UNIVERSITY OF WESTMINSTER

Machine Learning

6COSC020W - Lecture Week 07

Dr. Hamed Hamzeh

Outline

UNIVERSITY OF WESTMINSTER

- What is Machine Learning
  - Machine Learning Concepts
- Types of Machine Learning
  - Supervised Learning
    - Linear Regression
    - Logistic regression
    - Naive Bayes Classifier
    - Bayesian networks
    - Decision trees
    - Random Forests
    - Gradient Boosting
    - KNN
    - Support Vector Machine
    - Neural Networks
  - Unsupervised Learning
    - K-means Clustering
    - Hierarchical Clustering
    - Gaussian Mixture Model
    - DBSCAN
  - Reinforcement Learning
- Building Machine Learning Architecture

What is the Machine Learning?

What is the Machine Learning?

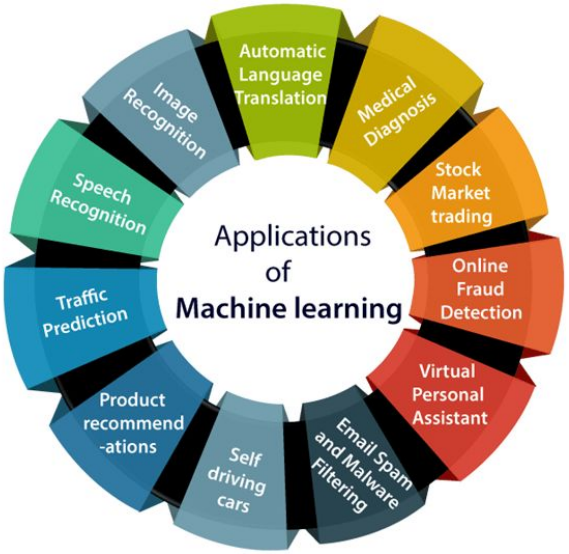
UNIVERSITY OF WESTMINSTER

- Machine Learning is a subset of **artificial intelligence** that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own.

Arthur Samuel:

(1959)

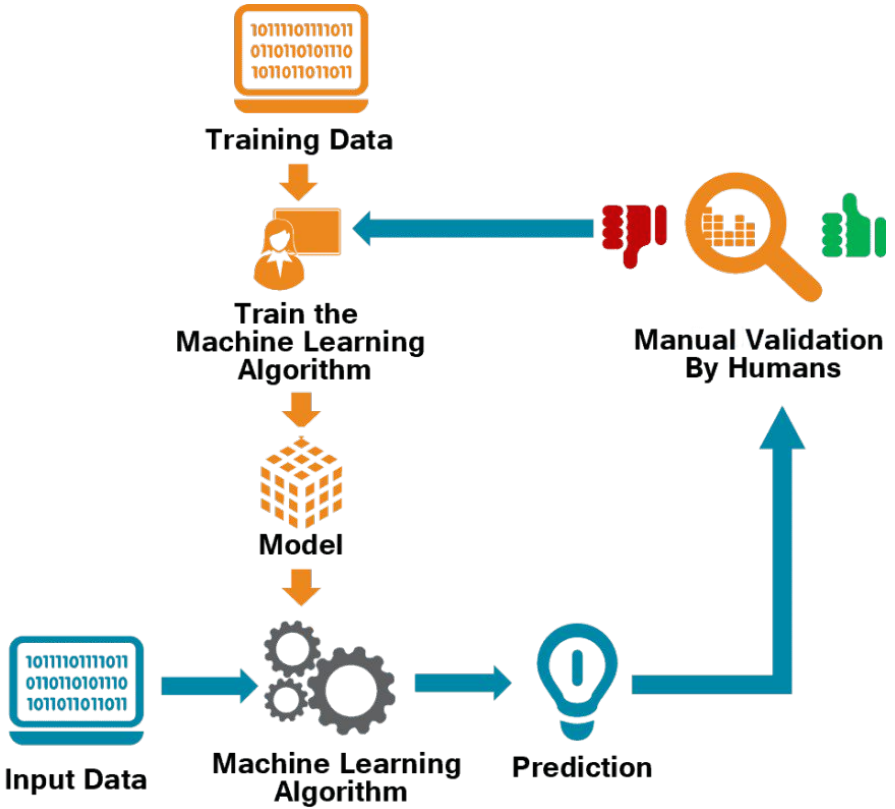
> "Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed."



How Does Machine Learning Work?

UNIVERSITY OF WESTMINSTER

- A Machine Learning system **learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it**.
- The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.



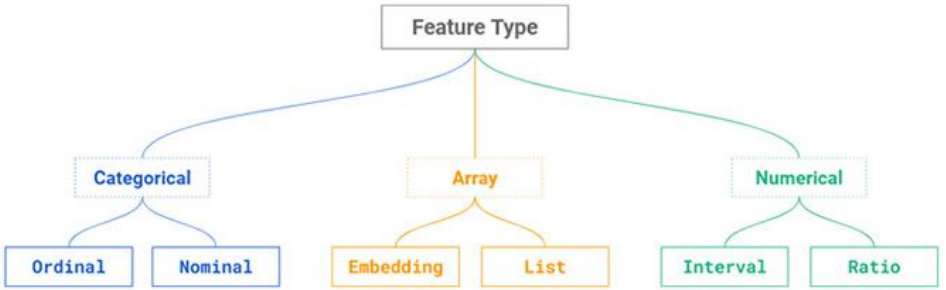
Machine Learning Concepts

UNIVERSITY OF WESTMINSTER

Feature: is an individual independent variable that acts as input in the system. While making the predictions, models use such features to make the predictions.

	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr. (	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, Mr	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, M	female	26	0	0	STON/O2. 31	7.925		S
5	4	1	1	Futrelle, Mrs	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr. Wi	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr. J	male		0	0	330877	8.4583		Q

- **Numerical Variables:** Variable with continuous values such as integer, float
- **Categorical Variables:** Variables with categorical values such as Boolean, ordinal, nominals.

