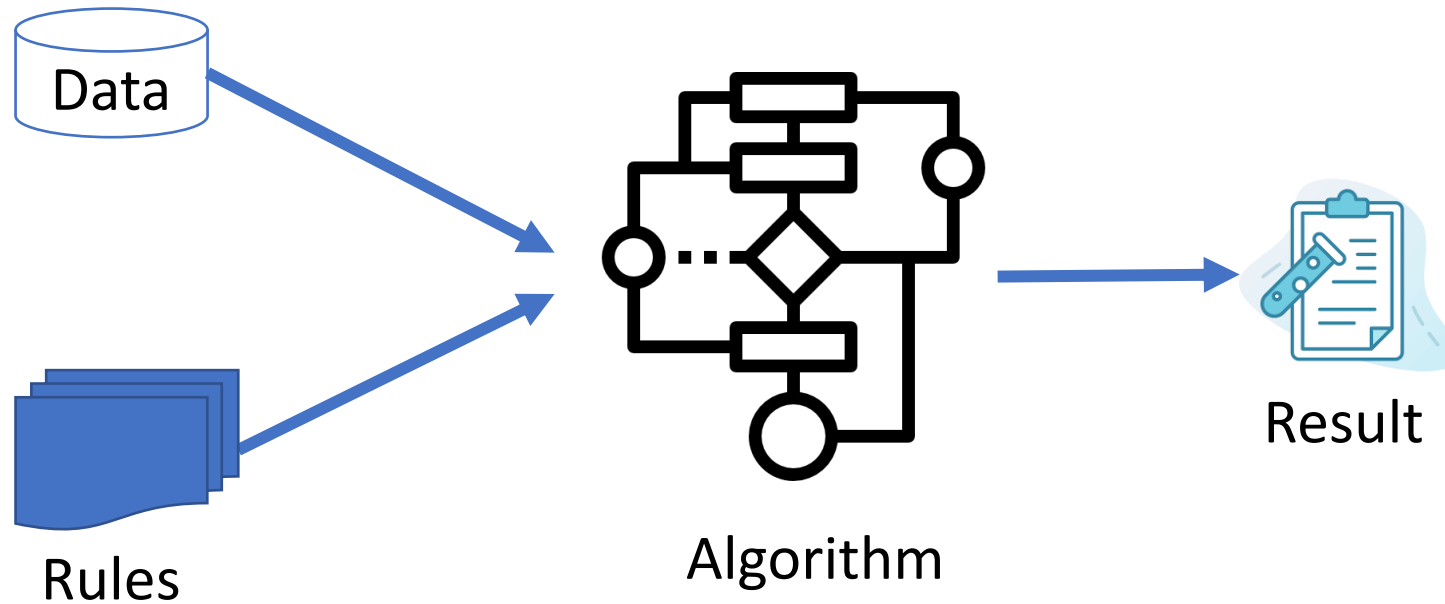
Age ,Weight, Diabetic,
Hypertensive,
Hypercholesterolemia,
and the Diagnosis



Predict the diagnosis:
Negative, or positive

Sex	Age	Weight	Hypertensive	Hypercholesterolemia	Smokeing	Diabetic	Family History	Diagnosis
F	77.00	73.00	1.00	1.00	0.00	0.00	0.00	Neg
F	55.00	70.00	1.00	1.00	0.00	0.00	1.00	Neg
F	67.00	63.00	1.00	1.00	0.00	0.00	0.00	Pos
F	51.00	89.00	1.00	1.00	0.00	1.00	1.00	Neg
F	72.00	56.00	1.00	1.00	0.00	1.00	0.00	Pos
M	67.00	76.00	0.00	1.00	0.00	0.00	0.00	Neg

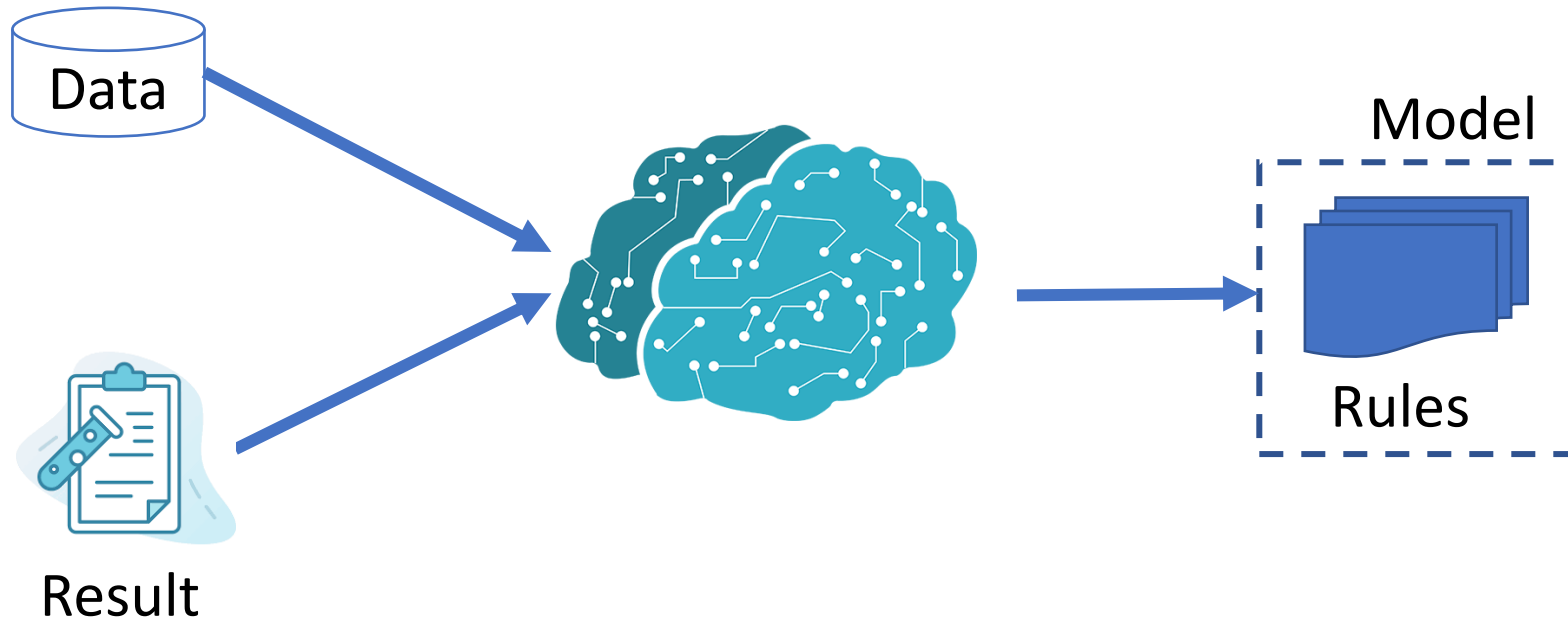
The difference between ML and traditional programming



