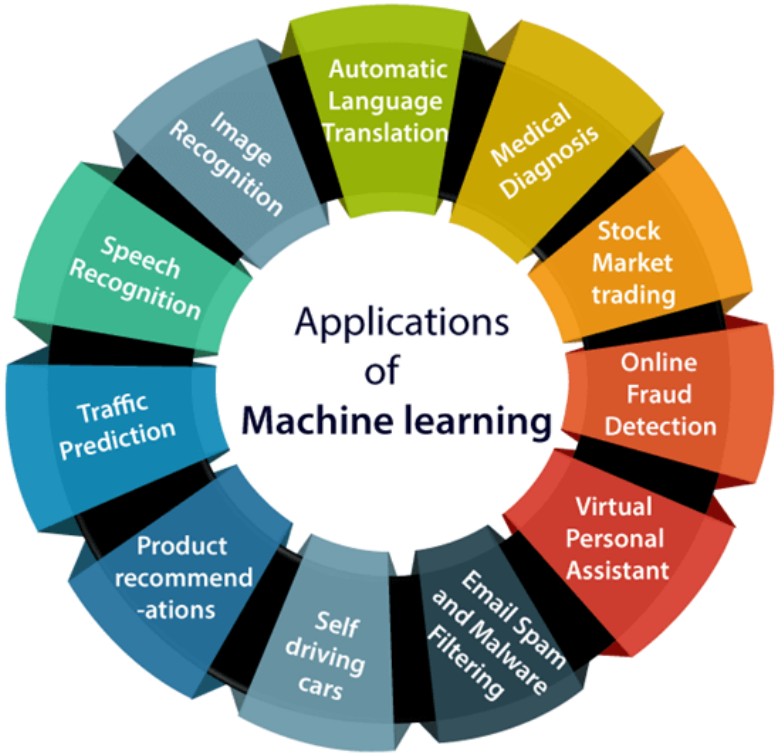
ML Basics

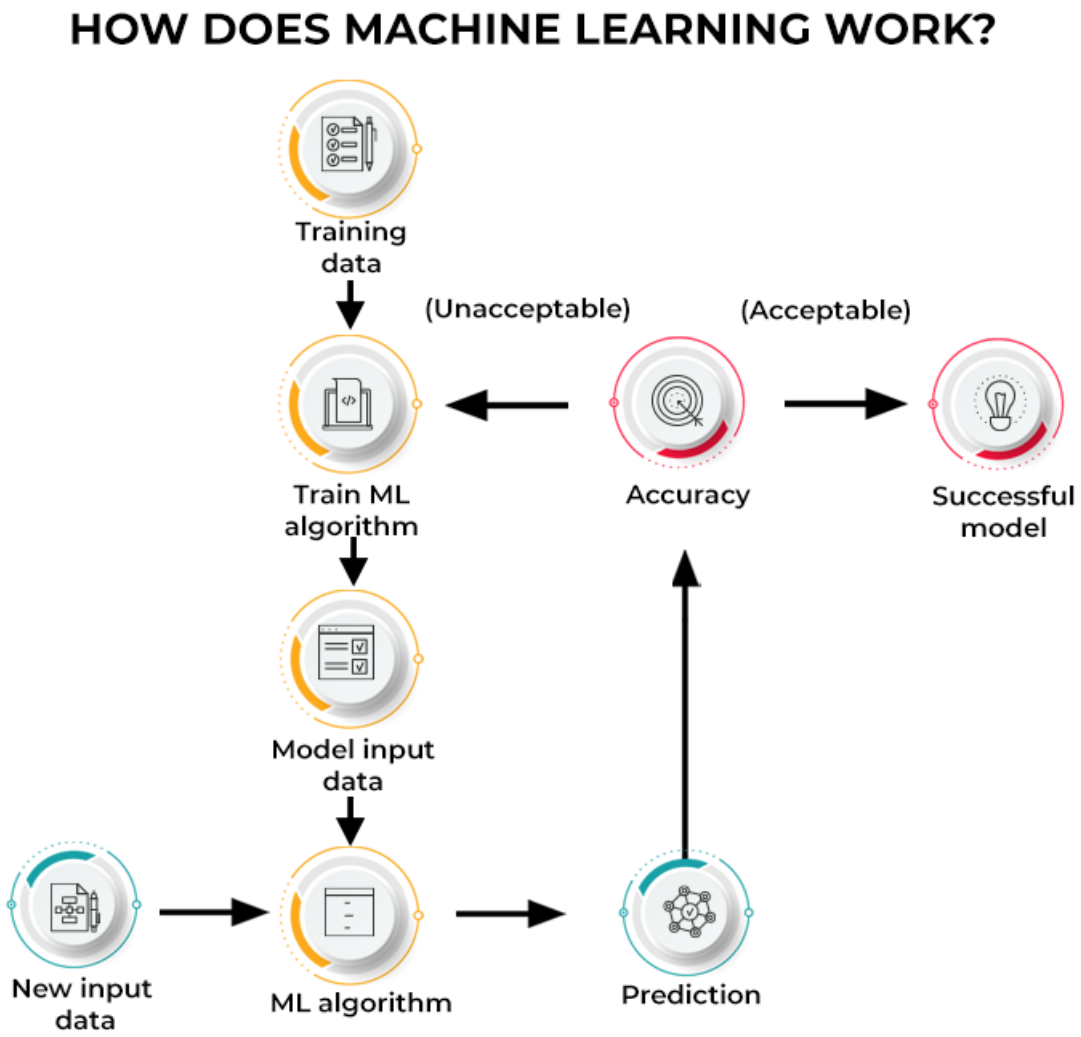
Sachin Tripathi

Overview of ML

Machine learning (ML) is a discipline of artificial intelligence (AI) that provides machines with the ability to automatically learn from data and past experiences while identifying patterns to make predictions with minimal human intervention.



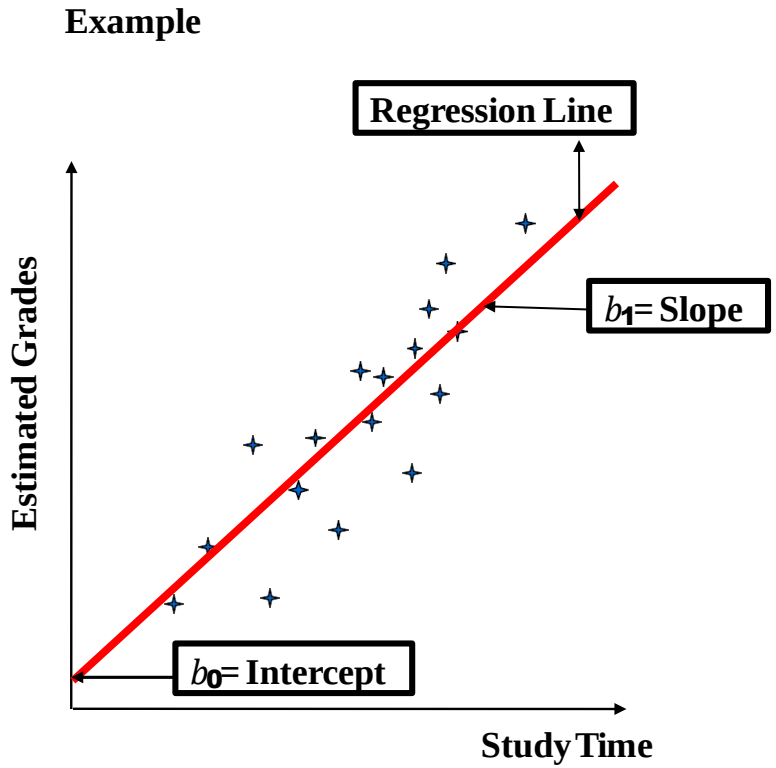
Overview of ML



Linear Regression

- ✓ Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable(s) (predictor).
- ✓ This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.
- ✓ For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

Linear Regression



$$y = b_0 + b_1 x$$

y = Estimated Grades
x = Study Time
b_0 = Intercept
b_1 = Slope

Linear Regression

Typically, a regression analysis is used for these purposes:

(1)Prediction of the target variable (forecasting).

(2)Modelling the relationships between the dependent variable and the explanatory variable.

(3)Testing of hypotheses.

Benefits

1.It indicates the strength of impact of multiple independent variables on a dependent variable.

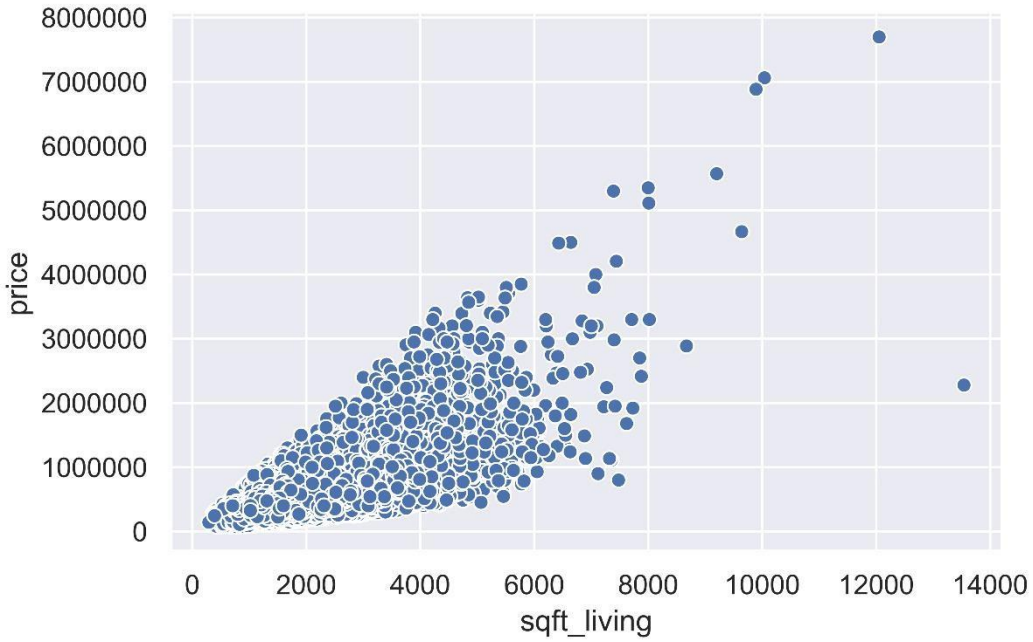
2.It indicates the significant relationships between dependent variable and independent variable.

These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set
of variables to be used for building predictive models.

Linear Regression

- It models the relationship between two variables x and y with a "**line**".
- Linear regression formula: $\mathbf{y = w_0 + w_1 x}$, where
  x: feature, attribute; y: target, outcome; w_0: intercept; w_1: slope

Example: How does the price of a house (y) change relate to its living square footage (x)?



