

Machine Learning Models

Machine Learning Definition

Simply says Finds pattern in data and uses those patterns to predict the future.

It allows us to discover patterns in existing data and create and make use of a model that identifies those patterns in innovative data.

What does it mean to learn?

Learning Process: Identifying patterns

Recognizing those pattern when you see them again

Why is machine learning so popular currently?

- Plenty of data
- Lots of computer power
- An effective machine learning algorithm

Some examples for ML and AI to make our life easy like

- Google Search
- Intelligent Gaming
- Stock Predictions
- Robotics

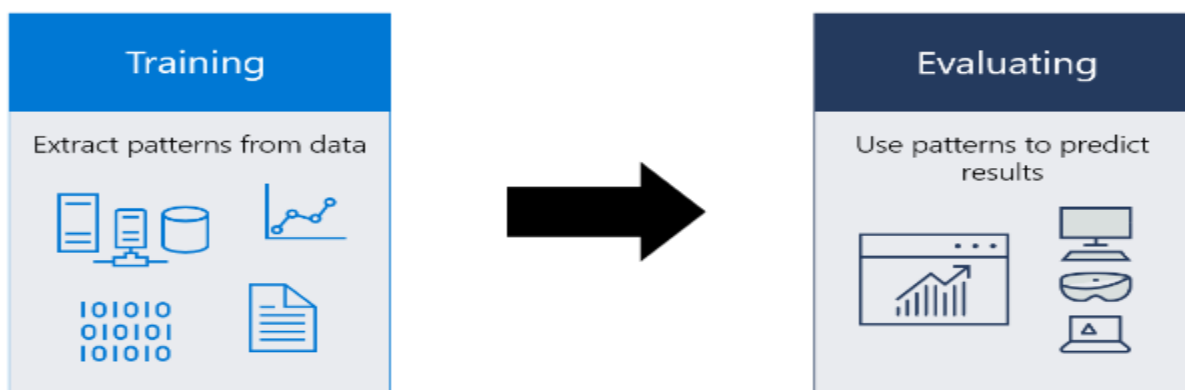
Top Machine Learning Companies

It is becoming an important part of our everyday life. It is really utilized in financial procedures, medical examinations, logistics, posting, and a variety of different fast-rising industries.

- Google – Neural Networks and machines
- Tesla – Autopilot
- Amazon – Echo Speaker Alexa
- Apple – Personalized Hey Siri
- TCS – Machine First Delivery Model with Robotics
- Facebook – Chatbot Army etc.

A machine learning model is a file that has been **trained to recognize certain types of patterns**. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data.

Once you have trained the model, you can use it to reason over data that it hasn't seen before, and make predictions about those data. For example, let's say you want to build an application that can recognize a **user's emotions based on their facial expressions**. You can train a model by providing it with images of faces that are each tagged with a certain emotion, and then you can use that model in an application that can recognize any user's emotion.



A machine learning model is the output of the training process and is defined as the mathematical representation of the real-world process. **The machine learning algorithms find the patterns in the training dataset, which is used to approximate the target function and is responsible for mapping the inputs to the outputs from the available dataset.**

When to use Machine Learning

Good machine learning scenarios often have the following common properties:

- They involve a repeated decision or evaluation which you want to automate and need consistent results.
- It is difficult or impossible to explicitly describe the solution or criteria behind a decision.
- You have labeled data, or existing examples where you can describe the situation and map it to the correct result.

All machine learning models are categorized as either supervised or unsupervised. If the model is a supervised model, it's then sub-categorized as either a regression or classification model.

Supervised Machine Learning

Supervised Machine Learning is defined as the subfield of machine learning techniques in which we used labelled datasets for training the model, making predictions of the output values and comparing its output with the intended, correct output, and then compute the errors to modify the model accordingly. Also, as the system is trained enough using this learning method, it becomes capable enough to provide the target values from any new input.

Think about **Gmail's spam recognition system**. Now there, it will take under consideration a collection of emails (a huge number, just like millions) which have recently been categorized because of **spam or not spam**, from this level, with the ability to identify what features an email that is spam or not spam display. Once gaining knowledge of this, with the ability to classify onset e-mails as spam or otherwise.

When we train the algorithm by providing the labels explicitly, it is known as supervised learning. This type of algorithm uses the available dataset to train the model. The model is of the following form.

$Y=f(X)$ where x is the input variable, y is the output variable, and $f(X)$ is the hypothesis.

The objective of Supervised Machine Learning Algorithms is to find the hypothesis as approx. as possible so that when there is new input data, the output y can be predicted. The application of supervised machine learning is to predict whether a mail is spam or not spam or face unlock in your smartphone.