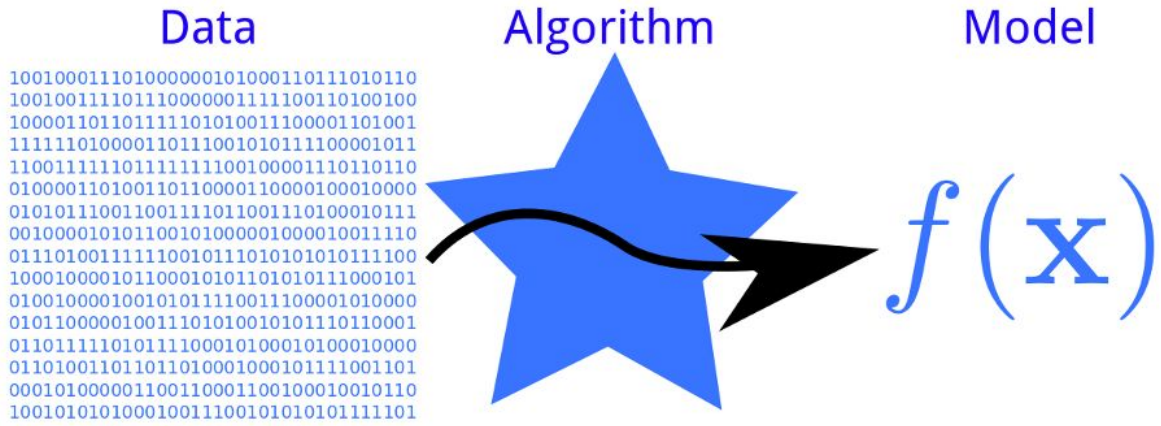


Machine Learning Concepts

UNIVERSITY OF WESTMINSTER

Algorithm: Refers to using any kind of machine learning methods to train the data.

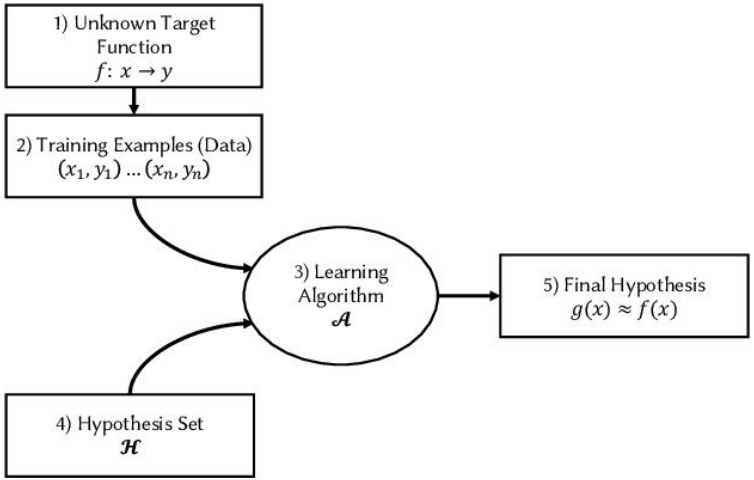
Model: Machine Learning models can be understood as a program/algorithm that has been trained to find patterns within new data and make predictions. These models are represented as a mathematical function that takes requests in the form of input data, makes predictions on input data, and then provides an output in response.

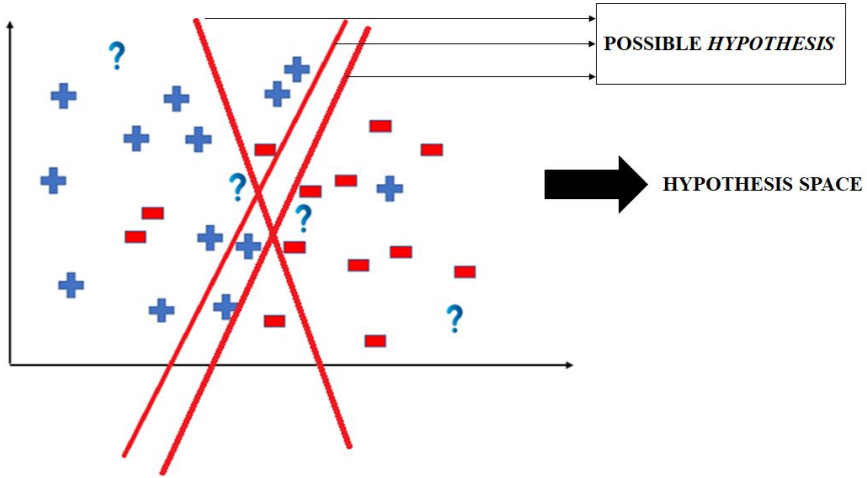


Machine Learning Concepts

UNIVERSITY OF WESTMINSTER

- In order to do machine learning, there should exist a relationship (pattern) between the input and output values. Let's say that the function, *y=f(x)* which is known as the **target function.**
- However, *f(.)* is unknown function. So machine learning algorithms try to guess a “hypothesis” function h(x) that approximates the unknown *f(.)*. The set of all possible hypotheses is known as the Hypothesis set H(.) The goal is the learning process is to find the final hypothesis that best approximates the unknown target function.
- **Learning algorithm**: Based on the input dataset, Learning algorithm learns the most optimal value of model parameters.



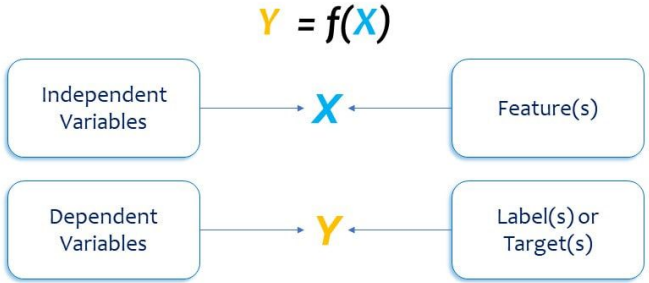


Machine Learning Concepts

UNIVERSITY OF WESTMINSTER

Independent Variables: The variable that are not affected by the other variables are called independent variables. For example age of a person, is an independent variable, two persons' born on same date will have same age irrespective of how they lived.

Dependent Variable: The variables which depend on other variables. We expect these variables to change when the independent variables, upon whom they depend, undergo a change. For example you have a test tomorrow, then, your test score is dependent upon the amount of time you studied, so the test score is a dependent variable, and amount of time independent variable in this case.

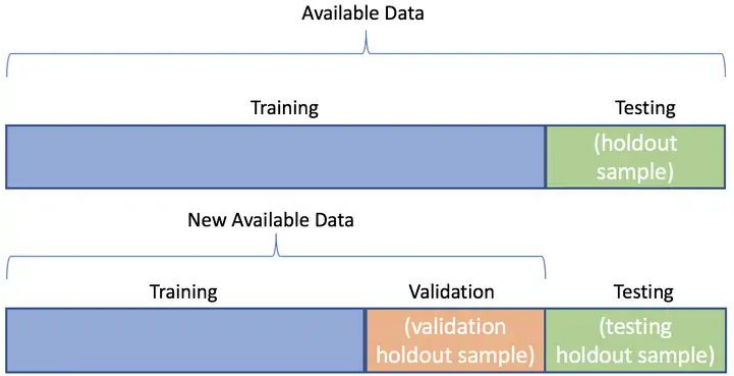


Height	Diameter	Altitude	Rainfall	Age
50	8	5000	12	80
56	9	4400	10	75
72	12	6500	18	60
47	10	5200	14	53

Machine Learning Concepts

UNIVERSITY OF WESTMINSTER

- **Train and Test data:**
  - **Training data:** Training data is typically larger than testing data. This is because we want to feed the model with as much data as possible to find and learn meaningful patterns.
  - **Testing data:** Once your machine learning model is built (with your training data), you need unseen data to test your model. This data is called testing data, and you can use it to evaluate the performance and progress of your algorithms' training and adjust or optimize it for improved results.