* Data source: King County, WA Housing Info.

Logistic Regression

- A statistical method used to model dichotomous or binary outcomes (but not limited to) using predictor variables.

What is the "Logistic" component?

Instead of modeling the outcome, Y, directly, the method models the log odds(Y) using the logistic function.

What is the "Regression" component?

Methods used to quantify association between an outcome and predictor variables. Could be used to build predictive models as a function of predictors.

Forms of Learning

A machine is said to be learning from past Experiences(data feed in) with respect to some class of tasks if its Performance in each Task improves with the Experience.

For example, assume that a machine must predict whether a customer will buy a specific product let's say "Antivirus" this year or not.
The machine will do it by looking at the previous knowledge/past experiences i.e., the data of products that the customer had bought every year and if he buys Antivirus every year, then there is a high probability that the customer is going to buy an antivirus this year as well.

This is how machine learning works at the basic conceptual level.

Learning is the process of converting experience into expertise or knowledge. Learning can be broadly classified into 4 categories: Supervised, Semi-supervised, Unsupervised, Transfer Learning.

Forms of Learning – Supervised Learning

- Supervised learning, as the name indicates, has the presence of a supervisor as a teacher.
- Basically, supervised learning is when we teach or train the machine using data that is well labeled.
- Which means some data is already tagged with the correct answer.
- After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labeled data.
- For instance, suppose you are given a basket filled with different kinds of fruits. Now the first step is to train the machine with all different fruits one by one like this:
  - If the shape of the object is rounded and has a depression at the top, is red in color, then it will be labeled as –**Apple**.
  - If the shape of the object is a long curving cylinder having Green-Yellow color, then it will be labeled as –**Banana**.

Forms of Learning – Supervised Learning

- Supervised learning is when the model is getting trained on a labelled dataset.
- A labelled dataset is one that has both input and output parameters.
- In this type of learning both training and validation, datasets are labelled as shown in the figure below:

User ID	Gender	Age	Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	1
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	1
15728773	Male	27	58000	1
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	1
15727311	Female	35	65000	0
15570769	Female	26	80000	1
15606274	Female	26	52000	0
15746139	Male	20	86000	1
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1

Figure A: CLASSIFICATION

Temperature	Pressure	Relative Humidity	Wind Direction	Wind Speed
10.69261758	986.882019	54.19337313	195.7150879	3.278597116
13.59184184	987.8729248	48.0648859	189.2951202	2.909167767
17.70494885	988.1119385	39.11965597	192.9273834	2.973036289
20.95430404	987.8500366	30.66273218	202.0752869	2.965289593
22.9278274	987.2833862	26.06723423	210.6589203	2.798230886
24.04233986	986.2907104	23.46918024	221.1188507	2.627005816
24.41475295	985.2338867	22.25082295	233.7911987	2.448749781
23.93361956	984.8914795	22.35178837	244.3504333	2.454271793
22.68800023	984.8461304	23.7538641	253.0864716	2.418341875
20.56425726	984.8380737	27.07867944	264.5071106	2.318677425
17.76400389	985.4262085	33.54900114	280.7827454	2.343950987
11.25680746	988.9386597	53.74139903	68.15406036	1.650191426
14.37810685	989.6819458	40.70884681	72.62069702	1.553469896
18.45114201	990.2960205	30.85038484	71.70604706	1.005017161
22.54895853	989.9562988	22.81738811	44.66042709	0.264133632
24.23155922	988.796875	19.74790765	318.3214111	0.329656571

Figure B: REGRESSION

Forms of Learning – Supervised Learning

Types of Supervised Learning Algorithms:
Classification:

- It is a Supervised Learning task where output is having defined labels(discrete value).
- For example, in **Figure A**, Output – Purchased has defined labels i.e., 0 or 1;
  - 1 means the customer will purchase and 0 means that customer won't purchase.
  - The goal here is to predict discrete values belonging to a particular class and evaluate them based on accuracy.
  - It can be either binary or multi-class classification. In binary classification, the model predicts either 0 or 1; yes or no but in the case of multi-class classification, the model predicts more than one class.
- Example: Gmail classifies mails in more than one class like social, promotions, updates, forums.

Regression:

- It is a Supervised Learning task where output is having continuous value.
- Example in **Figure B**, Output – Wind Speed is not having any discrete value but is continuous in the range.
- The goal here is to predict a value as much closer to the actual output value as our model can and then evaluation is done by calculating the error value.
- The smaller the error the greater the accuracy of our regression model.

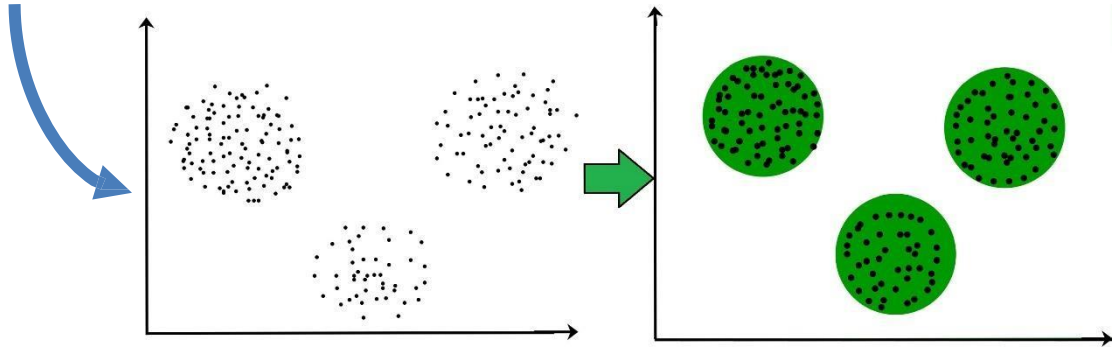
Forms of Learning – Unsupervised Learning

- Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance.
- Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.
- Unlike supervised learning, no teacher is provided, that means no training will be given to the machine.
- Therefore, the machine is restricted to find the hidden structure in unlabeled data by itself.
- For instance, suppose it is given an image having both dogs and cats which it has never seen.
  - Thus, the machine has no idea about the features of dogs and cats so we can't categorize it as 'dogs and cats '.
  - But it can categorize them according to their similarities, patterns, and differences, i.e., we can easily categorize the above picture into two parts.
  - The first may contain all pics having dogs in them and the second part may contain all pics having cats in them. Here we didn't learn anything before, which means no training data or examples.
- It allows the model to work on its own to discover patterns and information that was previously undetected. It mainly deals with unlabeled data.

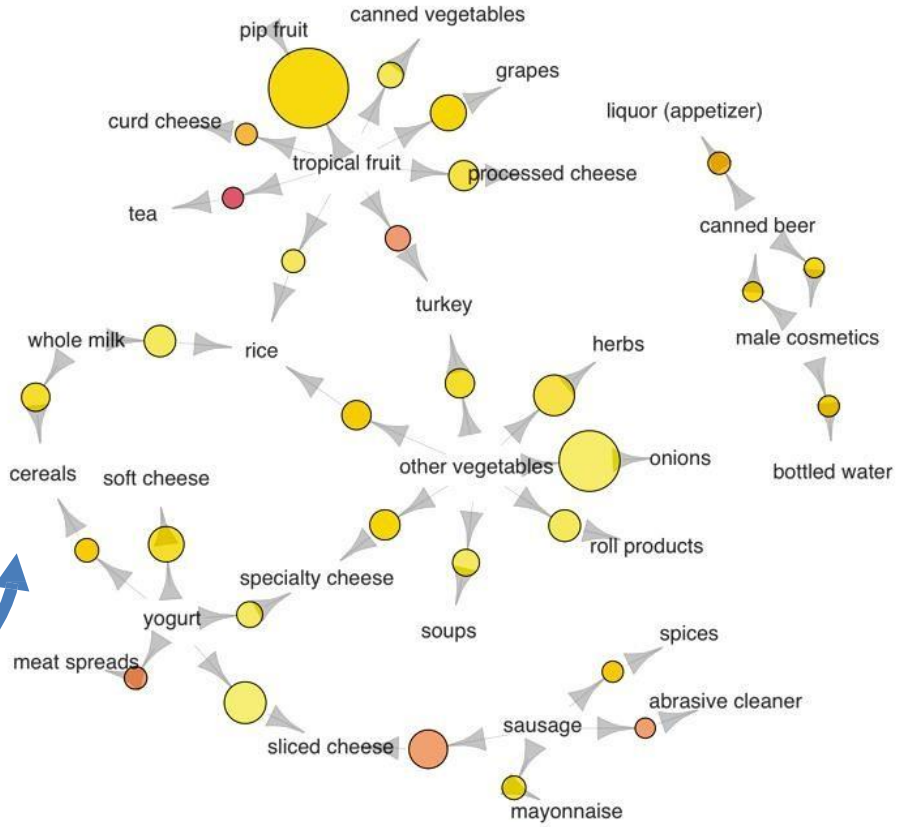
Forms of Learning – Unsupervised Learning

Types of Unsupervised Learning Algorithms:

- **Clustering:** A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.