Types of Supervised Machine Learning Algorithm

Supervised Machine Learning is divided into two parts based upon their output:

- 1. Regression**
- 2. Classification**

Regression

In Regression the output variable is numerical (continuous) i.e. we train the hypothesis($f(x)$) in a way to get continuous output(y) for the input data(x). Since the output is informed of the real number, the regression technique is used in the prediction of quantities, size, values, etc.

For Example, we can use it to predict the price of the house given the dataset containing the features of the house like area, floor, etc.

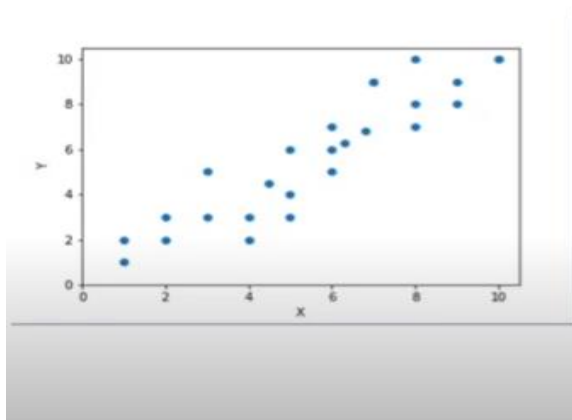
In regression problems the output values of target class are continuous. Target value must be **integer or float**. In this example target value of house price is an integer value depends on bed rooms in house, floors, condition and location. So, our model will predict an integer value of house price by using these features.

bathrooms	floors	condition	city	price
1.5	3	3	Shoreline	313000
2.5	2	5	Seattle	2384000
2	1	4	Kent	342000
2.25	1	4	Bellevue	420000
2.5	1	4	Redmond	550000
1	1	3	Seattle	490000
2	1	3	Redmond	335000
2.5	2	3	Maple Valley	482000
2.5	1	4	North Bend	452500
2	3	3	Seattle	640000
1.75	1	3	Lake Forest Park	463000
2.5	3	5	Seattle	1400000
1.75	1	3	Sammamish	588500
1	1	4	Seattle	365000



Int

For regression problem suppose you have only one feature x and target value y and you are going to plot in graph for x equal to 1 y also 1, for 3 y equal to 5 and that's going on. Finally, our data will look like this. We can draw a line or curve for predicting new data.



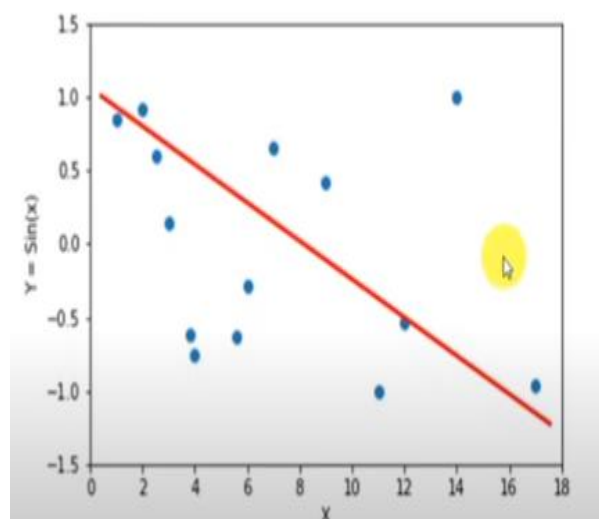
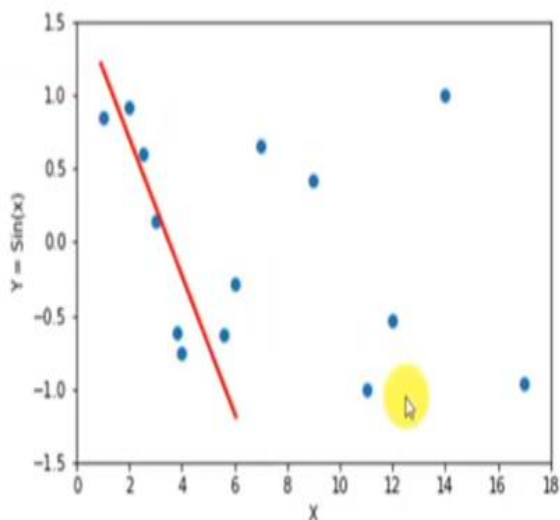
X	Y	X	Y
1	1	6	7
3	5	7	9
2	2	8	10
4	2	9	9
5	3	10	10
6	6	4.5	4.5
7	9	8	8
8	7	6.3	6.3
9	8	6	5
10	10	5	4
1	2	4.7	4.7
2	3		
3	3		
4	3		
5	6		