If we use this approach to detect the heart problem:

- Use the data to create an algorithm (IF THEN ELSE).
- Based on the rules the algorithm will detect if the patient is healthy or not.
- The algorithm will not change.

Machine learning approach



ML algorithm:

- Uses the input data to tune the parameter through matching the input results with the predicted results.
- The model decide the logic between the input variables .
- The model keeps changing the rules and the parameters until the difference between the actual results and the predicted results is minimised.
- Unlike a traditional algorithm, This model can be continuously trained and be used in the future to predict values.

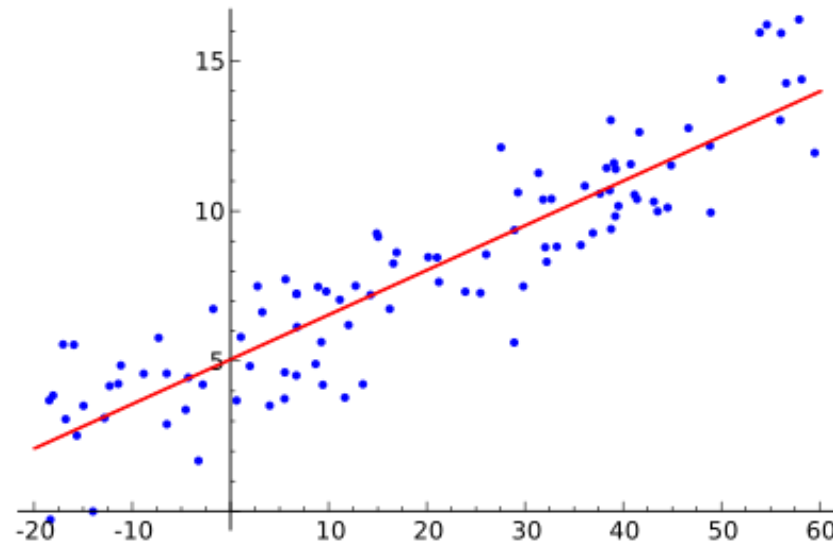
ML algorithms categories:

- Supervised learning:
 - The ML algorithm is trained on human-labelled data. The more samples you provide a supervised learning algorithm, the more precise it becomes in classifying new data.
- Unsupervised learning:
 - Is another type of machine language, relies on giving the algorithm unlabelled data and letting it find patterns by itself. You provide the input but not labels, and let the machine infer qualities that algorithm ingests unlabelled data, draws inferences, and finds patterns. This type of learning can be useful for clustering data.

- Reinforcement Learning:
 - Relies on providing a machine learning algorithm with a set of rules and constraints, and letting it learn how to achieve its goals. You define the state, the desired goal, allowed actions, and constraints.
 - The algorithm figures out how to achieve the goal by trying different combinations of allowed actions, and is rewarded or punished depending on whether the decision was a good one.
 - The algorithm tries its best to maximize its rewards within the constraints provided.
 - You could use Reinforcement Learning to teach a machine to play chess or navigate an obstacle course.

Supervised learning:

- It can be broken out to three types:
 1. Regression
 - Regression models estimates continuous values.
 - They are built by looking at the relationships between features x and the result y where y is a continuous variable.



- Classification
 - focuses on discrete values it identifies. We can assign discrete class labels y based on many input features x .

Sex	Age	Weight	Hypertensive	Hypercholesterolemia	Smokeing	Diabetic	Family History	Diagnosis
F	77.00	73.00	1.00	1.00	0.00	0.00	0.00	Neg
F	55.00	70.00	1.00	1.00	0.00	0.00	1.00	Neg
F	67.00	63.00	1.00	1.00	0.00	0.00	0.00	Pos
F	51.00	89.00	1.00	1.00	0.00	1.00	1.00	Neg
F	72.00	56.00	1.00	1.00	0.00	1.00	0.00	Pos
M	67.00	76.00	0.00	1.00	0.00	0.00	0.00	Neg

Data types and structure

- Categorical data:
 - Label for some value, such as the species of a plant or whether a photo is that of a cat or a dog. Food type: Indian, Italian, and Chinese
 - Categorical variables cannot be ordered and the only useful comparison to make between two categorical variables is whether they are equal or not
- Continuous data:
 - Can take any particular value, it includes things like weight, height, speed, ..etc
 - We can order the value, addition and subtraction make sense.
- Ordinal data:
 - Things are ranked according to magnitude, e.g. Bad→Good, First→Last, Rich→Poor, by year, by age.