Machine Learning Concepts

UNIVERSITY OF WESTMINSTER

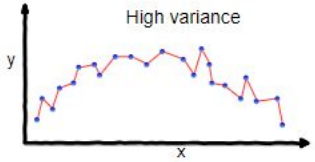
- **Variance:**
  - Variance is the amount that the prediction will change if different training data sets were used.
  - If the difference is high then it's called high variance and when the difference of errors is low then it's called low variance. Usually, we want to make a low variance for generalized our model.
- **Bias:**
  - While making predictions, a difference occurs between prediction values made by the model and actual values/expected values, and this difference is known as bias errors or Errors due to bias.
  - In other words, Bias is the simplifying assumptions made by a model to make the target function easier to learn.
  - A low bias model will make fewer assumptions about the form of the target function.
  - A model with a high bias makes more assumptions, and the model becomes unable to capture the important features of our dataset. A high bias model also cannot perform well on new data.

Machine Learning Concepts

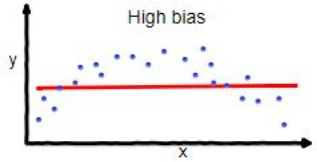
UNIVERSITY OF WESTMINSTER

Overfitting:

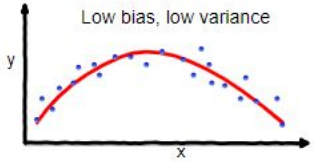
- A statistical model is said to be overfitted when the model does not make accurate predictions on testing data.
- When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set.There are some techniques to reduce overfitting:
  - Increase training data
  - Reduce model complexity
  - Feature Selection



overfitting



underfitting



Good balance

Underfitting:

- Underfitting occurs when we have a high bias and low variance in our data and it only performs well on training data but performs poorly on testing data.
- Techniques to overcome underfitting:
  - Increase model complexity
  - Increase the number of features, performing feature engineering
  - Remove noise from the data.

Machine Learning Concepts

UNIVERSITY OF WESTMINSTER

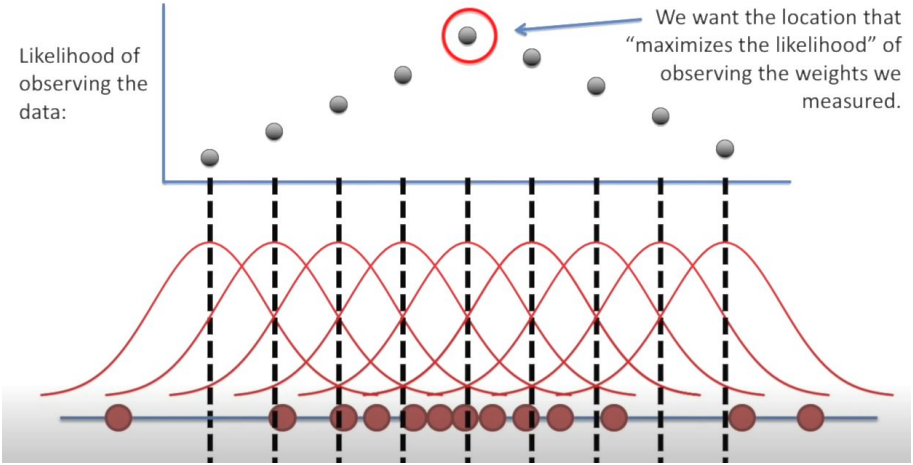
- **Hyper-parameters**
  - There are hidden elements that affect the machine learning optimization which is outside the model and have great influence on model behaviour. These hidden elements are referred to as hyperparameters, these are critical components for the optimization of any machine learning model.
- **Latent variables**
  - Latent variables/hidden are variables that can only be inferred indirectly through a mathematical model from other observable variables that can be directly observed or measured.
- **Inference**
  - Inference refers to the ability of a learning system, namely going from the "particular" (the examples) to the "general" (the predictive model). Inference would be performed in a single step:
    - we input training examples into a big black box containing all models, hyper-parameters, and parameters; outcomes the best possible trained model.
  - In practice, we often use two levels of inference:
    - we split the training data into a training set and a validation set. The training set serves the trains at the lower level (adjust the parameters of each model). The validation set serves to train at the higher level (select the model.)