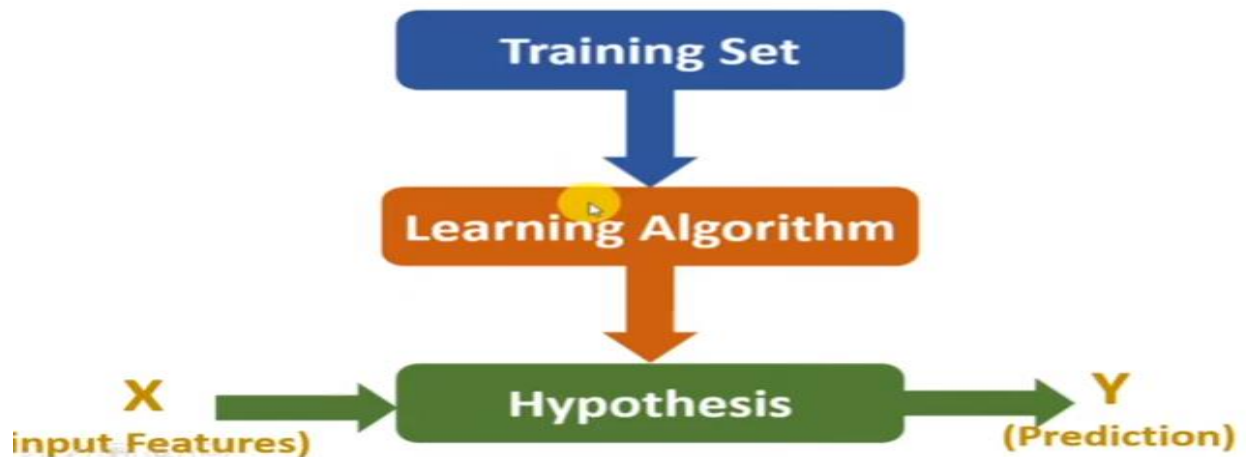
A few popular Regression Algorithm is:

- Linear Regression
- Support Vector Regression
- Poisson Regression

Linear regression is used for finding **linear relationship between target and one or more predictors**. There are two types of linear regression- Simple and Multiple.

The core idea is to obtain a line that **best fits** the data. The **best fit line** is the one for which total prediction error (all data points) are as small as possible. Error is the **distance between the point to the regression line**.





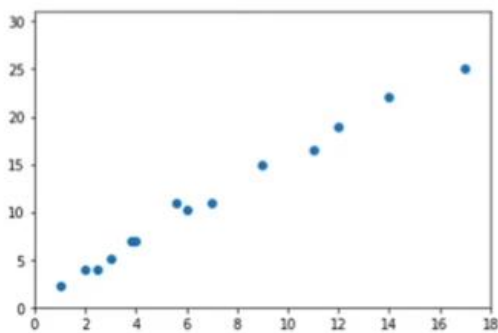
Hypothesis:

$$h_{\theta}(x) = \theta_0 + \theta_1 x \quad \text{Predicted Y}$$

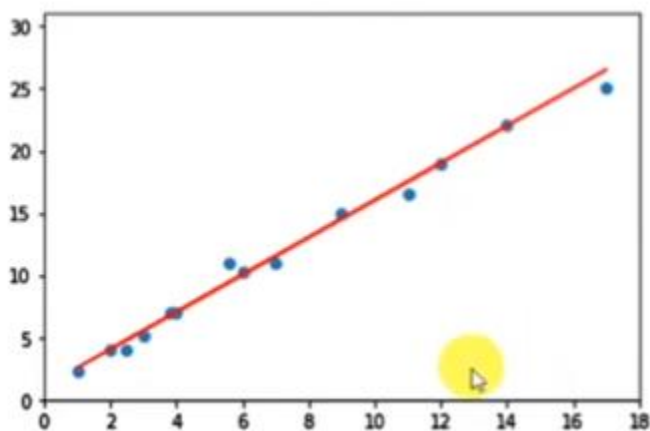
$$\theta_0 = 1, \theta_1 = 1.5$$

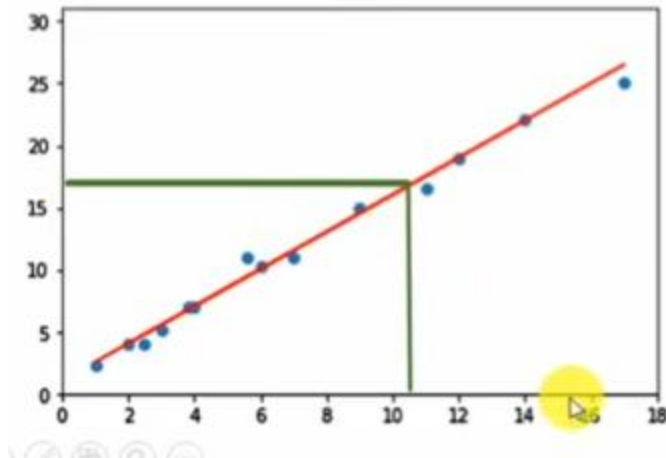
$$h_{\theta}(9) = 1 + 1.5 \times 9$$

$$h_{\theta}(9) = 14.5$$



X	Y	Pred_Y
7	11	11.5
2	4	4
17	25	26.5
9	15	
4	7	
11	16.5	
12	19	
6	10.2	
1	2.3	
3	5.1	
2.5	4	
3.8	7	
5.6	11	
14	22	





Optimizing using gradient descent:

In linear regression, the model targets to get the best-fit regression line to predict the value of y based on the given input value (x). While training the model, the model calculates the cost function which measures the Root Mean Squared error between the predicted value (pred) and true value (y). **The model targets to minimize the cost function.**

To minimize the cost function, the model needs to have the best value of θ_1 and θ_2 . Initially model selects θ_1 and θ_2 values randomly and then iteratively update these value in order to minimize the cost function until it reaches the minimum. By the time model achieves the minimum cost function, it will have the best θ_1 and θ_2 values. Using these finally updated values of θ_1 and θ_2 in the hypothesis equation of linear equation, model predicts the value of x in the best manner it can.

Cost Function:

Minimize $(h_{\theta}(x) - y)^2$:

minimize the difference between $h(x)$ and y for each/any/every example

$$\frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 \quad m = \text{number of iteration} \\ i = 1, 2, 3, 4, \dots, m$$

$$\frac{1}{2 \times 14} \{ (11.5 - 11)^2 + (4 - 4)^2 + (26.5 - 25)^2 + (14.5 - 15)^2 + \dots + (22 - 22)^2 \}$$

$$\frac{1}{28} (0.25 + 0 + 2.25 + 0.25 + 0 + 1 + 0 + 0.04 + 0.04 + 0.16 + 0.5625 + 0.09 + 2.56 + 0)$$

$$\frac{1}{28} \times 7.2025$$

$$\text{MSE} = 0.2572$$

$$\theta_0 = 1, \theta_1 = 1.5$$

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

X	Y	Pred_Y
7	11	11.5
2	4	4
17	25	26.5
9	15	14.5
4	7	7
11	16.5	17.5
12	19	19
6	10.2	10
1	2.3	2.5
3	5.1	5.5
2.5	4	4.75
3.8	7	6.7
5.6	11	9.4
14	22	22

Gradient Decent

$$\frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 = J(\theta)$$

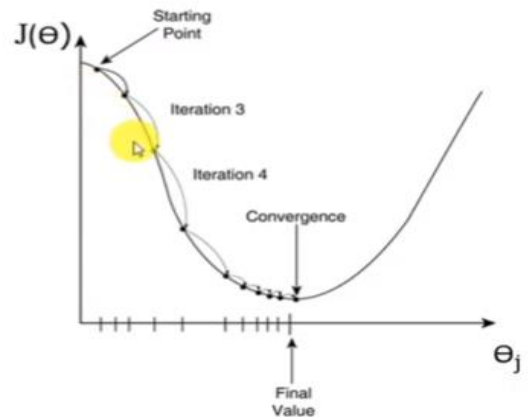
$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \quad \alpha = \text{Learning Rate}$$

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} \times \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

$$\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})$$

$$\theta_0 = \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})$$

$$\theta_1 = \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) \times x^{(i)}$$



Example:

1st iteration:

$$\theta_0 = 1, \theta_1 = 1$$

$$\therefore h_{\theta}(x) = \theta_0 + \theta_1 x$$

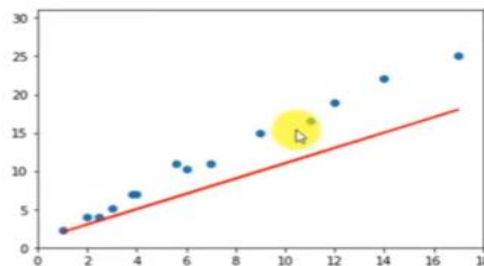
$$h_{\theta}(7) = \theta_0 + \theta_1 x = 1 + 1 \times 7 = 8$$

$$h_{\theta}(2) = \theta_0 + \theta_1 x = 1 + 1 \times 2 = 3$$

$$h_{\theta}(17) = \theta_0 + \theta_1 x = 1 + 1 \times 17 = 18$$

⋮
⋮
⋮
⋮
⋮

$$h_{\theta}(14) = \theta_0 + \theta_1 x = 1 + 1 \times 14 = 15$$



X	Y	Pred_Y
7	11	8
2	4	3
17	25	18
9	15	10
4	7	5
11	16.5	12
12	19	13
6	10.2	7
1	2.3	2
3	5.1	4
2.5	4	3.5
3.8	7	4.8
5.6	11	6.6
14	22	15

⏪ ⏩ 🔍 🔄 📄 ⌂

1st iteration: $\theta_0 = 1, \theta_1 = 1$

$$\theta_0 = \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})$$

$$\frac{1}{14} \left\{ (8 - 11) + (3 - 4) + (18 - 25) + (10 - 15) + (5 - 7) + (12 - 16.5) + (13 - 19) + (7 - 10.2) \right. \\ \left. + (2 - 2.3) + (4 - 5.1) + (3.5 - 4) + (4.8 - 7) + (6.6 - 11) + (15 - 22) \right\}$$

$$\frac{1}{14} (-3 + -1 + -7 + -5 + -2 + -4.5 + -6 + -3.2 + -0.3 + -1.1 + -0.5 + -2.2 + -4.4 + -7)$$

$$\frac{1}{14} \times -47.2 = -3.3714$$

$$\theta_0 = \theta_0 - \alpha \times (-3.3714)$$

$$\theta_0 = 1 - 0.01 \times (-3.3714)$$

$$\theta_0 = 1.0337$$

1st iteration: $\theta_0=1, \theta_1=1$

$$\theta_1 = \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) \times x^{(i)}$$

$$\frac{1}{14} \{ (8-11) \times 7 + (3-4) \times 2 + (18-25) \times 17 + (10-15) \times 9 + (5-7) \times 4 + (12-16.5) \times 11$$

$$+ (13-19) \times 12 + (7-10.2) \times 6 + (2-2.3) \times 1 + (4-5.1) \times 3 + (3.5-4) \times 2.5 + (4.8-7) \times 3.8$$

$$+ (6.6-11) \times 5.6 + (15-22) \times 14 \}$$

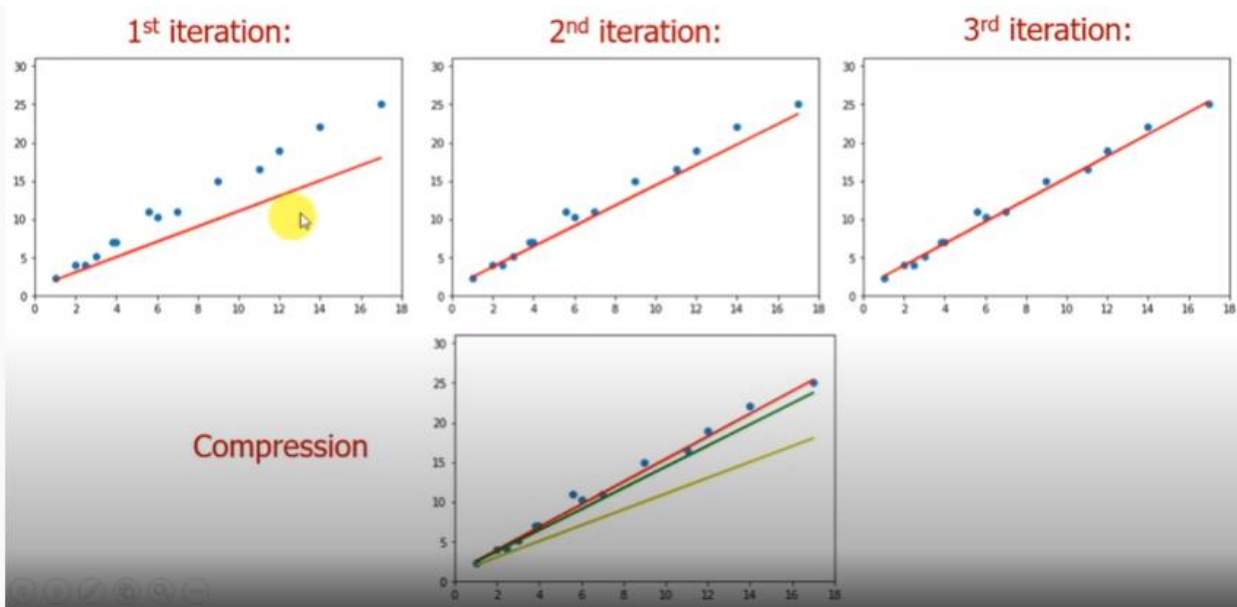
$$\frac{1}{14} (-21 + -2 + -119 + -45 + -8 + -49.5 + -72 + -19.2 + -0.3 + -3.3 + -1.25 + -8.36 + -24.64 + -98.)$$