Causality relation between variables

- Independent variables: They are also called features. They Can be manipulated to see how changing them affects another dependent (related) variable.
- Dependent variables: They are called labels or classes. They are measured or labelled to provide the ground truth for supervising learning algorithm. Usually the change of independent variable affects them.

Features:

- Features are distinctive properties that describe the patterns in data. Features represents the independent variables in the data. They help in determining the dependent variable or classes.
- Each column is a feature and each row is a data point.

Sex	Age	Weight	Hypertensive	Hypercholesterolemia	Smokeing	Diabetic	Family History	Diagnosis
F	77.00	73.00	1.00	1.00	0.00	0.00	0.00	Neg
F	55.00	70.00	1.00	1.00	0.00	0.00	1.00	Neg
F	67.00	63.00	1.00	1.00	0.00	0.00	0.00	Pos
F	51.00	89.00	1.00	1.00	0.00	1.00	1.00	Neg
F	72.00	56.00	1.00	1.00	0.00	1.00	0.00	Pos
M	67.00	76.00	0.00	1.00	0.00	0.00	0.00	Neg

Training classification algorithm

ML algorithm uses training data to learn how given input features relate to that ground truth class label. But, what exactly do we mean by training?

- Training refers to using labelled data by the learning algorithm to determine the parameters or calculate the weight of connection within the trained model.
- For example:
 - If we are training a model using the coronary artery disease data to predict whether the patient is healthy (negative) or nonhealthy (positive).
 - We need to show the algorithm data point from the actual data labelled positive, then show the algorithm again some data labelled negative.
 - We need to repeated this process until the whole data points in the trained data are showed to the algorithm.
 - The algorithm modifies internal values until it has learned to tell from data that indicates diagnosis is positive or negative.

Validation and testing data

- Data is divided randomly into three parts:
 - Training subset: we know already this one !
 - Validation subset : The validation subset is used to validate the results to monitor the performance during the training to avoid overfitting.
 - Testing set: The Testing data is the data that the model has never seen before and it is used to evaluate the performance of the model.
 - A good rule of thumb is to reserve about 20–25% of the dataset as testing data, and to keep back about 10% of the training data as validation data.

K-fold Cross Validation:

- Cross-validation (CV) is a technique used to assess the performance of ML model.
- For example if $K=5$, it means we are using 5-fold cross validation, the procedure is as following:
 1. Shuffle the dataset randomly.
 2. Split the dataset into k groups
 3. For each unique group:
 - Take the group as a hold out or test data set
 - Take the remaining groups as a training data set
 - Fit a model on the training set and evaluate it on the test set
 - Retain the evaluation score and discard the model
 4. Summarize the skill of the model using the sample of model evaluation scores

Performance Metrics in Machine Learning

- Classification metrics:
 - Confusion Matrix: Confusion Matrix is a tabular visualization of the ground-truth labels versus model predictions.

		Predicted	
		Positive	Negative
Ground Truth	Positive	TP	FP
	Negative	FN	TN

- Accuracy: $AC = \frac{TP+TN}{TP+FP+TN+FN}$
- Sensitivity=Recall=true positive rate $= \frac{TP}{TP+FN}$
- Specificity= true negative rate $= \frac{TN}{TN+FP}$
- Precision $= \frac{TP}{TP+FP}$
- F-score $= 2 \times \frac{Recall \times Precision}{Precision+Recall}$

Unsupervised Learning

- **K-Means Clustering algorithm**



Agenda for Section

What is Clustering ?

What is K-Means Clustering.

How dose K-Means Algorithm work?

Distance Measures

Advantages and Limitations

What is clustering?

Clustering is the process of dividing the data points into groups that consists similar points.

- Points in the same group should be as similar as possible.
- Points in different groups should be as dissimilar as possible.

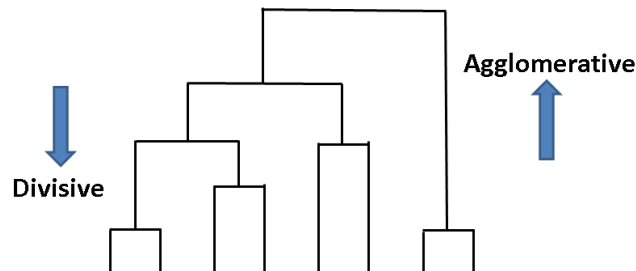
Where is it used?



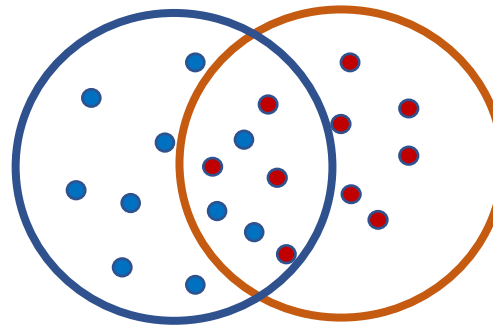
Types of Clustering

- Hard Clustering.
- Data points belong exclusively to one cluster.
- Exclusive.
- Overlapping Clustering.
- Hierarchical Clustering

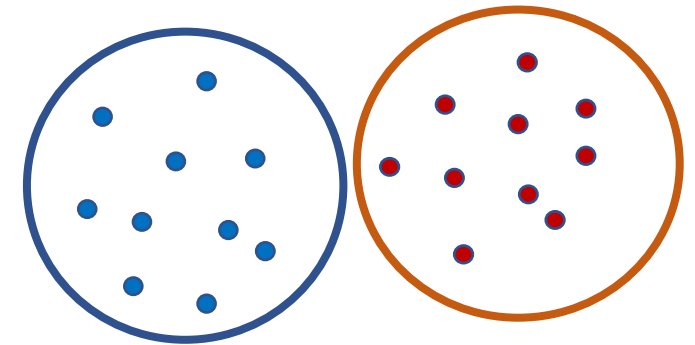
Hierarchical Clustering



Overlapping Clustering



Hard clustering

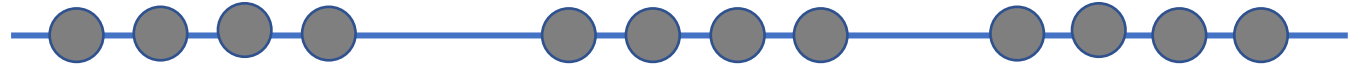


What is K-Means Clustering?