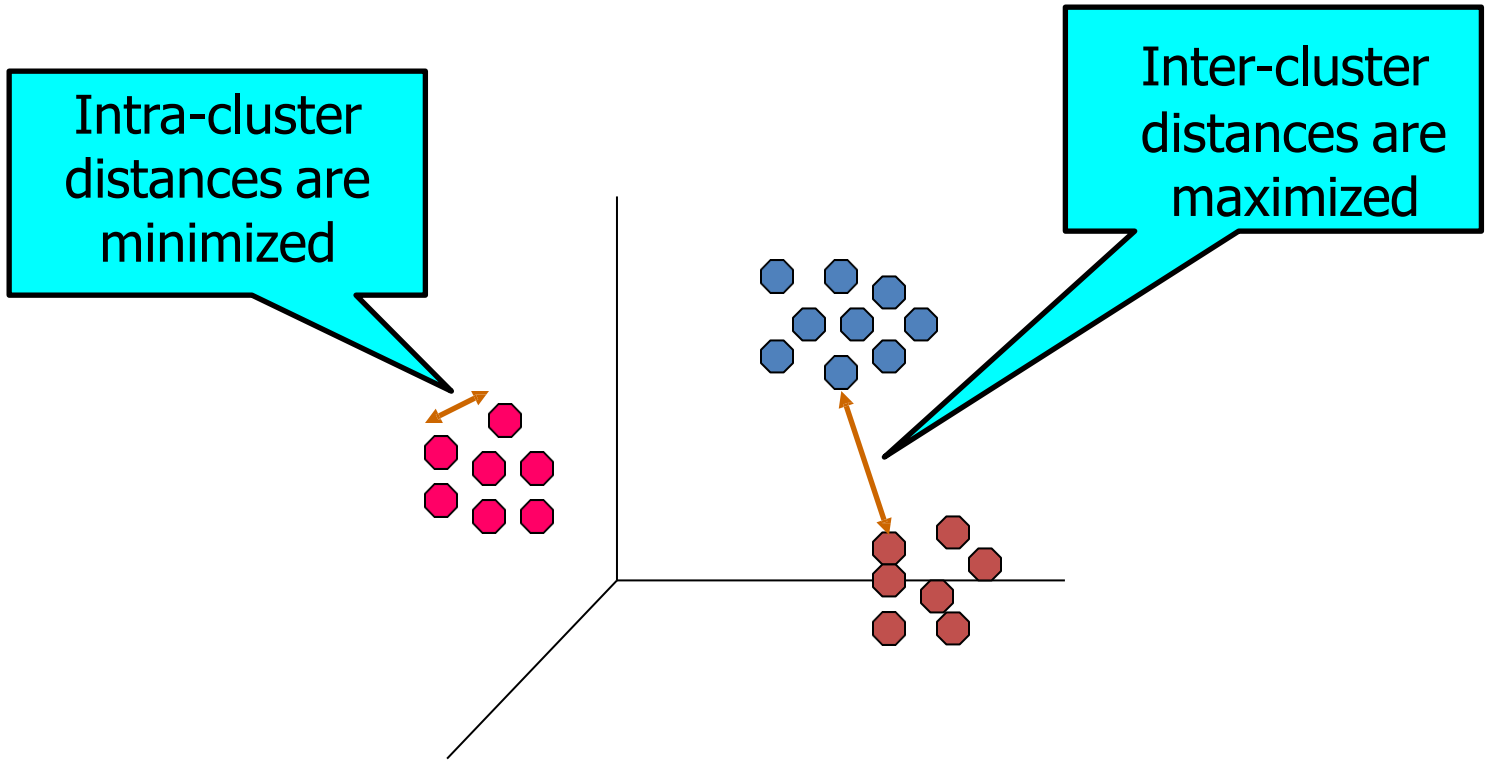


- **Association:** An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.



Forms of Learning – Unsupervised Learning - Clustering

In general a grouping of objects such that the objects in a group (cluster) are similar (or related) to one another and different from (or unrelated to) the objects in other groups



Forms of Learning – Unsupervised Learning - Clustering

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

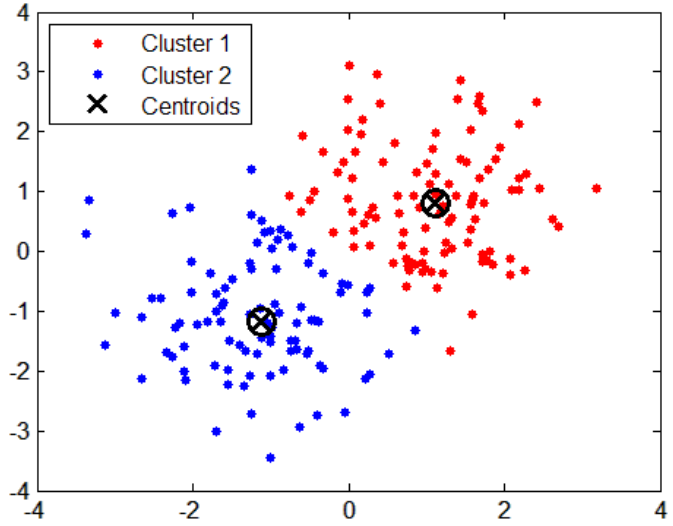
Two Clusters

Four Clusters

Forms of Learning – Unsupervised Learning - Clustering

K-Means Clustering

- K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science.
- K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.
- It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.
- It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

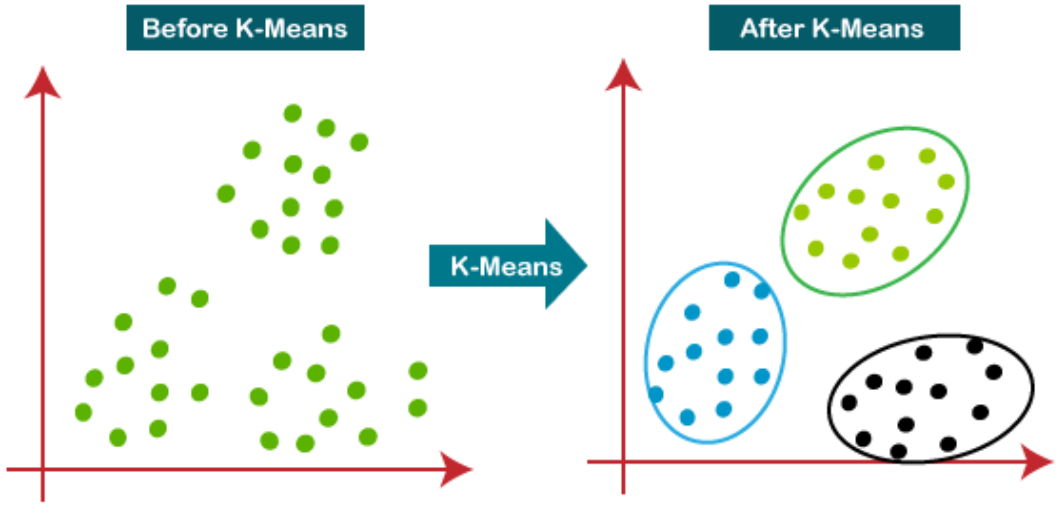


Forms of Learning – Unsupervised Learning - Clustering

K-Means Clustering

- The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

- 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.
- Hence each cluster has datapoints with some commonalities, and it is away from other clusters.



Forms of Learning – Unsupervised Learning - Clustering

K-Means Clustering

The working of the K-Means algorithm is explained in the below steps:

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other from the input dataset).

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

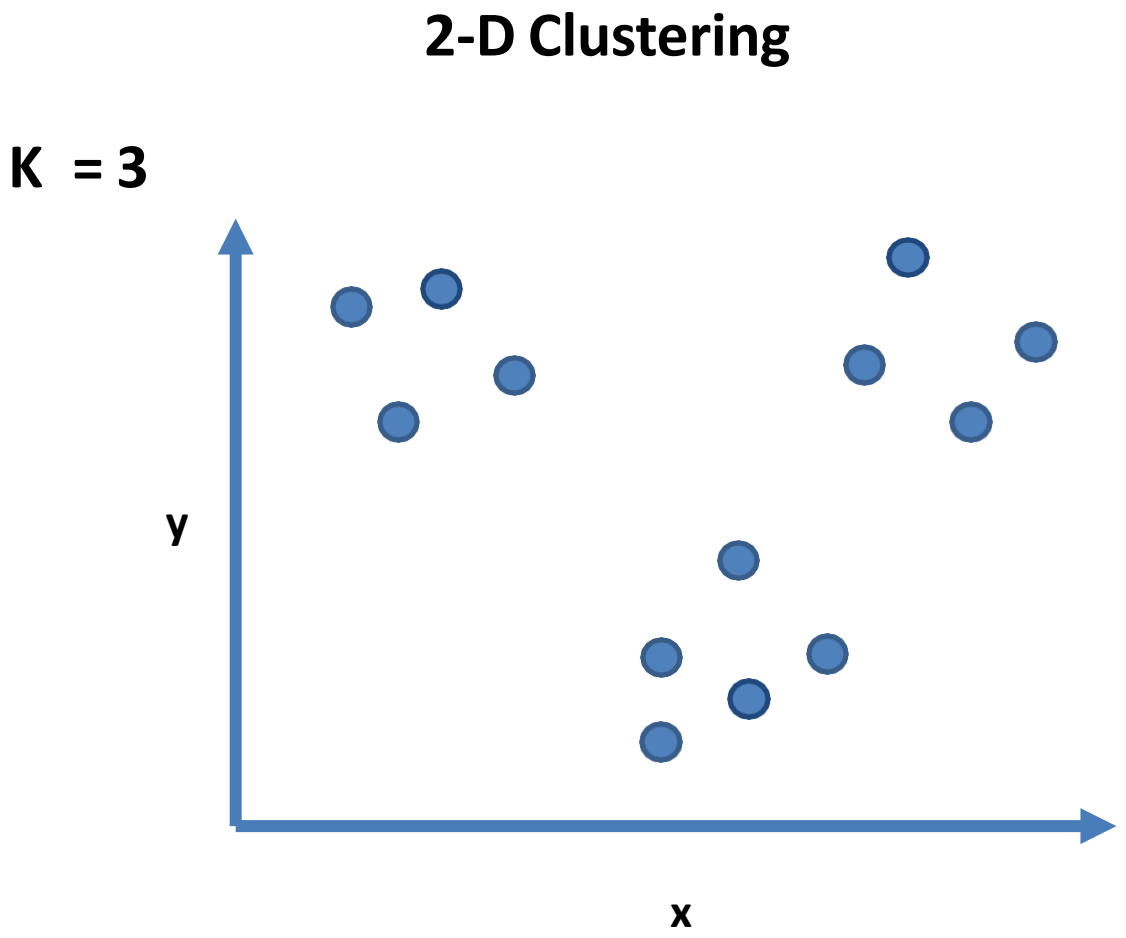
Step-4: Calculate the variance and place a new centroid of each cluster.

Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

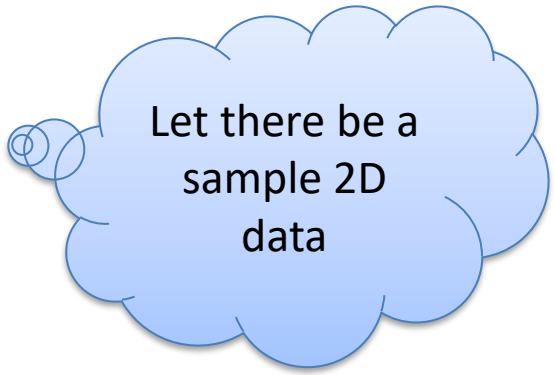
Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.

Step-7: The model is ready.