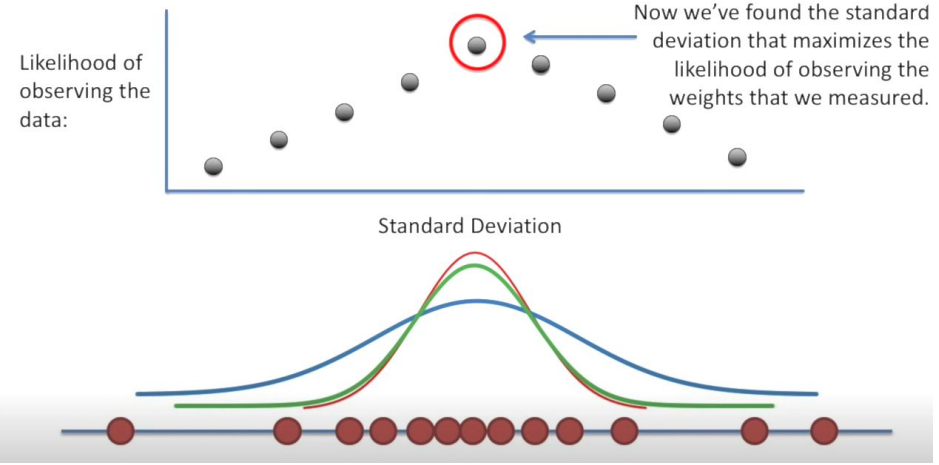
Machine Learning Concepts

UNIVERSITY OF WESTMINSTER

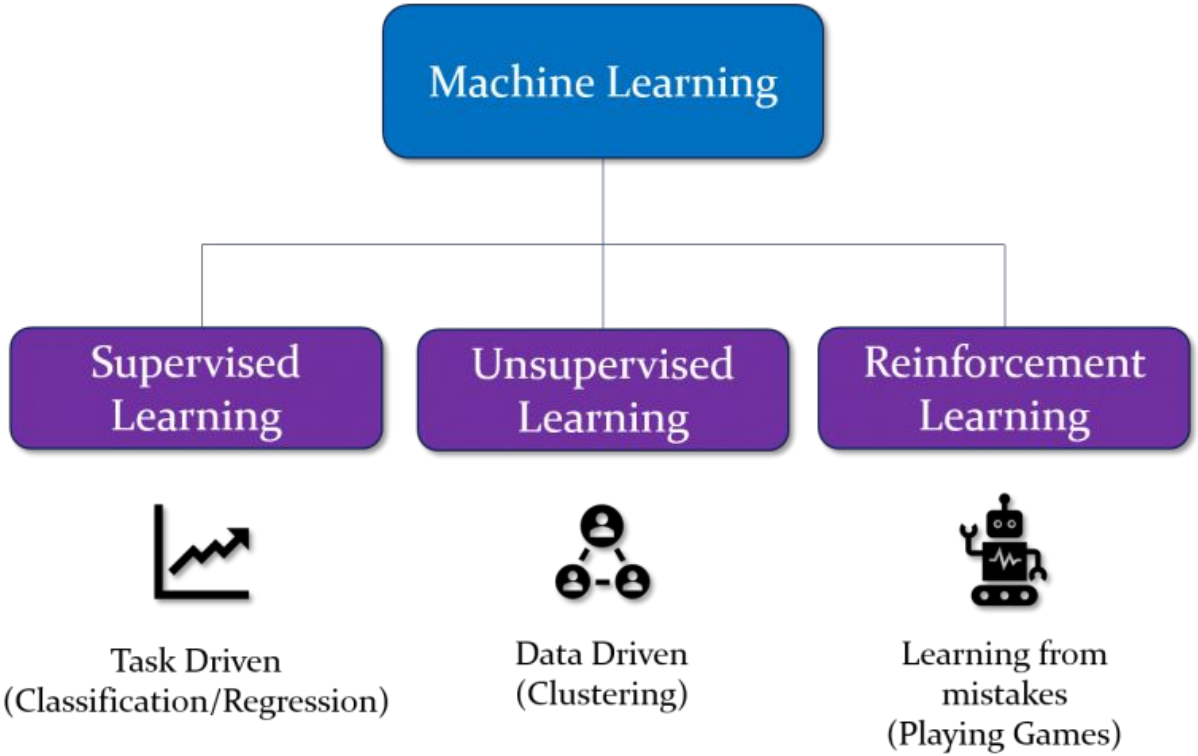
Maximum Likelihood and the Expectation Maximization:

- Maximum likelihood is a method of estimating model parameters which maximize the likelihood that the model generates the observed data that are around the mean and standard deviation with respect to the normal distribution.
- The goal of maximum likelihood is to find the optimal way to fit a distribution to the data.
- The Maximum Likelihood Estimation (MLE) algorithm is an iterative method of computing maximum likelihood estimations for probabilistic models with latent variables that are inferred from observed data.





Types of Machine Learning



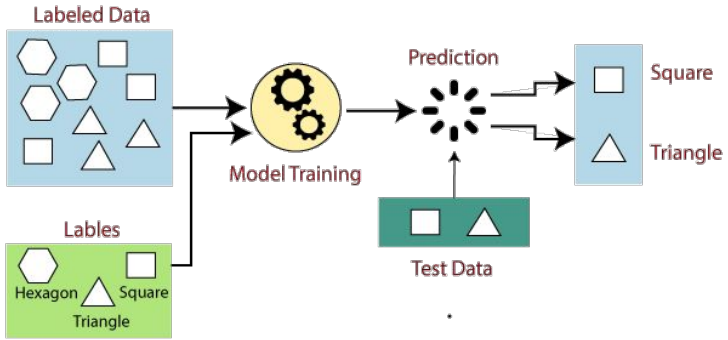
Supervised Learning Algorithms

Supervised Learning

UNIVERSITY OF WESTMINSTER

- Supervised learning is a type of machine learning method in which we provide sample **labeled data** to the machine learning system in order to train it, and on that basis, it predicts the output.
- The goal of supervised learning is to map input data with the output data. The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher. The example of supervised learning is **spam filtering**.
- Supervised Learning is categorized into **Classification** and **Regression**.
- The procedure of Supervised Learning can be described as the follows: we use $x(i)$ to denote the input variables, and y(i) to denote the output variable. A pair $(x(i), y(i))$ is a training example, and the training set that we will use to learn is $\{(x(i), y(i)), i = 1, 2, ..., m\}$. (i) in the notation is an index into the training set. We use X denote the space of input values, and Y as the space of output values. The goal is, given a training set, to learn a function $h : X \rightarrow Y$ *so that* $h(x)$ is a good predictor for the corresponding value of y.

What is the h?



Supervised Learning

UNIVERSITY OF WESTMINSTER

- let's say we want to predict whether a person on board the Titanic survived or not:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...

- We are predicting whether a passenger survived or not in the second example. In the 'Survived' variable, 0 represents that the person did not survive and 1 means the person did make it out alive. The independent variables here include Pclass, Sex, Age, Fare, etc.
- So, when we are given a target variable (count and Survival in the above two cases) which we have to predict based on a given set of predictors or independent variables (season, holiday, Sex, Age, etc.), such problems are called supervised learning problems.

Linear Regression

UNIVERSITY OF WESTMINSTER

- Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables.
- Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables.
- Linear Regression is of two types: Simple and Multiple
- Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as **temperature, age, salary, price,** etc.
- LR can be used in different sectors such as:
    - The most common use of regression analysis in business is for forecasting future opportunities and threats. Demand analysis, for example, forecasts the amount of things a customer is likely to buy.
    - Sales prediction
    - Predicting the Salary of a person based on years of experience

Independent Variables

Years of Experience	Salary in 1000$
2	15
3	28
5	42
13	64
8	50
16	90
11	58
1	8
9	54

Dependent Variables

Linear Regression

UNIVERSITY OF WESTMINSTER

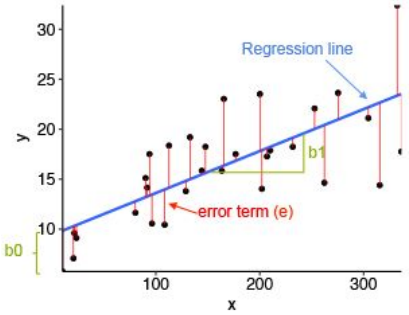
- **Simple Linear Regression** is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable.

$$y = b_0 + b_1 x$$

- In **Multiple Linear Regression** there are more than one independent variables for the model to find the relationship.

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

- A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.