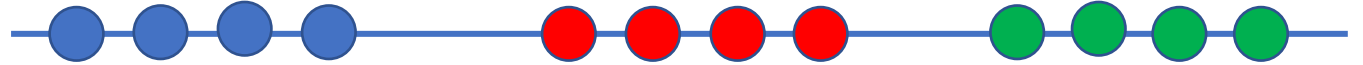
- It is an algorithms that group similar the data points into hard clusters.
- 'k' indicate the number of clusters.

K-Means Algorithm?

One dimensional dataset consists of 12 points

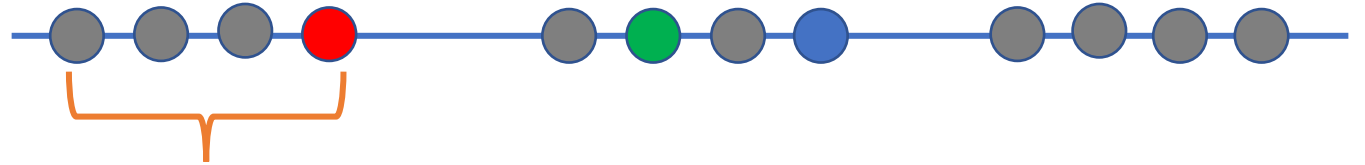


Expected three clusters



K-Means Algorithm?

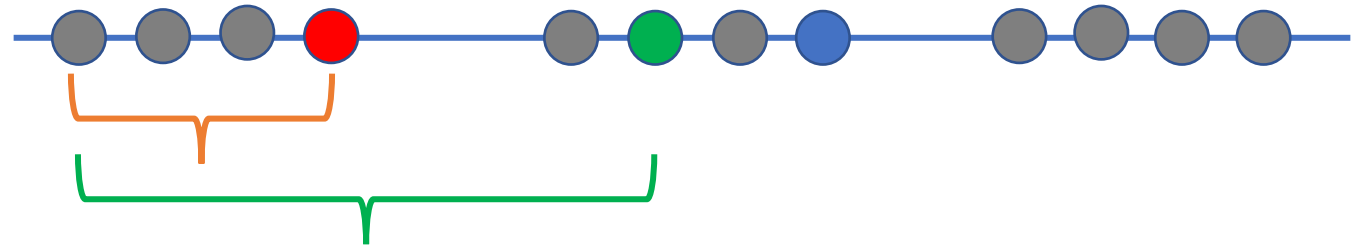
- Step 1: $K=3$
- Step 2: Select randomly 3 distinct data points, which will work as initial centroid of each cluster.
- Step 3: Measure the distance between each point and 3 selected clusters centroid.



Distance between
point 1 and the red
cluster

K-Means Algorithm?

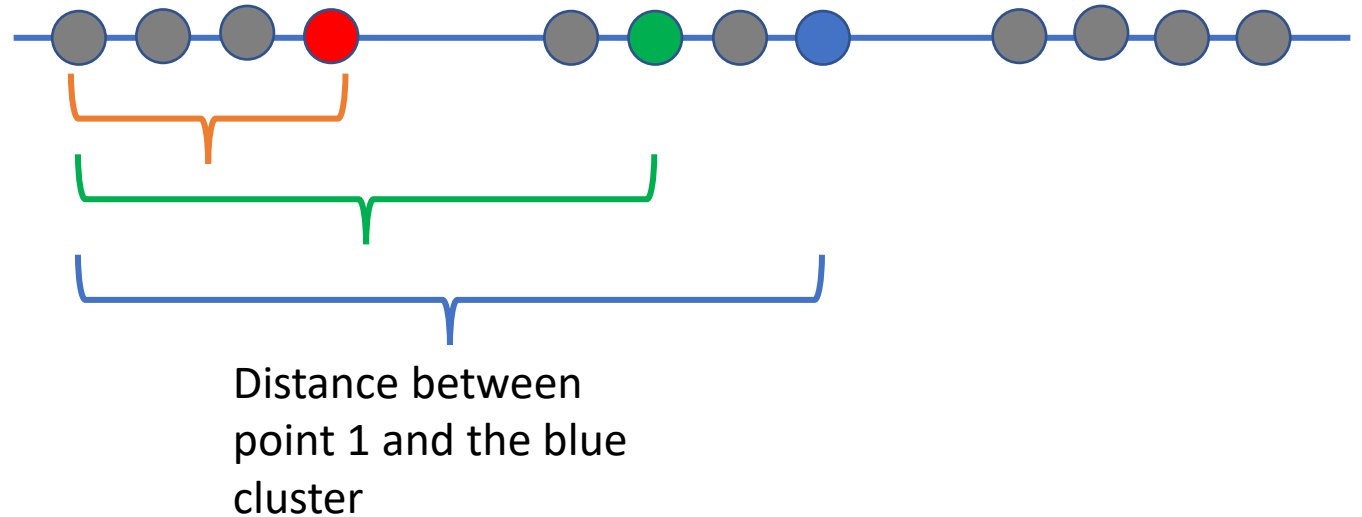
- Step 1: $K=3$
- Step 2: Select randomly 3 distinct data points, which will work as initial centroid of each cluster.
- Step 3: Measure the distance between each point and 3 selected clusters centroid.