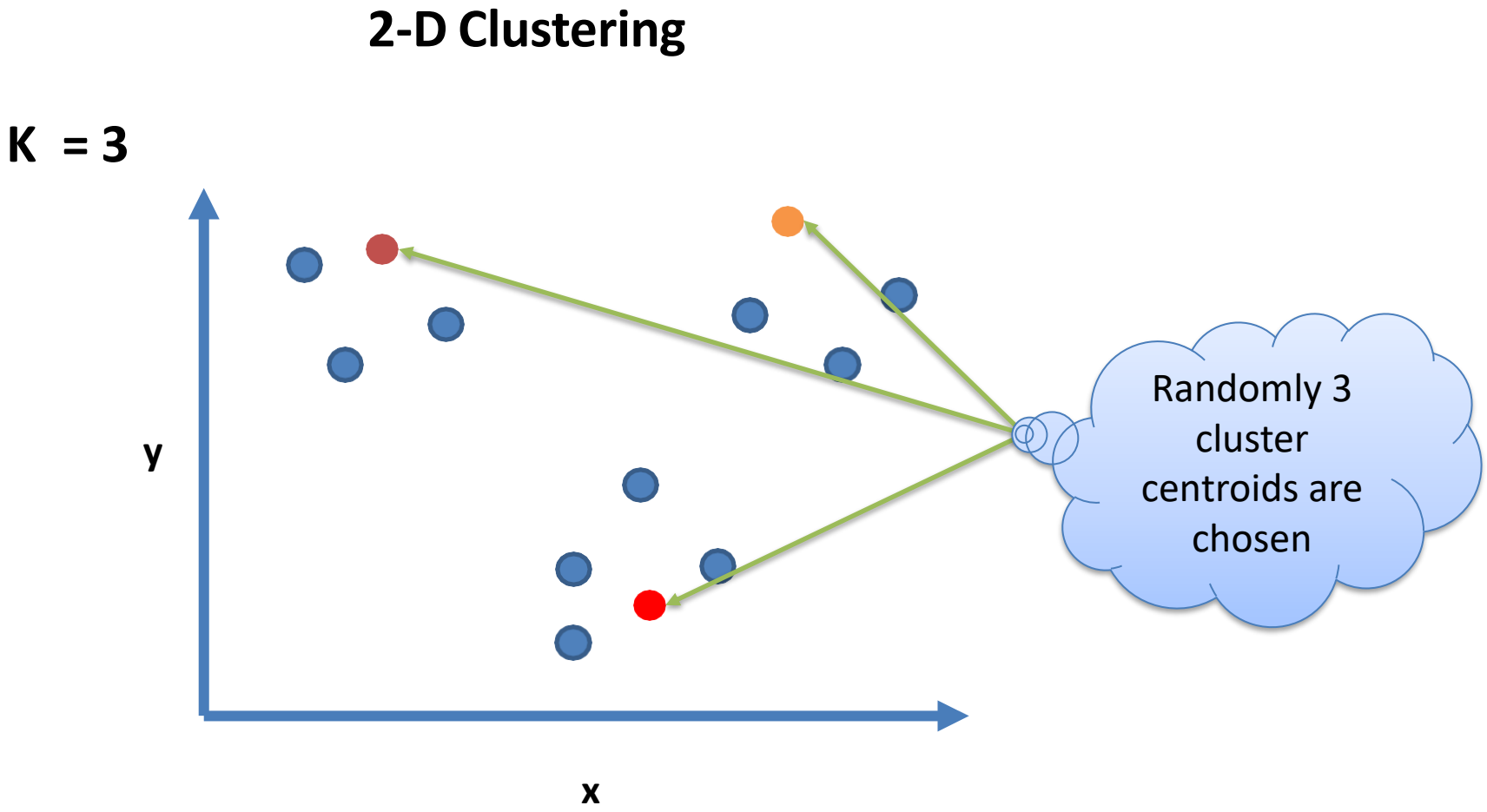
Forms of Learning – Unsupervised Learning - Clustering



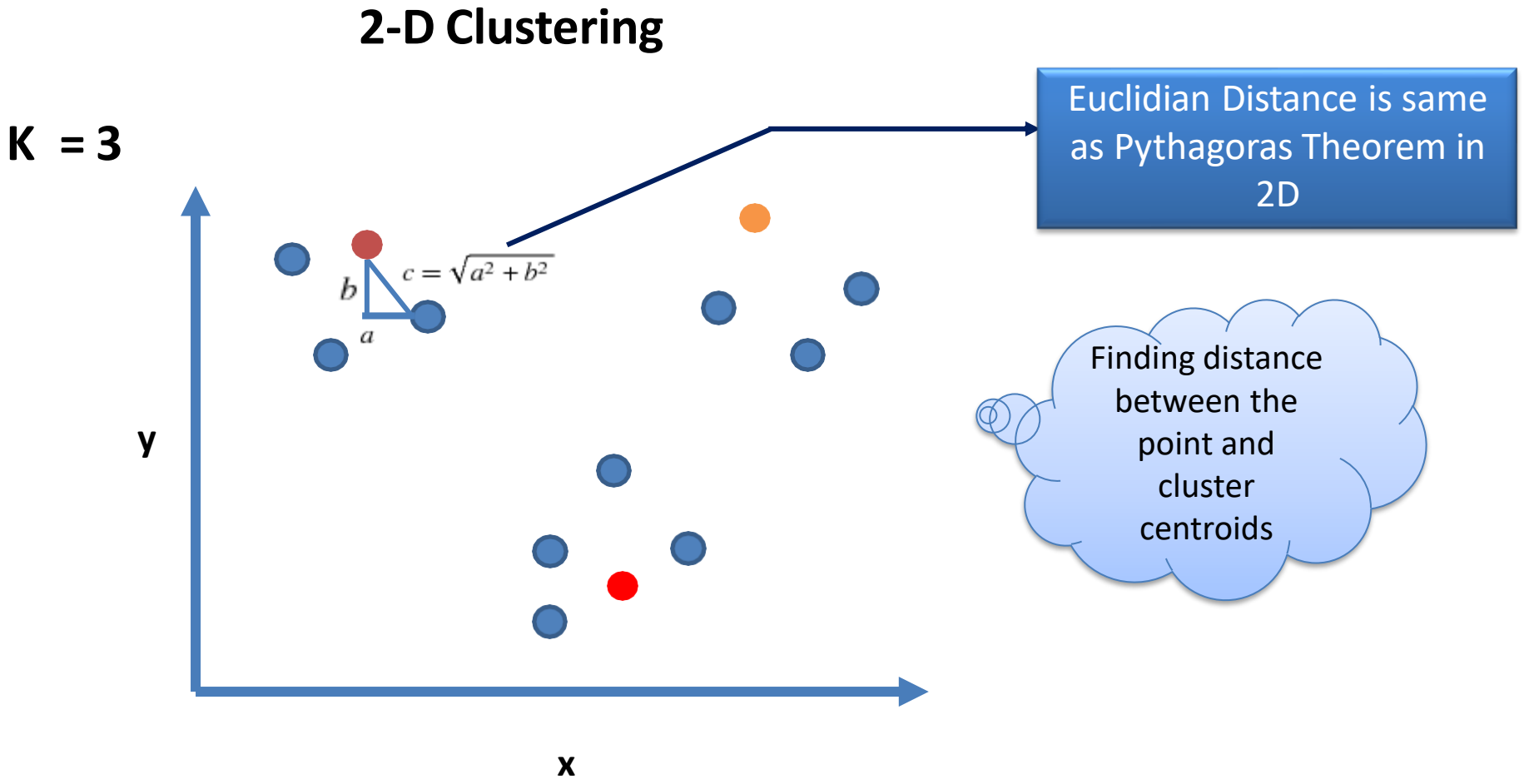
Interactive Clustering – https://kkevsterrr.github.io/K-Means/



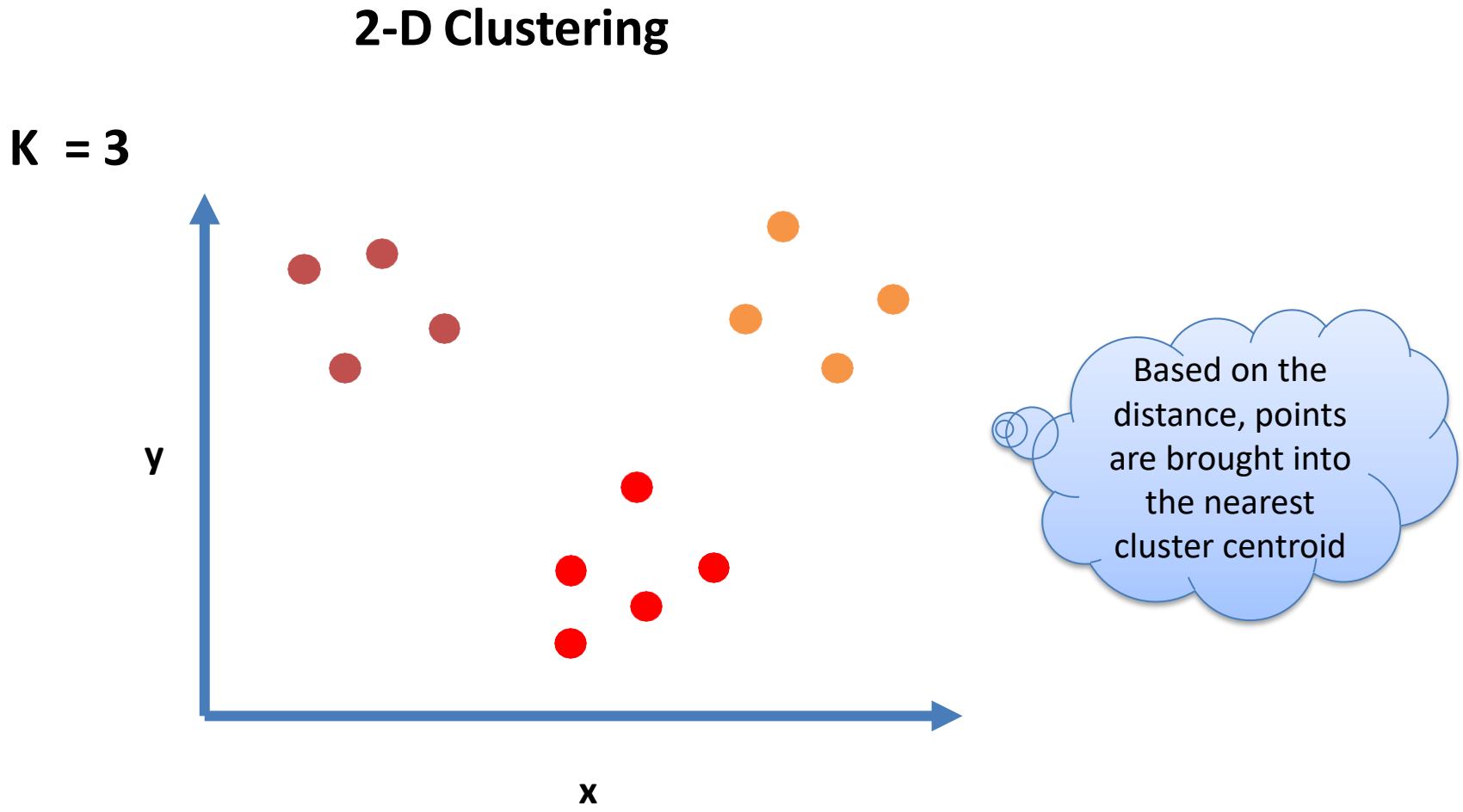
Forms of Learning – Unsupervised Learning - Clustering