Regression Analysis

UNIVERSITY OF WESTMINSTER

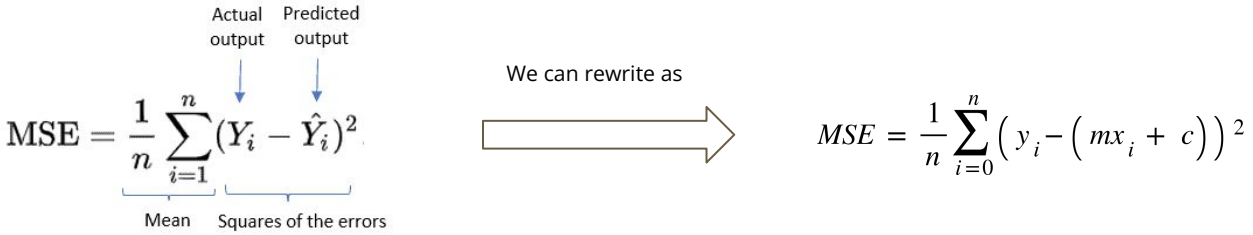
- **R squared or Coefficient of Determination:** The most commonly used metric for model evaluation in regression analysis is R squared. It can be defined as a Ratio of variation to the Total Variation. The value of R squared lies between 0 to 1, the value closer to 1 the better the model.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- **Residual Sum of Squares (RSS):** In general terms, the sum of squares is a statistical technique used in regression analysis to determine the dispersion of data points. The RSS also measures the amount of error remaining between the regression function and the data set after the model has been run.

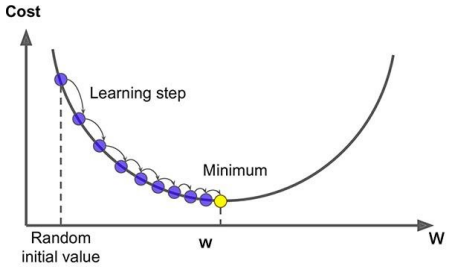
$$RSS = \sum_{i=1}^{n} \left(y_i - \tilde{y}_i \right)^2$$

- **Mean Squared Error (MSE):** Another Common metric for evaluation is Mean squared error which is the mean of the squared difference of actual vs predicted values.



Regression Analysis

- **Gradient Descent:** Gradient descent is an optimisation method which is used to minimize the MSE by calculating the gradient of the cost function. A regression model uses gradient descent to update the coefficients of the line by reducing the cost function. It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.



$$f_b(x) = b_0 + b_1 x$$

$$b_f = b_f - \alpha \frac{\delta}{\delta b_f} f(b_0, b_1) \quad (f \text{ or } f = 0 \text{ and } f = 1)$$

$$f = 0 : \frac{\delta}{\delta b_0} f(b_0, b_1) = \frac{1}{m} \sum_{i=1}^{m} \left(h_b(x^{(i)}) - y^{(i)} \right)$$

$$f = 1 : \frac{\delta}{\delta b_1} f(b_0, b_1) = \frac{1}{m} \sum_{i=1}^{m} \left(h_b(x^{(i)}) - y^{(i)} \right) .x^{(i)}$$

- **Root Mean Squared Error (RMSE):** It is the root of MSE i.e Root of the mean difference of Actual and Predicted values. RMSE penalizes the large errors whereas MSE doesn't.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

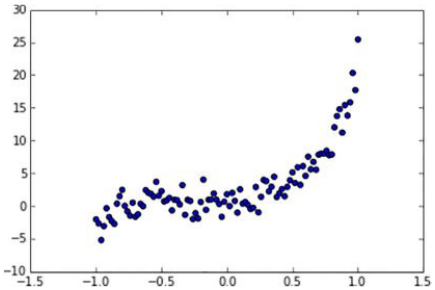
Linear Regression

UNIVERSITY OF WESTMINSTER

What is a Polynomial Model?

- Linear models may be an intuitive first guess, but rarely are real-world correlations so simple. For example, the trajectory of a missile through space is curved relative to the observer on Earth. The change in height of a flower over its lifetime certainly isn't linear.
- When data points appear to form smooth curves rather than straight lines, you need to change your regression model from a straight line to something else. One such approach is to use a polynomial model. A polynomial is a generalization of a linear function. The *nth* degree polynomial looks like the following:

$$f(x) = w_n x^n + \ldots + w_1 x + w_0$$



Regularization

UNIVERSITY OF WESTMINSTER

- When we increase the degree of freedom (increasing polynomials in the equation) for regression models, they tend to overfit. Using the regularization techniques we can overcome the overfitting issue.
- Two popular methods for that is lasso and ridge regression.
- **Lasso Regression:**
  - Is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean.
  - It is used when we have more features because it automatically performs feature selection.
  - It uses L1 regularization. In this technique we set penalties to coefficients . Larger penalties result in coefficient values that are closer to zero.

$$Cost\ (W) = RSS(W) + \lambda * (sum\ of\ absolute\ value\ of\ weights)$$

$$= \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j\, x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} |w_j|$$

- **Ridge Regression:**
  - The main purpose of Ridge Regression is, to find the coefficients that minimize the sum of error squares by applying a penalty to these coefficients. It performs L2 regularization, by adding penalty equivalent to square of the magnitude of coefficients.

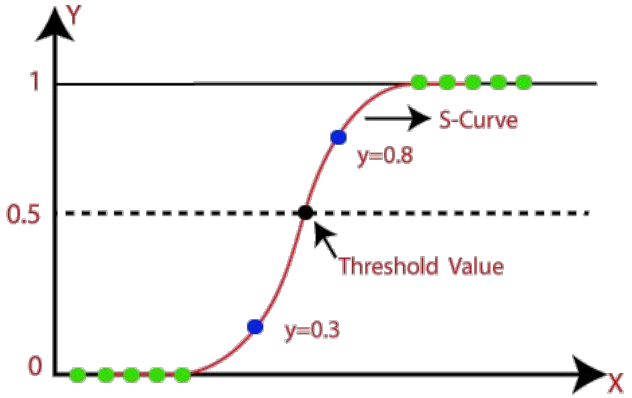
$$SSE_{L_2} = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 + \lambda \sum_{j=1}^{P} {\beta_j}^2$$

Logistic Regression

UNIVERSITY OF WESTMINSTER

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- Logistic Regression Use cases (but not limited to):
    - Credit Scoring
    - Medicine
    - Text Editing
    - Hotel Booking

User ID	Gender	Age	EstimatedSalary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	0
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	0
15728773	Male	27	58000	0
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	0
15727311	Female	35	65000	0



Logistic Regression

UNIVERSITY OF WESTMINSTER

- In the logistic regression the constant (b0) moves the curve left and right and the slope (b1) defines the steepness of the curve. By simple transformation, the logistic regression equation can be written in terms of an odds ratio.