Distance between
point 1 and the green
cluster

K-Means Algorithm?

- Step 1: $K=3$
- Step 2: Select randomly 3 distinct data points, which will work as initial centroid of each cluster.
- Step 3: Measure the distance between each point and 3 selected clusters centroid.



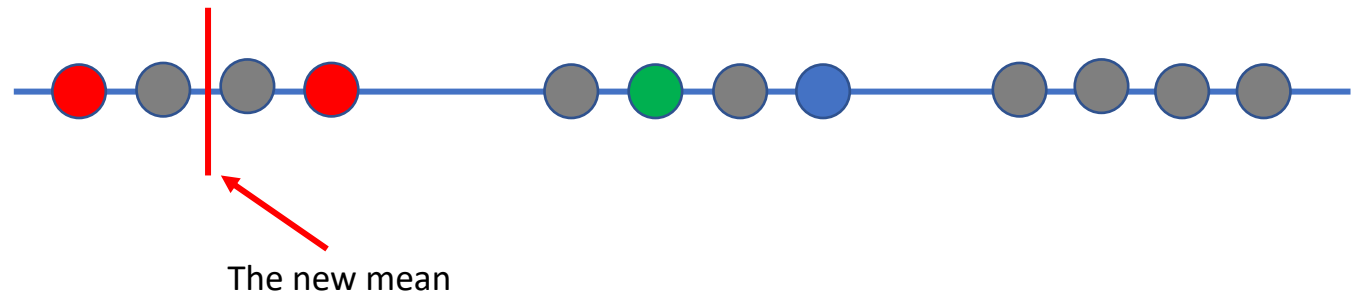
K-Means Algorithm?

- Step 1: $K=3$
- Step 2: Select randomly 3 distinct data points, which will work as initial centroid of each cluster.
- Step 3: Measure the distance between each point and 3 selected clusters centroid.
- Step 4: Assign the point to the nearest cluster (which the red cluster for first point).



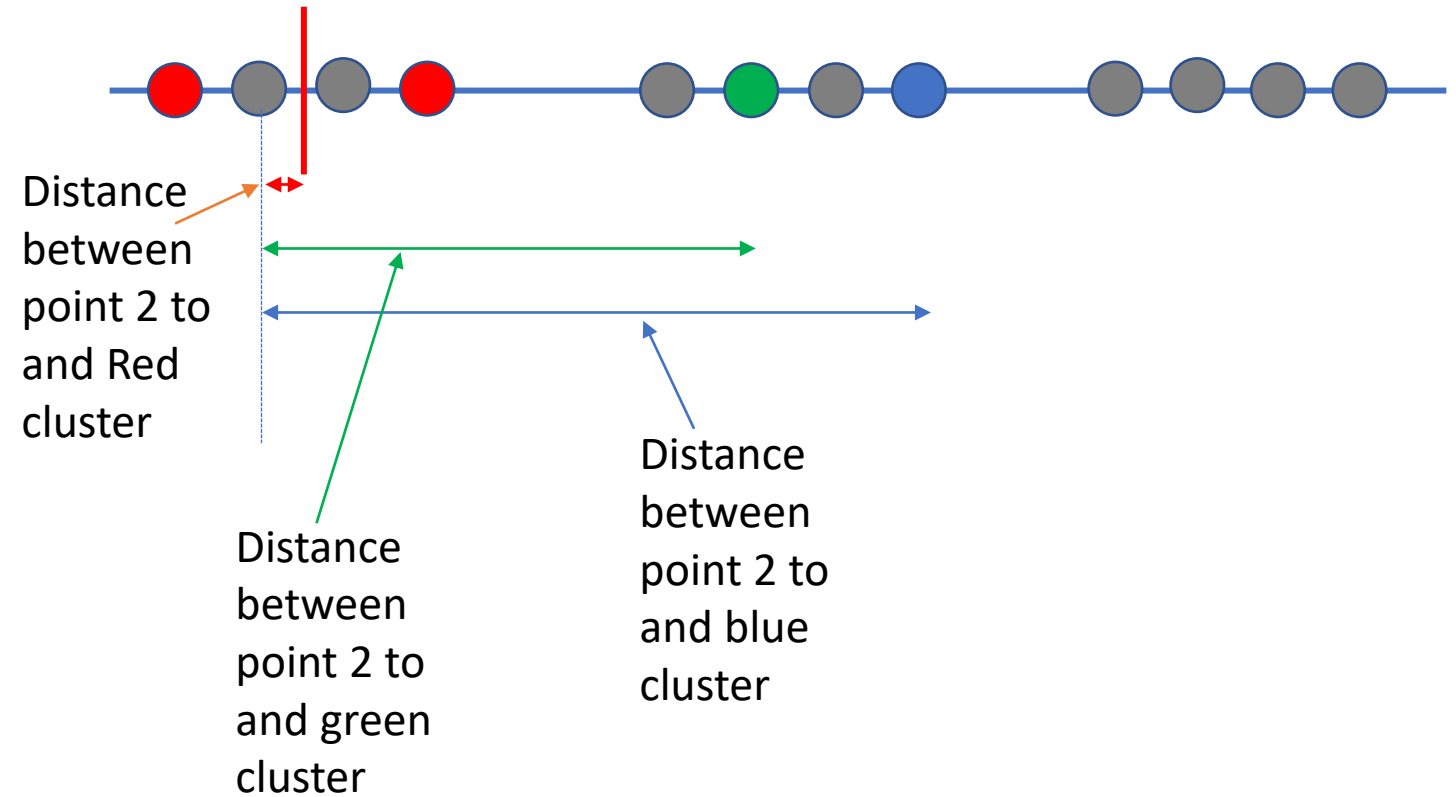
K-Means Algorithm?