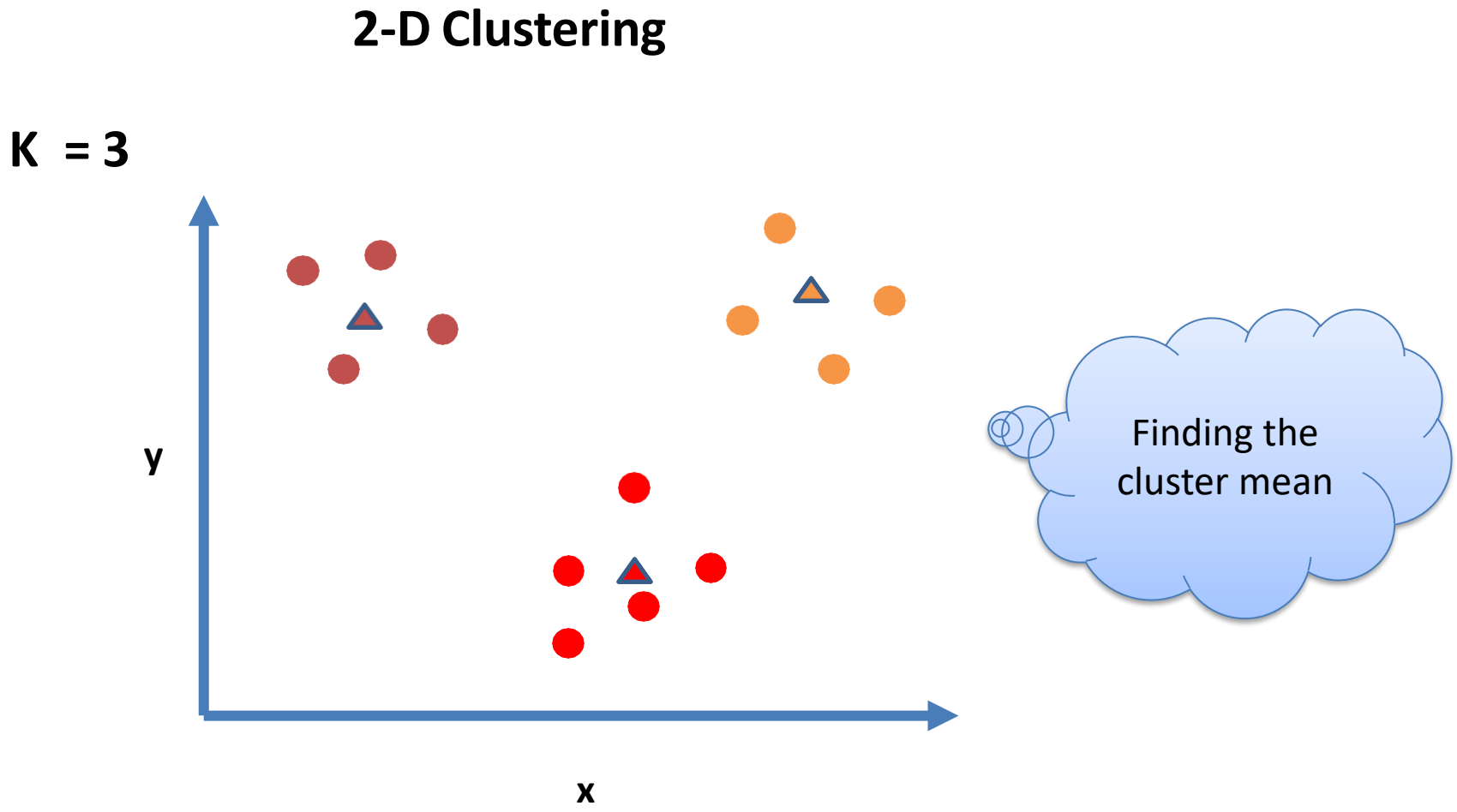
Forms of Learning – Unsupervised Learning - Clustering



Forms of Learning – Unsupervised Learning - Clustering



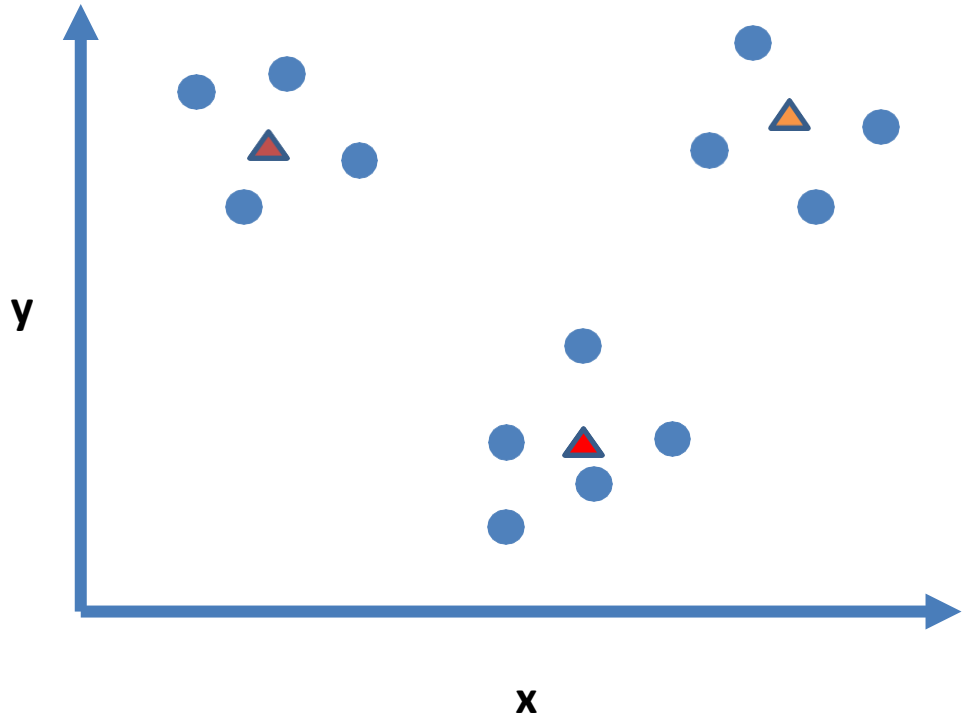
Forms of Learning – Unsupervised Learning - Clustering



Forms of Learning – Unsupervised Learning - Clustering

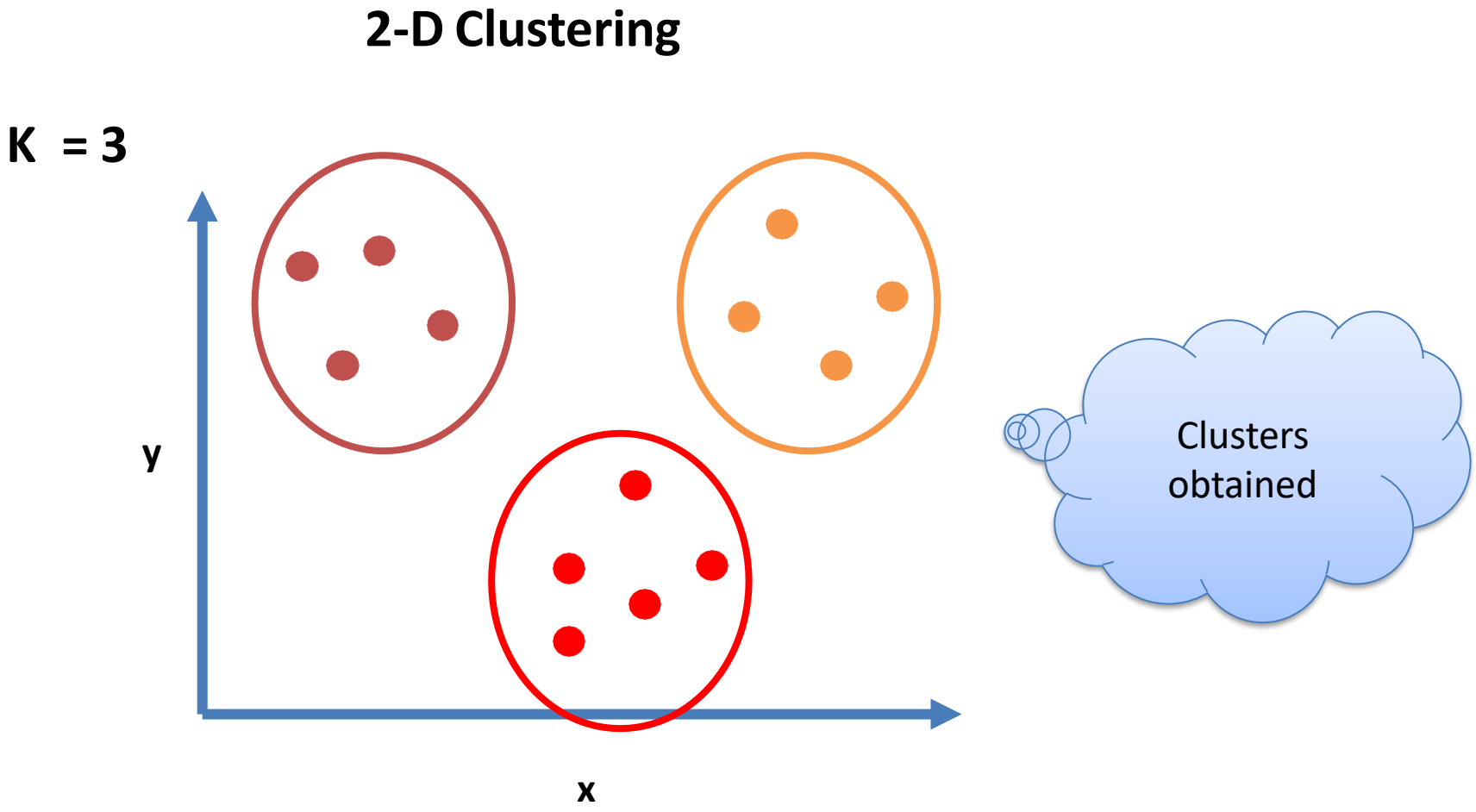
2-D Clustering

K = 3



Reiterating the process of finding distance and stopping if clusters don't change