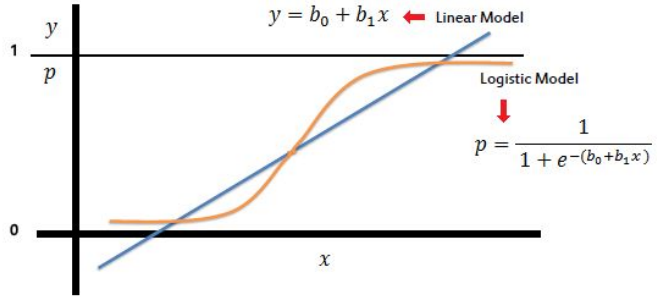
$$\frac{p}{1-p} = \exp(b_0 + b_1 x)$$

- Finally, taking the natural log of both sides, we can write the equation in terms of log-odds (logit) which is a linear function of the predictors. The coefficient (b1) is the amount the logit (log-odds) changes with a one unit change in x.

$$ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

- logistic regression can handle any number of numerical and/or categorical variables.

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p)}}$$



Naive Bayes Classifier

UNIVERSITY OF WESTMINSTER

- Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

- Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. The assumption made here is that the predictors/features are independent.
- Based on the example, the formula could be written as follows:

$$P(y|X) = \frac{P(X|y)\,P(y)}{P(X)}$$

Where:

$$X = (x_1, x_2, \ldots x_n)$$

Then:

$$(y|x_1, \ldots, x_n) = \frac{P(x_1|y)\,P(x_2|y)\ldots P(x_n|y)\,P(y)}{P(x_1)\,P(x_2)\ldots P(x_n)}$$

	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY GOLF
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes

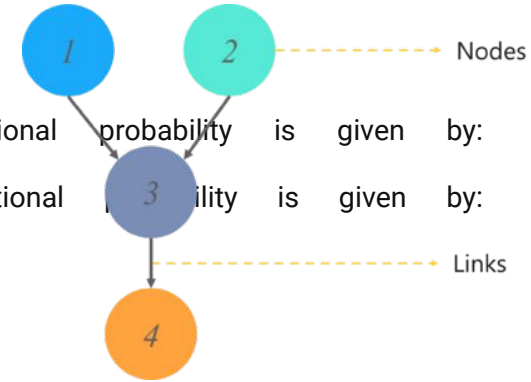
- Applications:
  - Credit scoring
  - Weather forecasting
  - Spam Filtering

Bayesian Networks

UNIVERSITY OF WESTMINSTER

- A **Bayesian network** is a probabilistic graphical model that measures the conditional dependence structure of a set of random variables based on the Bayes theorem and that is used to compute uncertainties by using the concept of probability.
- Bayesian networks are graphical models that contain information about causal probability relationships between variables and are often used to aid in decision making
- Each node corresponds to a random variable and each edge represents the conditional probability for the corresponding random variables
- **Joint Probability**: is a statistical measure of two or more events happening at the same time, i.e., $P(A, B, C)$, The probability of event A, B and C occurring. It can be represented as the probability of the intersection two or more events occurring.
- **Conditional Probability:** Conditional Probability of an event X is the probability that the event will happen given that an event Y has already happened.

$p(X|\ Y)$ is the probability of event X occurring, given that event, Y occurs:



- If X and Y are dependent events then the expression for conditional probability is given by: $P\ (X|\ Y) = P\ (X\ and\ Y) / P\ (Y)$
- If A and B are independent events then the expression for conditional probability is given by: $P(X|\ Y) = P\ (X)$
- The predicted output for the probabilistic model is represented as follows:

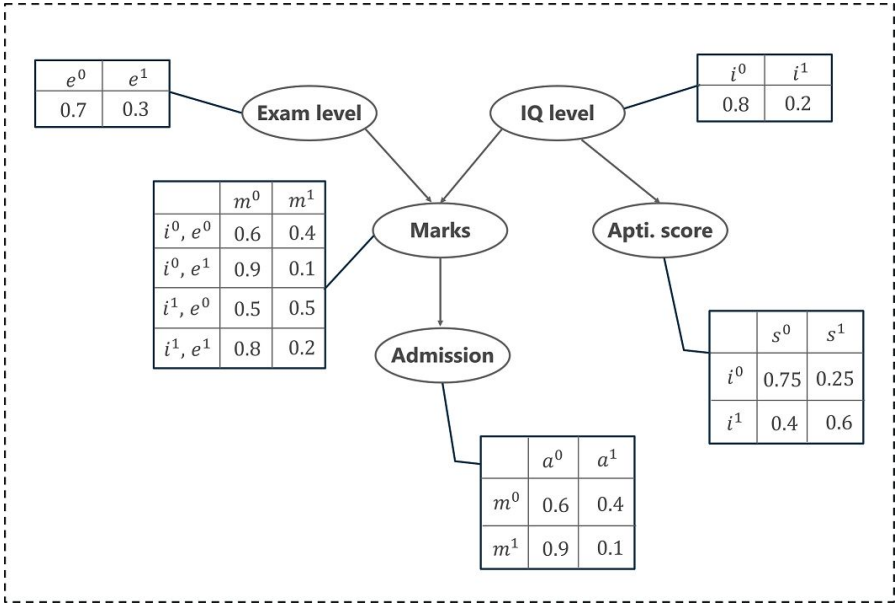
$$P\left(X_1, \ldots, X_n\right) = \pi_{i=1}^{n} p\left(X_i \mid Parents\left(X_i\right)\right)$$

Bayesian Networks

UNIVERSITY OF WESTMINSTER

$$p(a, m, i, e, s) = p(a \mid m)\, p(m \mid i, e)\; p(i)\; p(e)\; p(s \mid i)$$

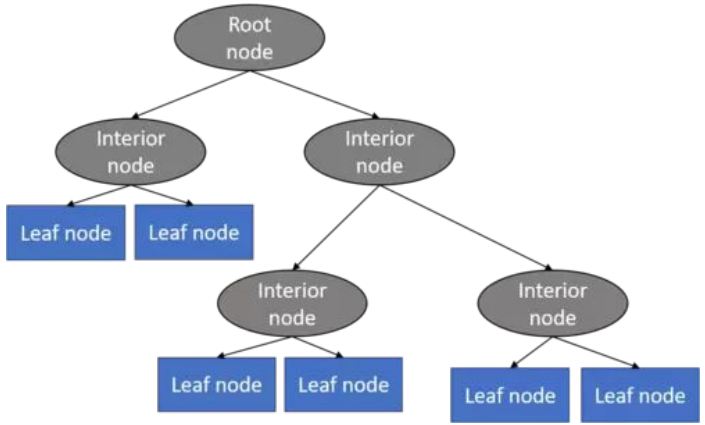
- ***p(a | m)*** represents the conditional probability of a student getting an admission based on his marks.
- ***p(m | I, e)*** represents the conditional probability of the student's marks, given his IQ level and exam level.
- ***p(i)*** denotes the probability of his IQ level (high or low)
- ***p(e)*** denotes the probability of the exam level (difficult or easy)
- ***p(s | i)*** denotes the conditional probability of his aptitude scores, given his IQ level
- Applications:
    a. Medicine
    b. Spam filtering
    c. Document Classification



Decision Tree

UNIVERSITY OF WESTMINSTER

- Decision trees are supervised learning algorithms used for both, classification and regression tasks
- In principle, decision trees can be used to predict the target feature of a unknown query instance by building a model based on existing data for which the target feature values are known (supervised learning).
- It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.
- The leaf nodes contain the predictions we will make for new query instances presented to our trained model. This is possible since the model has kind of learned the underlying structure of the training data and hence can, given some assumptions, make predictions about the target feature value (class) of unseen query instances.
- Applications:
    - Marketing
    - Fraud Detection
    - Customers Retention



Decision Trees

UNIVERSITY OF WESTMINSTER

How Decision Trees Work?