- Step 1: Select the number of clusters, i.e. $K=3$ in this case.
- Step 2: Select randomly 3 distinct data points, which will work as initial centroid/ mean of each cluster.
- Step 3: Measure the distance between each point and 3 selected clusters centroid.
- Step 4: Assign the point to the nearest cluster (which the red cluster for first point).
- Step 5: Calculate the new mean of the cluster after adding the new point is added to red cluster.



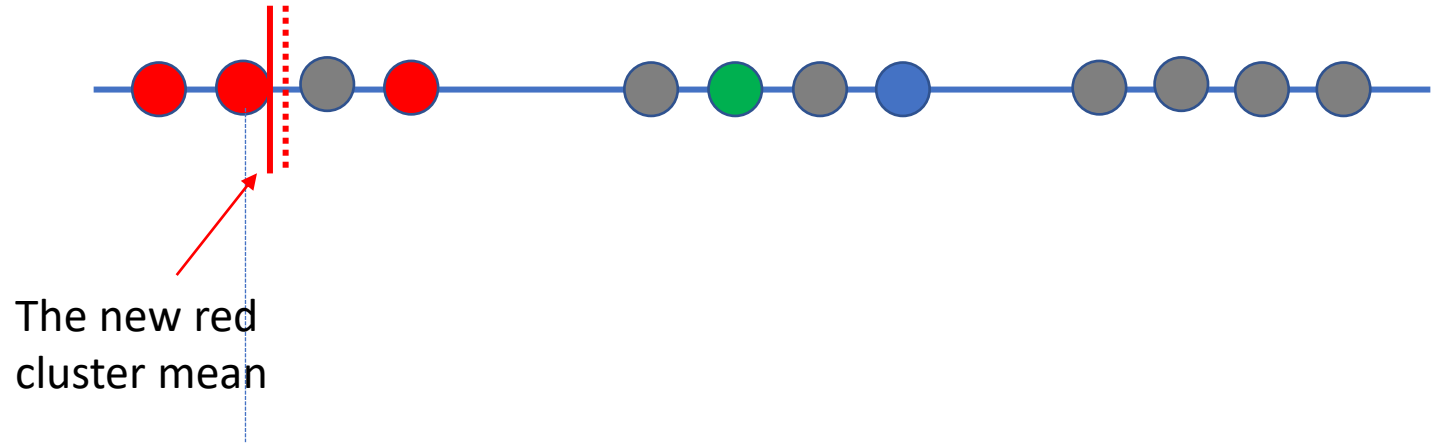
K-Means Algorithm?

- The same steps are repeated to find out to which cluster is the point 2 is belong to ?



K-Means Algorithm?

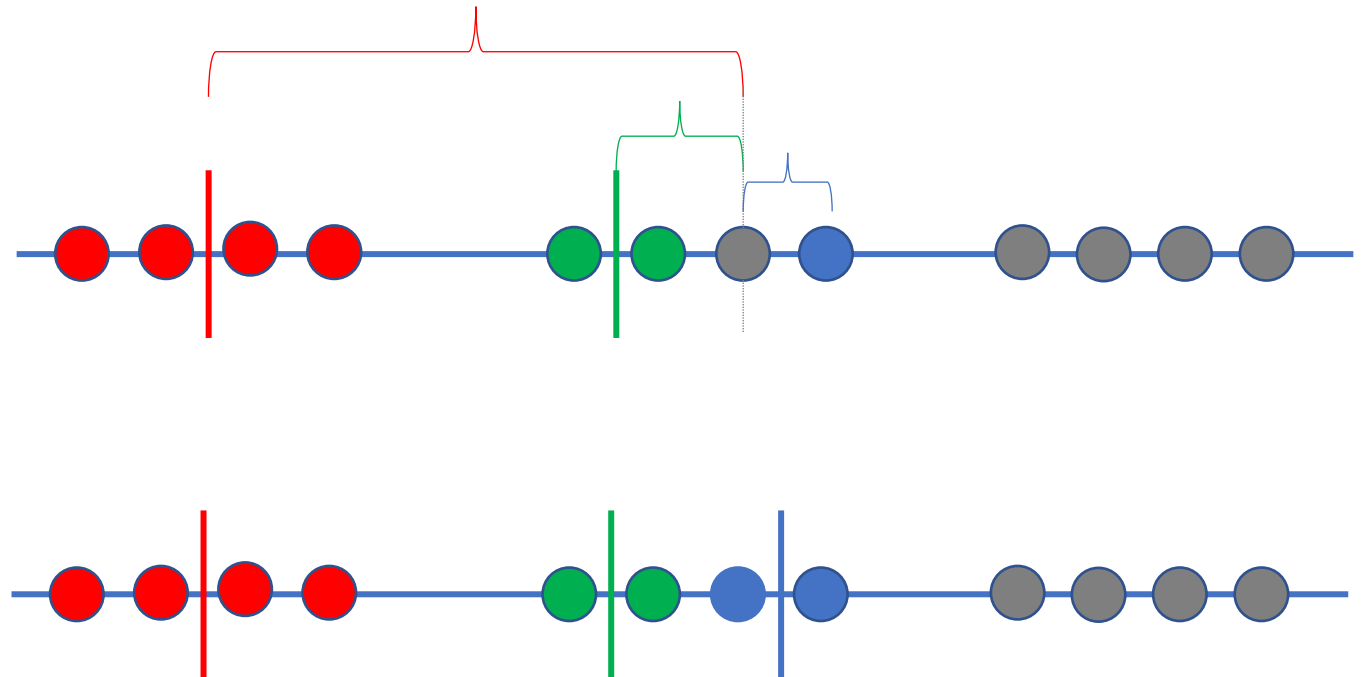
- Assign the point 2 to the nearest cluster
- Calculate the new mean again for the red cluster including the point 2.