Forms of Learning – Unsupervised Learning - Clustering

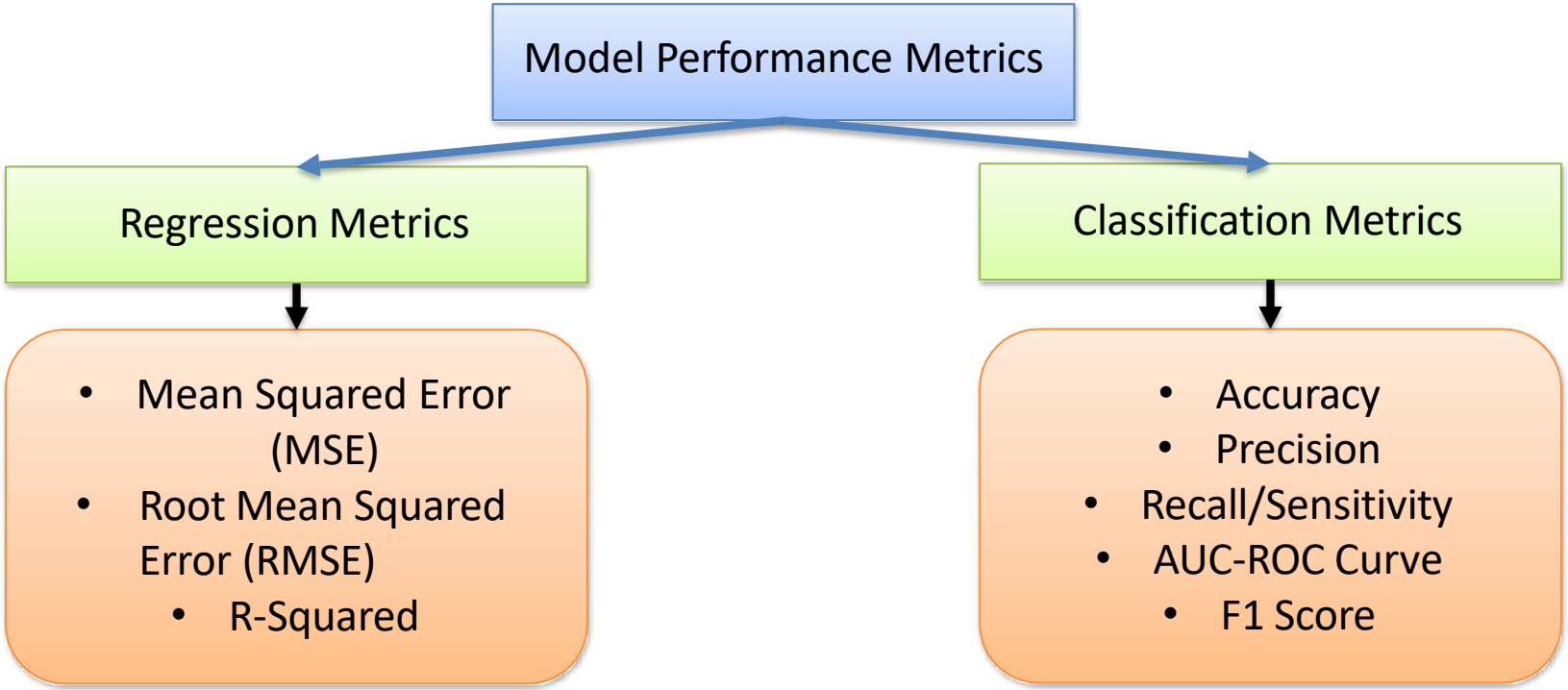


Forms of Learning – Semi-supervised Learning

- If some learning samples are labeled, but some other are not labeled, then it is semi-supervised learning.
- It makes use of a large amount of unlabeled data for training and a small amount of labeled data for testing.
- Semi-supervised learning is applied in cases where it is expensive to acquire a fully labeled dataset while more practical to label a small subset.
- For example, it often requires skilled experts to label certain remote sensing images, and lots of field experiments to locate oil at a particular location, while acquiring unlabeled data is relatively easy.
- Semi-supervised learning falls somewhere in the middle of supervised and unsupervised learning.
- It is used because many problems that AI is used to solving require a balance of both approaches.
- In many cases the reference data needed for solving the problem is available, but it is either incomplete or somehow inaccurate.
- This is when semi-supervised learning is summoned for help since it can access the available reference data and then use unsupervised learning techniques to do its best to fill the gaps.
- Unlike supervised learning which uses labelled data and unsupervised which is given no labelled data at all, Semi-supervised learning uses both.

Model Evaluation and Performance Metrics

- Performance metrics are a part of every machine learning pipeline.
- Every machine learning task can be broken down to either Regression or Classification, just like the performance metrics.
- Metrics are used to monitor and measure the performance of a model (during training and testing)



Regression Metrics

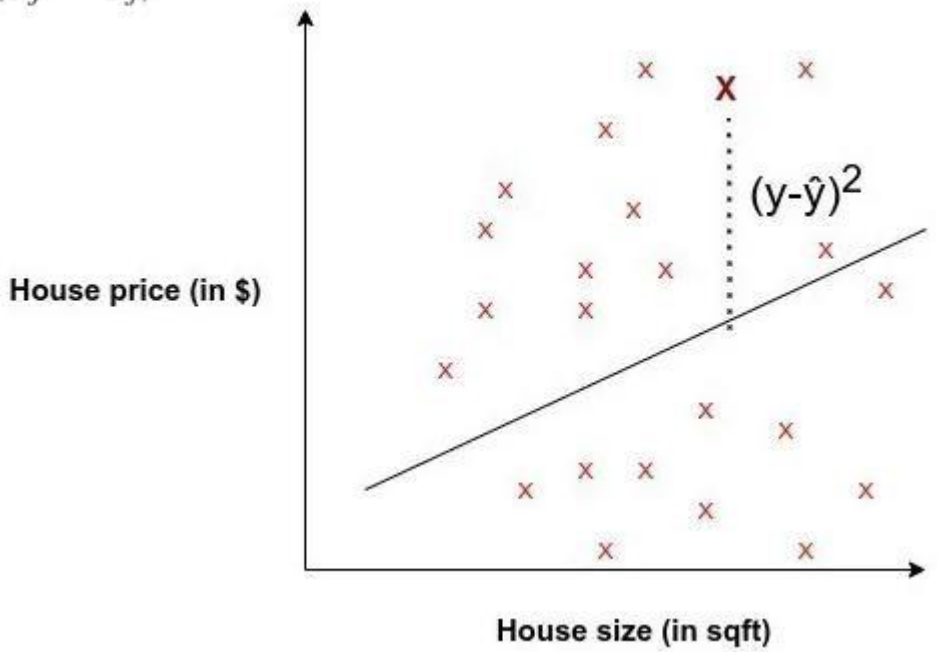
MEAN SQUARED ERROR (MSE)

- Mean squared error is perhaps the most popular metric used for regression problems. It essentially finds the average of the squared difference between the target value and the value predicted by the regression model.

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(y_j - \check{y}_j)^2$$

Where:

- y_j: ground-truth value
- y_hat: predicted value from the regression model
- N: number of datums



Regression Metrics

ROOT MEAN SQUARED ERROR (RMSE)

- Root Mean Squared Error corresponds to the square root of the average of the squared difference between the target value and the value predicted by the regression model. Basically, sqrt(MSE). Mathematically it can be represented as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_j - \check{y}_j)^2}$$

Where:

- y_j: ground-truth value
- y_hat: predicted value from the regression model
- N: number of datums

It addresses a few downsides in MSE.

Few key points related to RMSE:

- It retains the differentiable property of MSE.
- It handles the penalization of smaller errors done by MSE by square rooting it.
- Error interpretation can be done smoothly, since the scale is now the same as the random variable.
- Since scale factors are essentially normalized, it's less prone to struggle in the case of outliers.

Regression Metrics

ROOT MEAN SQUARED ERROR (RMSE)

RMSE is the most popular evaluation metric used in regression problems. It follows an assumption that error are unbiased and follow a normal distribution. Here are the key points to consider on RMSE:

- The power of ‘square root’ empowers this metric to show large number deviations.
- The ‘squared’ nature of this metric helps to deliver more robust results which prevents cancelling the positive and negative error values. In other words, this metric aptly displays the plausible magnitude of error term.
- It avoids the use of absolute error values which is highly undesirable in mathematical calculations.
- When we have more samples, reconstructing the error distribution using RMSE is more reliable.
- RMSE is highly affected by outlier values. Hence, make sure you’ve removed outliers from your data set prior to using this metric.
- As compared to mean absolute error, RMSE gives higher weightage and punishes large errors.

Regression Metrics

R-SQUARED (Coefficient of Determination)

- R^2 - Coefficient of determination works as a post metric, meaning it's a metric that's calculated using other metrics.

- The point of even calculating this coefficient is to answer the question **"How much (what %) of the total variation in Y(target) is explained by the variation in X(regression line)"**. This is calculated using the sum of squared errors.

- R-squared is a statistical measure of how close the data are to the fitted regression line(For measuring the **goodness of fit**). It is also known as the coefficient of determination.

- Firstly, we calculate distance between actual values and mean value and calculate distance between estimated value and mean value.

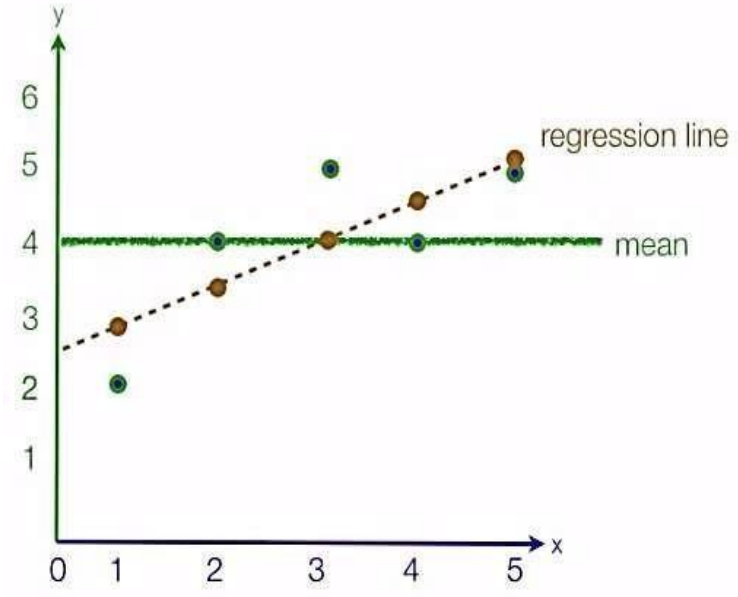
- Then compare both the distances.