$$\frac{1}{14} \times -471.55 = -33.6821$$

$$\theta_1 = \theta_1 - \alpha \times (-33.6821)$$

$$\theta_1 = 1 - 0.01 \times (-33.6821)$$

$$\theta_1 = 1.336$$



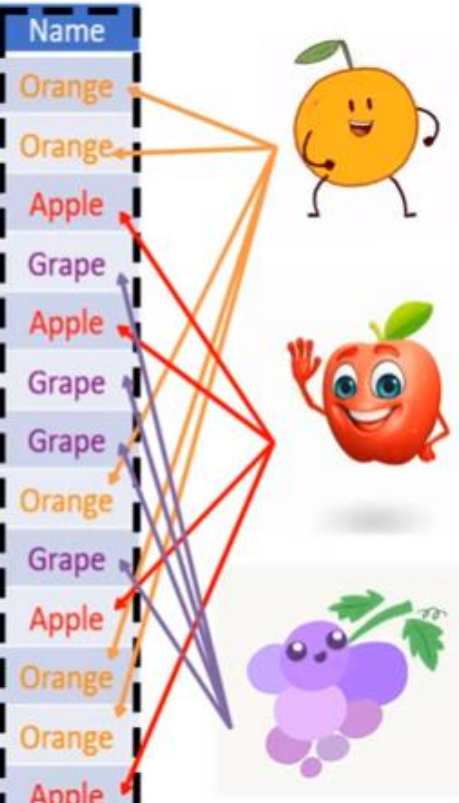
Classification

In classification, the output variable is discrete. i.e. we train the hypothesis($f(x)$) in a way to get discrete output(y) for the input data(x). The output can also be termed as a class. For example, by taking the above example of house price, we can use classification to predict whether the house price will be above or below **instead of getting the exact value**. So we have two classes, one if the **price is above** and the other if **it is below**.

In classification problems the output values of target class are discrete. Target value must be belonging to a class. In this example target values are a fruit name

which depends on color, diameter, and weight. In classification we classify our data into classes and predict class for new data.

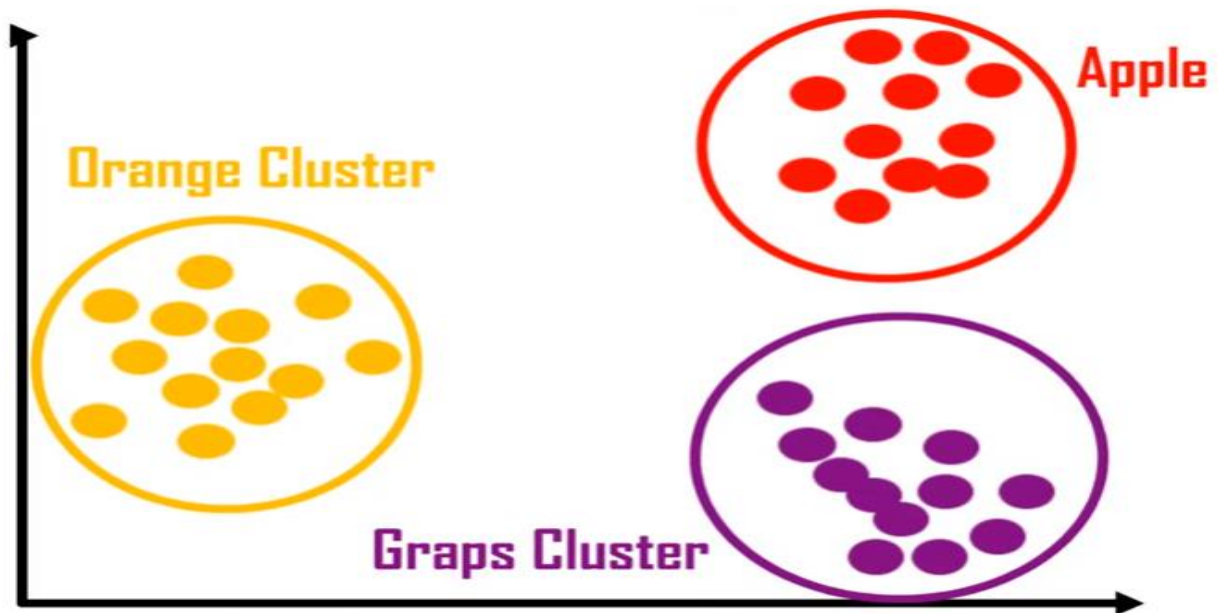
Diameter	Weight	Red	Green	Blue	Name
2.96	86.76	172	85	2	Orange
3.91	88.05	166	78	3	Orange
5.43	108.54	157	98	2	Apple
5.51	109.49	150	98	5	Grape
11.06	191.08	151	57	6	Apple
11.06	191.08	151	57	6	Grape
11.06	191.08	151	57	6	Grape
11.06	191.08	151	57	6	Orange
13.17	223.49	162	79	13	Grape
13.17	223.51	163	74	23	Apple
13.17	223.52	140	66	22	Orange
13.17	223.55	165	75	26	Orange
13.17	223.56	125	69	24	Apple