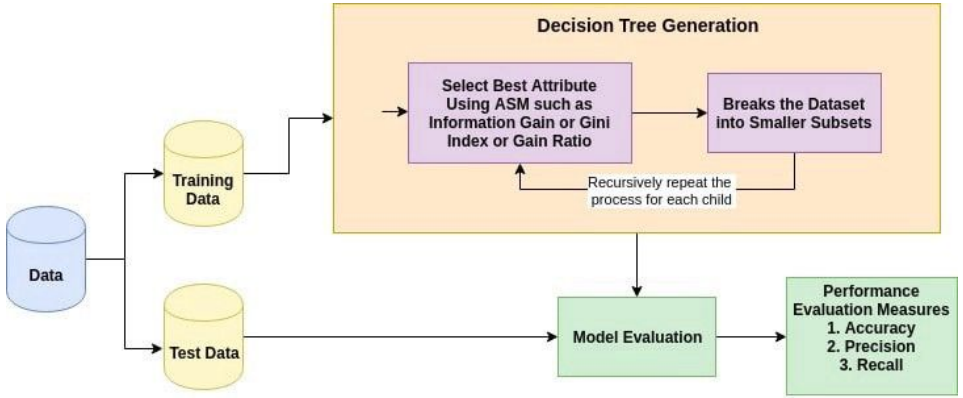
1. The root node feature is selected based on the results from the Attribute Selection Measure(ASM).
2. The ASM is repeated until a leaf node, or a terminal node cannot be split into sub-nodes.

The two main ASM techniques are

1. Gini index
2. Information Gain(ID3)

Gini Index: The measure of the degree of probability of a particular variable being wrongly classified when it is randomly chosen is called the Gini index or Gini impurity. The data is equally distributed based on the Gini index.

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2$$



Decision Trees

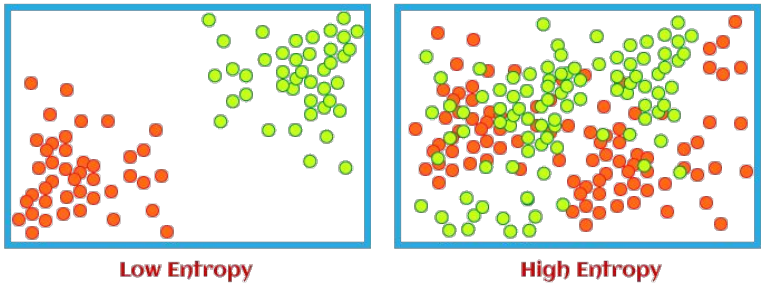
UNIVERSITY OF WESTMINSTER

How to predict the output from Decision Tree?

Entropy: Entropy is a measure of disorder or uncertainty and the goal of machine learning models and Data Scientists in general is to reduce uncertainty.

$$H(p_1, p_2, \ldots, p_n) = -K \sum_{i=1}^{n} p_i \log(p_i)$$

$$Entropy = \frac{-p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$



Information Gain: Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute. It calculates how much information a feature provides us about a class. According to the value of information gain, we split the node and build the decision tree.

$$I(Attribute) = \sum \frac{p_i + n_i}{p+n} Entropy(Attribute) \quad \Rightarrow \quad IG = H(S) - I(Attribute)$$

Decision Trees

UNIVERSITY OF WESTMINSTER

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y

First Step: Calculate Entropy for the Entire Dataset

p = no of positive cases(Loan_Status accepted) , n = number of negative cases(Loan_Status not accepted). *In the data set, we have p = 332 , n=148, p+n=480*

$$E(s) = (-332/480)\log(332/480)-(148/480)\log(148/480)$$

$$= 0.8912402012$$

Second Step: Find the the entropy and gain for each column

Male: p = 278, n=116 , p+n=489 Entropy(G=Male) = 0.87

Female: p = 54 , n = 32 , p+n = 86 Entropy(G=Female) = 0.95

Gain(Gender) = E(s) — I(Gender)

$$I(Gender) = (Entropy(G = Male)* p+n/480)+(Entropy(G=Female)*p+n/480$$

$$= [0.87*(278+116)/480]+[0.95*(54+32)/480]$$

$$= 0.88$$

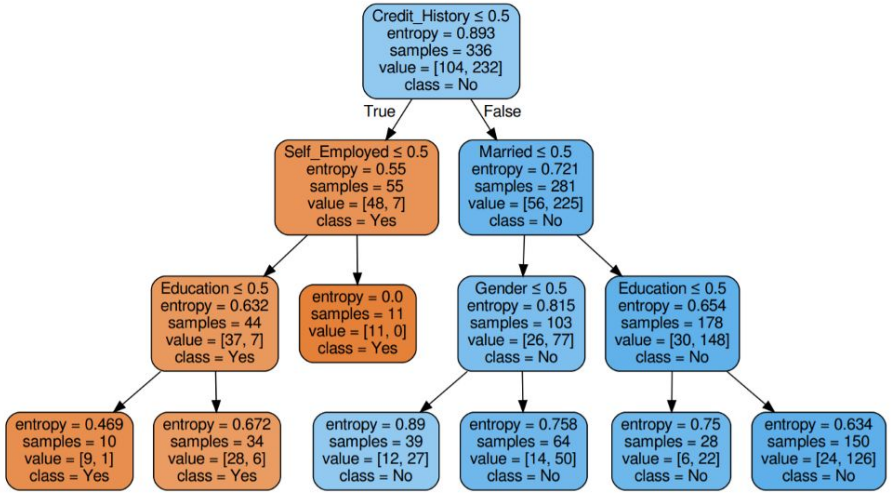
Gain = 0.89–0.88=0.001

Decision Trees

UNIVERSITY OF WESTMINSTER

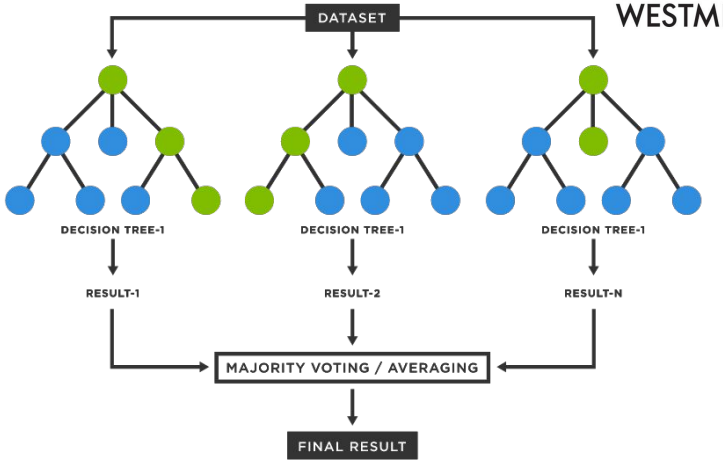
After calculating information gain for all features, we can select the one with the highest IG.