Regression Metrics

R-SQUARED (Coefficient of Determination) – Formula

$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}}$$

$$\frac{\text{MSE(model)}}{\text{MSE(baseline)}} \qquad \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (\overline{y}_i - \hat{y}_i)^2}$$

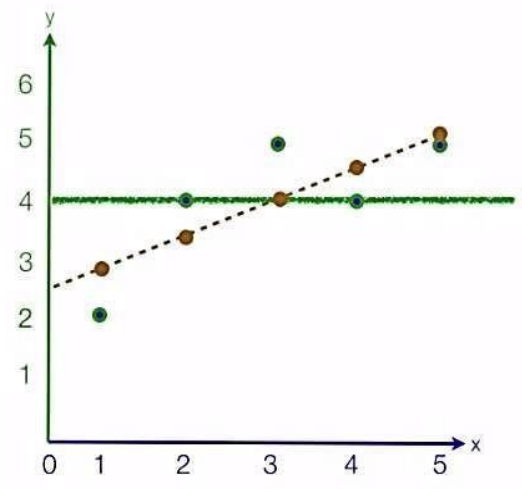


distance actual - mean compare distance estimated - mean

R^2

Regression Metrics

R-SQUARED (Coefficient of Determination) – Manual Calculation Example

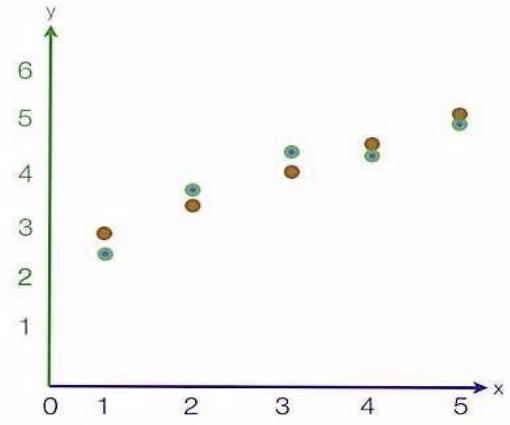


x	y	$y - \bar{y}$	$(y - \bar{y})^2$	$\hat{y}$	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$
1	2	-2	4	2.8	-1.2	1.44
2	4	0	0	3.4	-.6	.36
3	5	1	1	4	0	0
4	4	0	0	4.6	.6	.36
5	5	1	1	5.2	1.2	1.44
			6			3.6
mean	4					

$$R^2 = \frac{3.6}{6} = .6 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2}$$

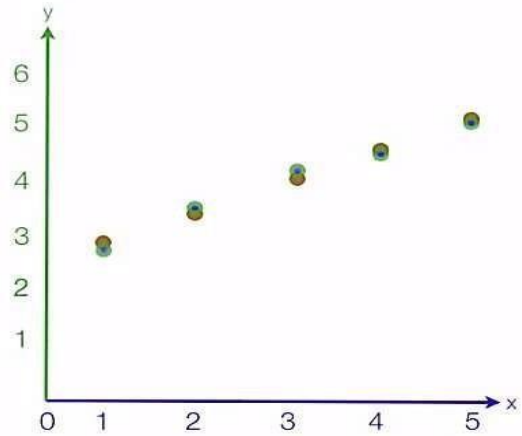
Regression Metrics

R-SQUARED(Coefficient of Determination) – Performance Comparison



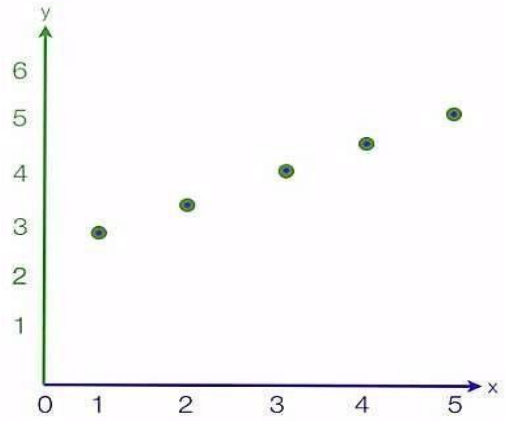
60 %

R^2 =.6



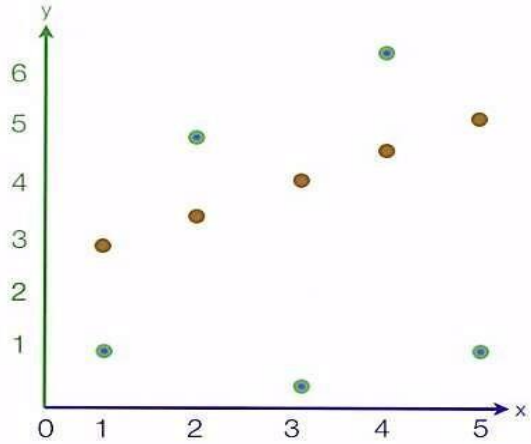
90 %

R^2 =.90



100 %

R^2 = 1



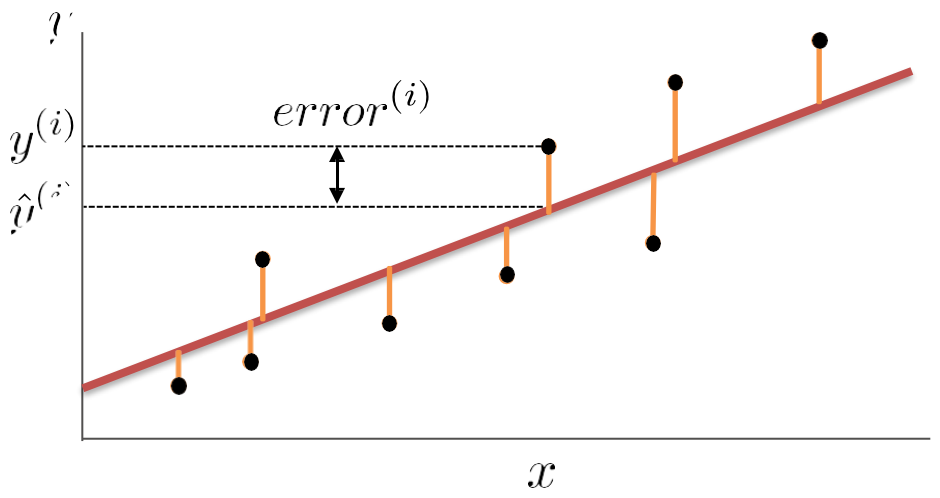
20 %

R^2 = .02

Regression Metrics

Metrics	Equations
Mean Squared Error (MSE)	$MSE = \frac{1}{n}\sum_{i=0}^{n}(y^{(i)} - \hat{y}^{(i)})^2$
Root Mean Squared Error (RMSE)	$RMS = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(y^{(i)} - \hat{y}^{(i)})^2}$
R Squared (R^2)	$R^2 = 1 - \frac{\sum_{i=0}^{n}(y^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=0}^{n}(y^{(i)} - \bar{y})^2}$

$y^{(i)}$: Data values
$\hat{y}^{(i)}$: Predicted values
$\bar{y}$: Mean value of data values, $\bar{y} = \frac{1}{n}\sum_{i=0}^{n} y^{(i)}$
n : Number of data records

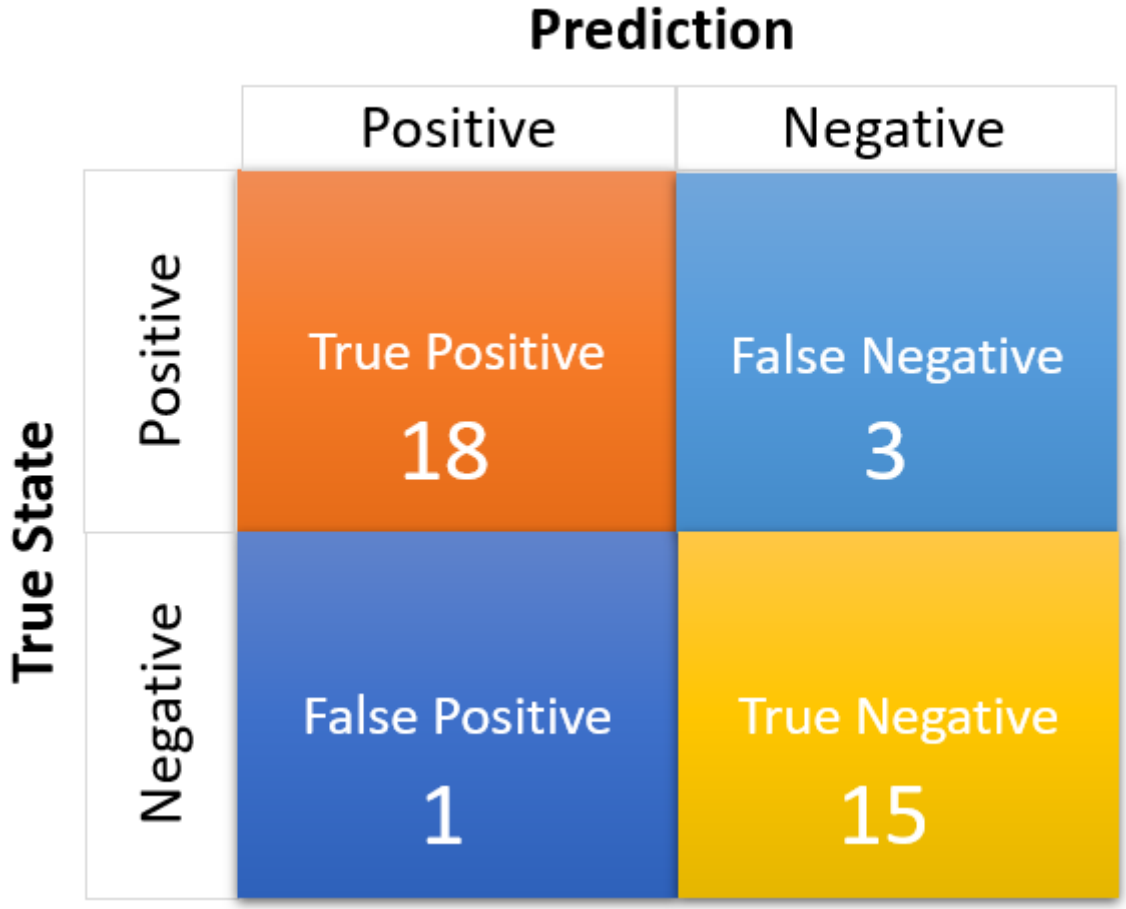


Classification Metrics

- Classification problems are one of the world's most widely researched areas.
- Use cases are present in almost all production and industrial environments. Speech recognition, face recognition, text classification – the list is endless.
- Classification models have discrete output, so we need a metric that compares discrete classes in some form.
- Classification Metrics evaluate a model's performance and tell you how good or bad the classification is, but each of them evaluates it in a different way.