K-Means Algorithm?

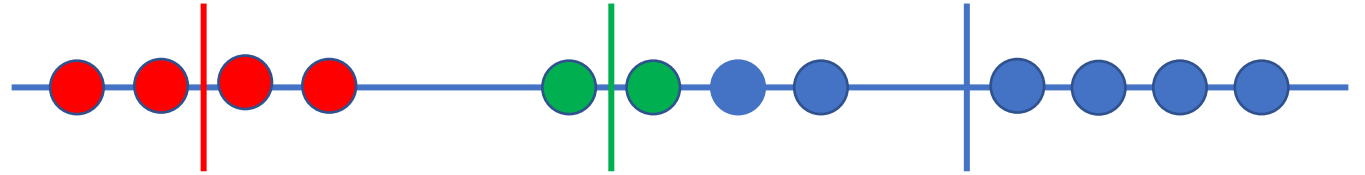
Repeat the same process again and gain until all the data points are assigned into one of the three clusters.

- Measure the distance
- Assign the point to the nearest cluster
- Calculate the new cluster mean using the new point

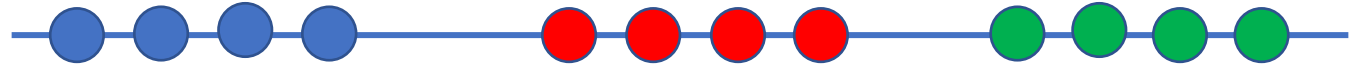


K-Means Algorithm?

The results of first nitration



The expected results



Evaluating the quality of the produced clusters:

Summing up the variation within all the clusters



K-Means Algorithm?

Total variation within the clusters

K-Means algorithm iterates over again and again until the data points within each cluster stop changing

K-Means Algorithm?

Second iteration:

Repeat the same procedure, which starts from selecting randomly the initial cluster centroid. Make sure the selected points are completely different from the ones that have been selected in the previous iterations.



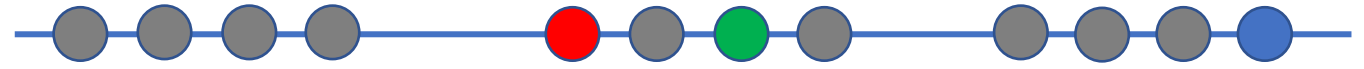
The initial cluster means

- Step 3: Measure the distance between each point and 3 selected clusters centroid.
- Step 4: Assign the point to the nearest cluster.
- Step 5: Calculate the new mean of the cluster after adding the new point is added to the cluster.

K-Means Algorithm?

Second iteration:

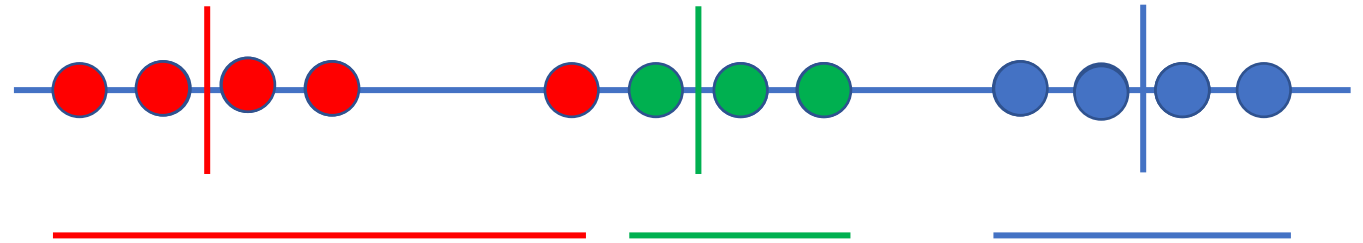
Repeat the same procedure, which starts from selecting randomly the initial cluster centroid. Make sure the selected points are completely different from the ones that have been selected in the previous iterations.



The initial cluster means

K-Means Algorithm?

Second iteration:

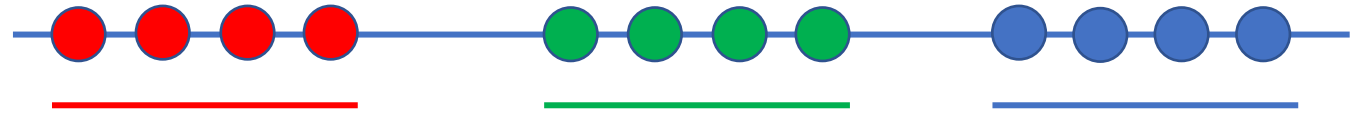


Total variation within the cluster

K-Means Algorithm?

Third iteration:

At every iteration we will start from selecting randomly the initial clusters means. Then we repeat the same steps.



The algorithm is able to compare the results and select the one with the best variation