The diagram illustrates the classification process. Arrows connect specific rows in the 'Name' column to corresponding fruit images:

- Orange (Row 1) and Orange (Row 2) point to an orange fruit image.
- Apple (Row 3) and Apple (Row 5) point to an apple fruit image.
- Grape (Row 4), Grape (Row 6), Grape (Row 7), and Grape (Row 9) point to a grape fruit image.
- Orange (Row 8), Orange (Row 12), and Orange (Row 13) point to an orange fruit image.
- Apple (Row 11) and Apple (Row 14) point to an apple fruit image.

In classification problem we will separate our data and make classes and our data should look like this. In this example based on values of x we make clusters for fruits target classes could be 2 or more than 2.



Classification is used in speech recognition, image classification, NLP, etc.

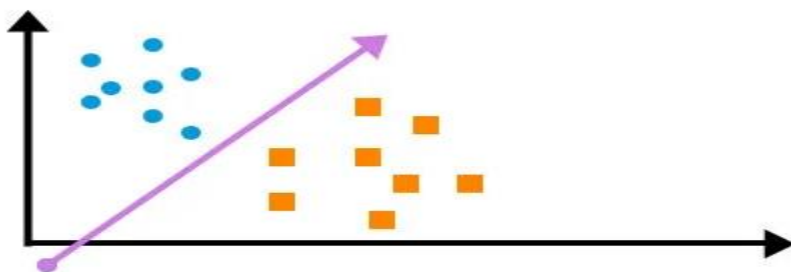
Few Popular Classification Algorithm is:

- Logistic Regression
- Neural Network
- Decision Tree
- Naïve Bayes Classifier

The classification of supervised learning algorithms is used to group similar objects into unique classes.

- **Binary classification** – If the algorithm is trying to group 2 distinct groups of classes, then it is called binary classification.
- **Multiclass classification** – If the algorithm is trying to group objects to more than 2 groups, then it is called multiclass classification.
- **Strength** – Classification algorithms usually perform very well.
- **Drawbacks** – Prone to overfitting and might be unconstrained. For Example – Email Spam classifier

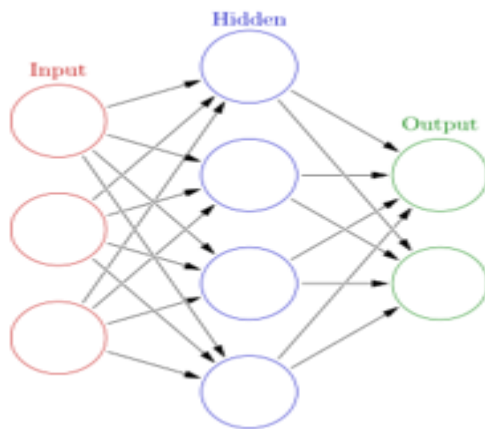
- **Logistic regression/classification** – When the Y variable is a binary categorical (i.e. 0 or 1), we use Logistic regression for the prediction. For Example – Predicting if a given credit card transaction is fraud or not.
- **Naïve Bayes Classifiers** – The Naïve Bayes classifier is based on the Bayesian theorem. This algorithm is usually best suited when the dimensionality of the inputs is high. It consists of acyclic graphs that are having one parent and many children nodes. The child nodes are independent of each other.
- **Decision Trees** – A decision tree is a tree chart like structure that consists of an internal node (test on attribute), a branch that denotes the outcome of the test and the leaf nodes, representing the distribution of classes. The root node is the topmost node. It is a very widely used technique which is used for classification.
- **Support Vector Machine** – In the support vector machine, we use a hyperplane to classify the dependent variable when we have only two dependent variables, i.e. only two classes to predict; then, this hyperplane is nothing but a straight line. The objective of SVM is to get the hyperplane in a way that all the independent variables of one class should be on one side. An optimal SVM function will result in a hyperplane that is at an equal distance from both the class. Fields where SVMs are extensively used, are biometrics, pattern recognition, etc.