Classification Metrics

IMPORTANT CONCEPTS

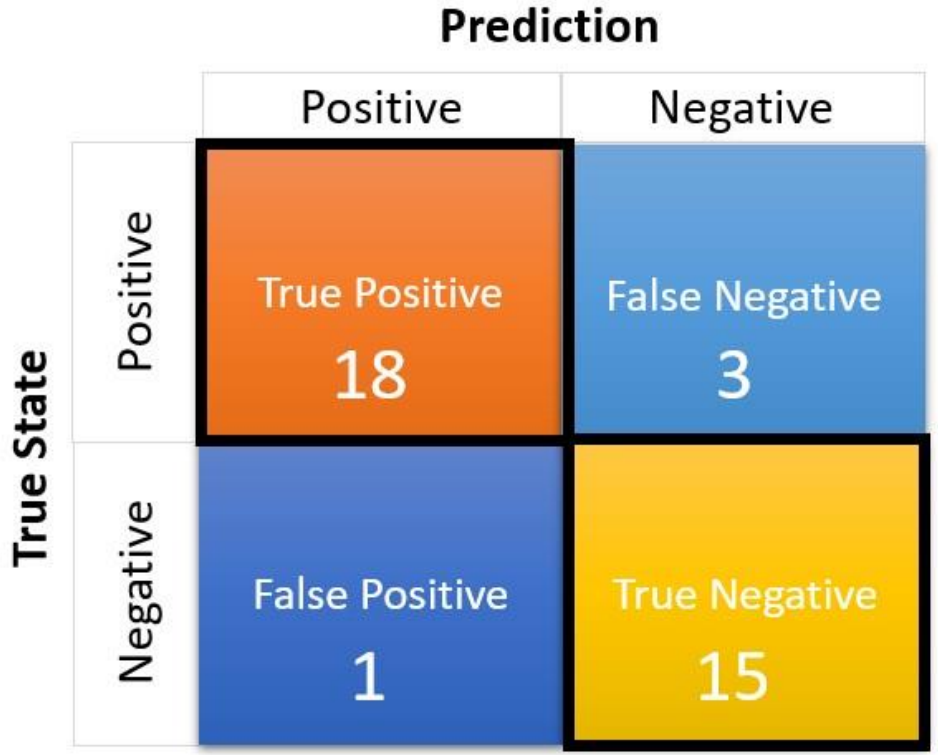


True Positive: Predicted 'Positive'
when the actual is 'Positive'
False Positive: Predicted 'Positive'
when the actual is 'Negative'
False Negative: Predicted 'Negative'
when the actual is 'Positive'
True Negative: Predicted 'Negative'
when the actual is 'Negative'

Classification Metrics

ACCURACY

- Accuracy is the proximity of measurement results to the true value. It tell us how accurate our classification model can predict the class labels given in the problem statement.



Accuracy*: The percent (ratio) of cases classified correctly

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

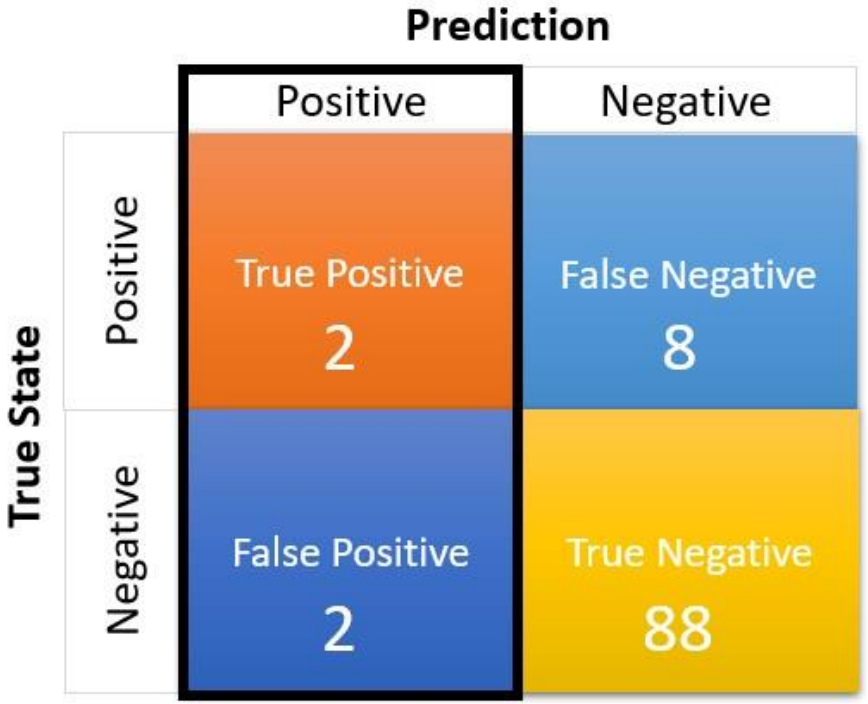
$$Accuracy = \frac{18 + 15}{18 + 1 + 3 + 15} = 0.89$$

$^{*}(bad)\ 0 \leq Accuracy \leq 1\ (good)$

Classification Metrics

PRECISION

- Precision is the ratio of true positives and total positives predicted



Precision*: Accuracy of a predicted positive outcome

$$Precision = \frac{TP}{TP + FP}$$

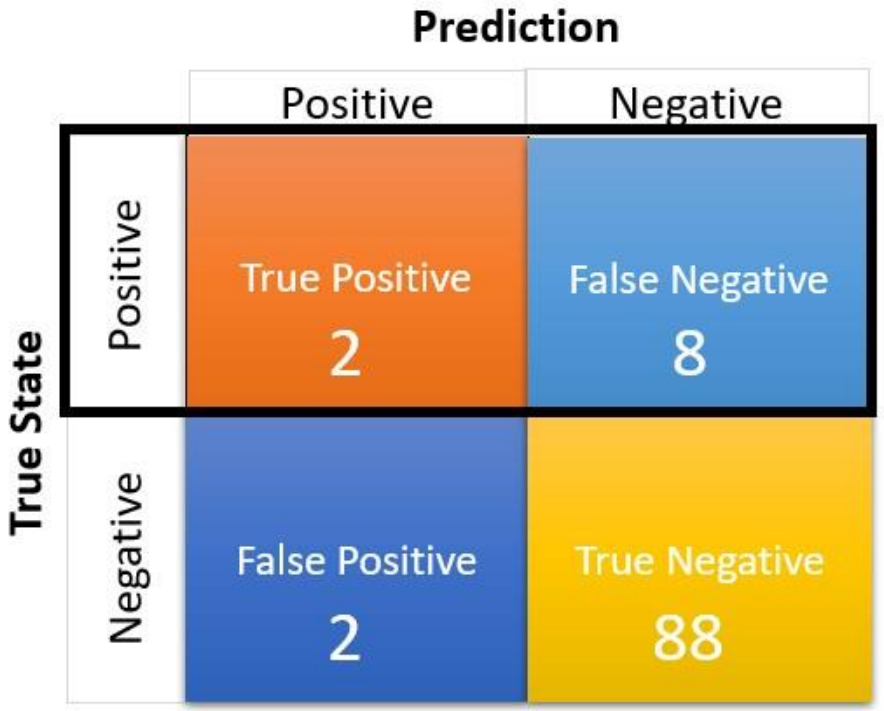
$$Precision = \frac{2}{2+2} = 0.50$$

$^{*}(bad)\ 0 \leq Precision \leq 1\ (good)$

Classification Metrics

RECALL/SENSITIVITY

- A Recall is essentially the ratio of true positives to all the positives in ground truth.



Recall*: Measures model's ability to predict a positive outcome

$$Recall = \frac{TP}{TP + FN}$$

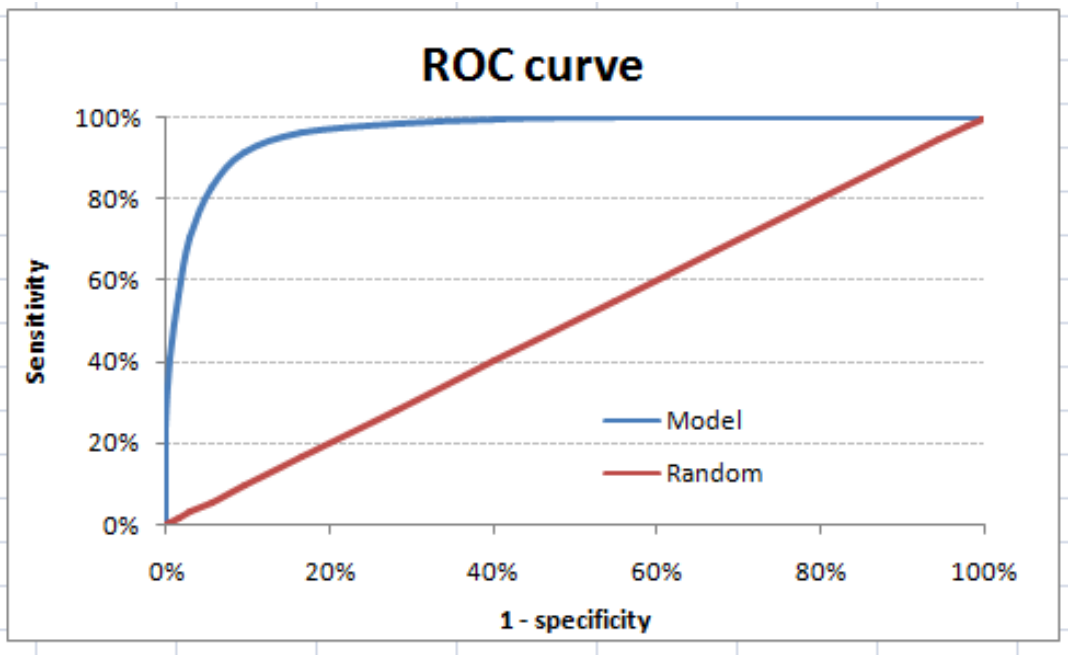
$$Recall = \frac{2}{2 + 8} = 0.20$$

$^{*}(bad)\ 0 \leq Recall \leq 1\ (good)$

Classification Metrics

AREA UNDER CURVE – RECEIVER OPERATOR CHARACTERISTICS (AUC – ROC)

- The ROC curve is the plot between sensitivity and (1- specificity). (1- specificity) is also known as false positive rate and sensitivity is also known as True Positive rate.
- Note that the area of entire square is 1*1 = 1. Hence AUC itself is the ratio under the curve and the total area.
- Following are a few thumb rules:
  - AUC value falls between:
    - .90 - 1 = excellent (A)
    - .80 - .90 = good (B)
    - .70 - .80 = fair (C)
    - .60 - .70 = poor (D)
    - .50 - .60 = fail (F)



Classification Metrics

F1 SCORE