1st Iteration



2nd Iteration

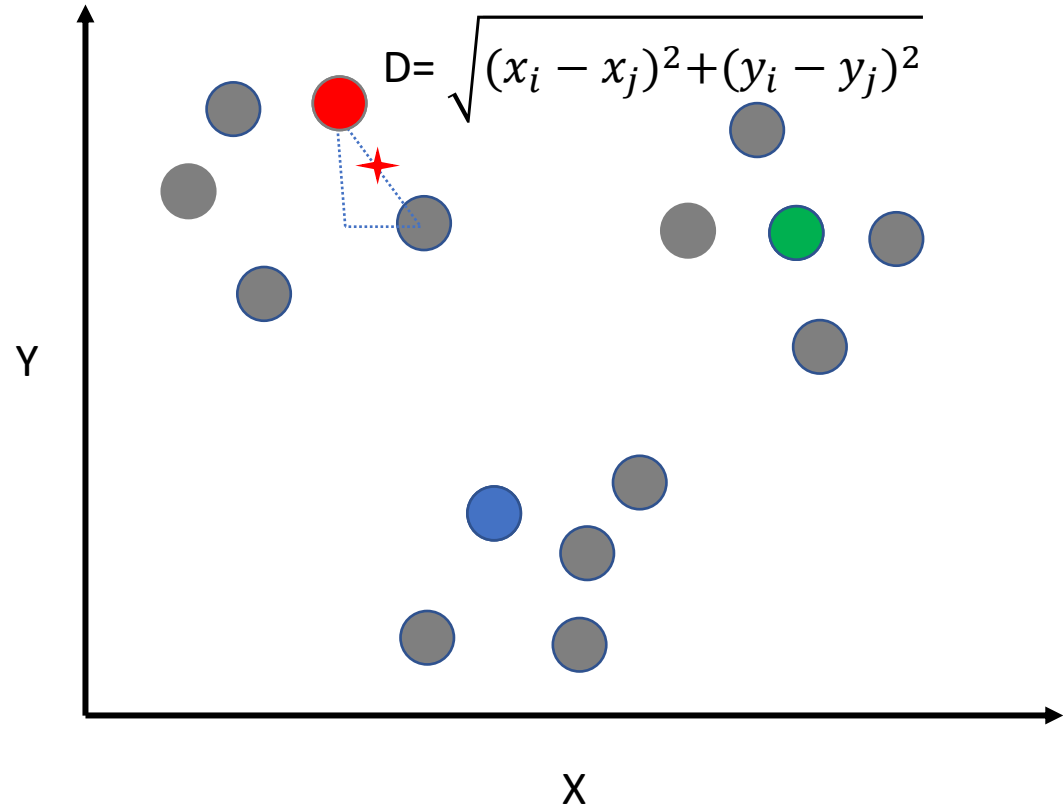


3rd Iteration



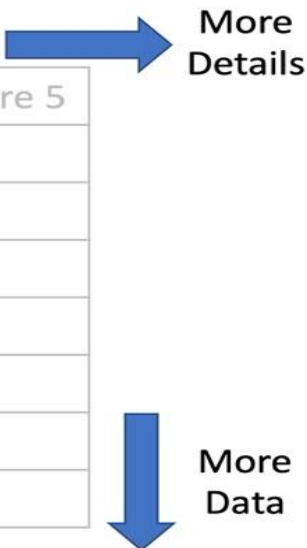
High-dimensional data:

K-Means
Algorithm?



Clustering Distance Measures

- Dimensionality of the data:



Customer	Sat	Sun	Mon	Tue	Feature 5
Jo	3	3	0	0	
Sarah	2	2	0	0	
Davis	5	5	0	0	
Chris	0	0	4	4	
Sophia	0	0	3	3	
Mike	0	0	5	5	

Every data point (observation) can be represented as point in N dimensional space, where N is the number of features

$$X = \{x_1, x_2, \dots, x_N\},$$

Clustering Distance Measures

- Euclidean Distance

The Euclidean distance, also known as L2 norm, between two data point x_i and x_j is

defined as: $D(X_i, X_j) = \sqrt{\sum_{l=1}^N (x_{il} - x_{jl})^2}$

- Where x_{il} and x_{jl} represent the l^{th} dimension of X_i and X_j respectively.
- If one of the input attributes has a relatively large range, then it can overcome the other attributes.

- Manhattan Distance

Manhattan distance also known as a city-block, rectilinear or L1 distance, is mathematically defined as:

$$D(X_i, X_j) = \sum_{l=1}^N |x_{il} - x_{jl}|$$

- The clusters formed using Manhattan distance tend to form rectangular shaped clusters.
- Its advantage over Euclidean distance is the reduction in computational time
- The disadvantage is that it depends upon the rotation of the coordinate system.

- Hamming Distance

- Calculates the distance between two vectors.
- It measures the minimum number of *substitutions* required to change one array into the other, or the minimum number of *errors* that could have transformed one string into the other.
- It works well with binary data.

- Cosine Distance (Similarity)

- It is a measure of similarity between two non-zero vectors.
- It works well with documents similarity.

$$D = \frac{\sum_{l=1}^N x_{jl} x_{il}}{\sqrt{\sum_{l=1}^N x_{il}^2} \sqrt{\sum_{l=1}^N x_{jl}^2}}$$

$$X=[x_1,x_2,x_3,\dots,x_n]$$

$$Y=[y_1,y_2,y_3,\dots,y_n]$$

$$(x_1*y_1+x_2*y_2+x_3*y_3+\dots+x_n*y_n)$$

$$(x_1*x_1+x_2*x_2+x_3*x_3+\dots+x_n*x_n)$$

$$= \frac{\sum_{l=1}^N x_{jl} x_{il}}{\sqrt{\sum_{l=1}^N x_{il}^2} \sqrt{\sum_{l=1}^N x_{jl}^2}}$$

Advantages:

- k-Means is simple and can be used for a wide variety of object types.
- It is also efficient both from storage requirement and execution time point of views. By saving distance information from one iteration to the next, the actual number of distance calculations, that must be made can be reduced (specially, as it reaches towards the termination).

Limitations:

- The k-Means is not suitable for all types of data. For example, k-Means does not work on categorical data because mean cannot be defined.
- k-means finds a local optima and may actually minimize the global optimum.

- K-means has trouble clustering data that contains outliers. When the SSE is used as objective function, outliers can influence the cluster that are produced. More precisely, in the presence of outliers, the cluster centroids, in fact, not truly as representative as they would be otherwise. It also influence the SSE measure as well.
- K-means algorithm are not so practical for large databases.