Neural Network

The neural network is a classification algorithm that has a minimum of 3 layers. Input – Hidden – Output.

The number of hidden layers may vary based upon the application of the problem. Each hidden layer tries to detect a pattern on the input. As the pattern is detected, it gets forwarded to the other hidden layer in the network till the output layer. Please see the figure below.



The circle in the figure determines the neuron which stores the features (only in the input layer), i.e. the independent variable.

Suppose we have a 32×32 image of a number, then in order to classify the number, we can use a neural network. We will pass the images to the input layers. Since the number of pixels in the image is $32 \times 32 = 1024$, we will have 1024 neurons in the input layers. The number of neurons in hidden layers can be tweaked, but for the output layer, it has to be 10 since, in this example, the number can be anything between 0-9.

The node of the output layers contains the probability. Whichever node has the highest probability that the node is supposed to be the resultant class?

Unsupervised machine learning

Unsupervised Machine Learning is one of the three main techniques of machine learning. It's a self-organized learning algorithm in which we don't need to supervise the data by providing a labelled dataset as it can find a previously unknown pattern in the unlabelled dataset on its own to discover useful

information by performing complex tasks (such as principal component analysis and cluster analysis) as compared to the other machine learning techniques like supervised learning.

Unsupervised learning simply works with the input data. It's essentially ideal for the incoming data going to enable it to be more understandable and organized. Mainly, it studies the input data to discover behavior or commonalities or flaws to your prospects. Possibly considered how Amazon or any type of other online stores can recommend many you can purchase?

This really is because of unsupervised machine learning. Web sites like these consider the prior acquisitions, and they are capable of recommending other activities that you might be thinking about too.

Types of Unsupervised Machine Learning

Unsupervised learning tasks can be broadly divided into 3 categories:

1. Clustering
2. Association rule mining
3. Recommendation system

Clustering

Clustering can be done any data where we do not have the class or label information. We want to group the data such that the observations with similar properties belong to the same cluster/group, and inter-cluster distance should be maximum. At the same time, the intra-cluster distance should be minimum. We can cluster the voter's data to determine the opinion about the government or cluster products based on their features and usage. Segment population based on income features or use clustering in sales and marketing.

We can use K-Means, K-Means++, K-Medoids, Fuzzy C-means (FCM),

Expectation-Maximisation (EM), Agglomerative Clustering, DBSCAN, Hierarchical Clustering types as single linkage, complete linkage, median linkage, Ward's method algorithms for clustering.

Association Rule Mining

When we have transactional data for something, it can be for products sold or any transactional data for that matters; I want to know, is there any hidden relationship between buyer and the products or product to product, such that I can somehow leverage this information to increase my sales. Extracting these relationships is the core of Association Rule Mining. We can use the AIS, SETM, Apriori, FP growth algorithms for extracting relationships.

Recommendation System

Recommendation System is basically an extension of Association rule mining in a sense; we are extracting relationships in ARM. In the Recommendation System, we are using these relationships to recommend something which is having higher acceptance chances by the end-user. Recommendation systems have gained popularity after Netflix announced a grand prize of US\$1,000,000 prize in 2009.

Recommendation Systems works on transactional data, be it financial transaction, e-commerce, or grocery shop transactions. Nowadays, giant players in the e-commerce industry are luring customers by making a customized recommendations for each user based on their past purchase history and similar behaviour purchase data from other users.

Methods to develop Recommendation Systems can be broadly divided into Collaborative filtering and Content-Based filtering. In Collaborative filtering, we have user-user Collaborative filtering and Item-Item Collaborative filtering, which are memory-based approaches & Matrix factorization and Singular Value Decomposition (SVD) model-based approaches.

Reinforcement Learning

Reinforcement Learning enables systems to understand depending on previous benefits for its activities. Whenever a system requires a resolution, it can be penalized or honored for its activities. For every action, it should get good feedback, which this discovers if this worked an incorrect or corrective action. This kind of machine learning is usually purely focused on the boosted effectiveness of the function.