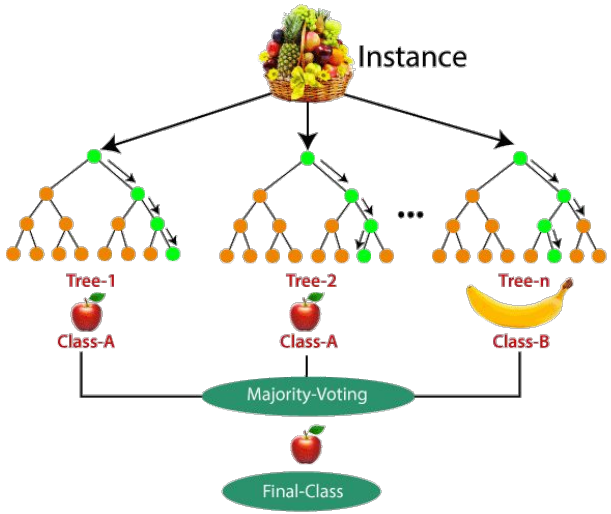
```
Gain(Gender)          = 0.01
Gain(Married)         = 0.01
Gain(Education)       = 0.01
Gain(Self-Employed)   = 0.01
Gain(Credit Score)    = 0.2
```



Random Forests

- Random forest is a supervised machine learning algorithm. It is one of the most used algorithms due to its accuracy, simplicity, and flexibility.
- The data from these trees are then merged together to ensure the most accurate predictions. While a solo decision tree has one outcome and a narrow range of groups, the forest assures a more accurate result with a bigger number of groups and decisions.
- The data from these trees are then merged together to ensure the most accurate predictions. While a solo decision tree has one outcome and a narrow range of groups, the forest assures a more accurate result with a bigger number of groups and decisions.

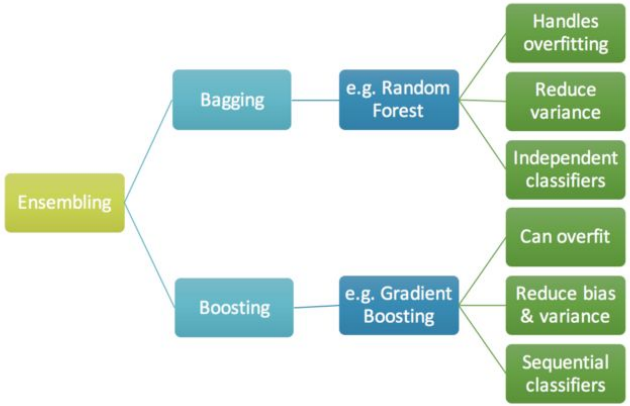




Gradient Boosting

UNIVERSITY OF WESTMINSTER

- When we try to predict the target variable using any machine learning technique, the main causes of difference in actual and predicted values are noise, variance, and bias. Ensemble helps to reduce these factors.
- An ensemble is just a collection of predictors which come together (e.g. mean of all predictions) to give a final prediction.
- Ensembling techniques are further classified into Bagging and Boosting.
    - **Bagging** is a simple ensembling technique in which we build many independent predictors/models/learners and combine them using some model averaging techniques. (e.g. weighted average, majority vote or normal average).
    - **Boosting** is an ensemble technique in which the predictors are not made independently, but sequentially. This technique employs the logic in which the subsequent predictors learn from the mistakes of the previous predictors.



Gradient Boosting

UNIVERSITY OF WESTMINSTER

- Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.
- we have mean squared error (MSE) as loss defined as:

$$Loss = MSE = \sum (y_i - y_i^p)^2$$

where, y_i = ith target value, y_i^p = ith prediction, $L(y_i, y_i^p)$ is Loss function

- We want our predictions, such that our loss function (MSE) is minimum. By using gradient descent and updating our predictions based on a learning rate, we can find the values where MSE is minimum.

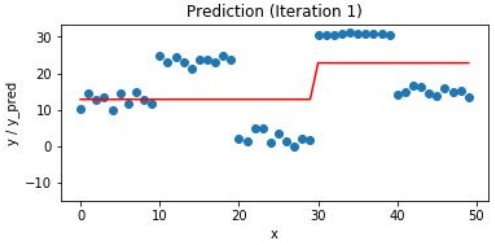
$$y_i^p = y_i^p + \alpha * \delta \sum (y_i - y_i^p)^2 / \delta y_i^p$$

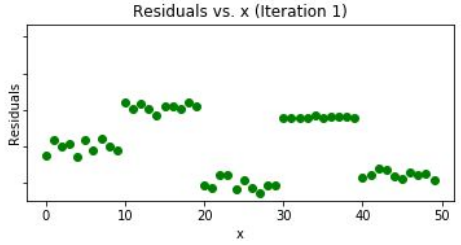
which becomes, $y_i^p = y_i^p - \alpha * 2 * \sum (y_i - y_i^p)$

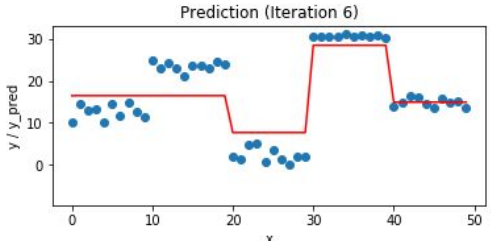
where, α is learning rate and $\sum (y_i - y_i^p)$ is sum of residuals

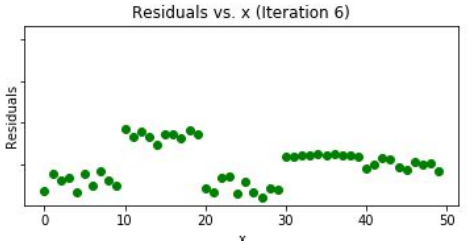
Gradient Boosting

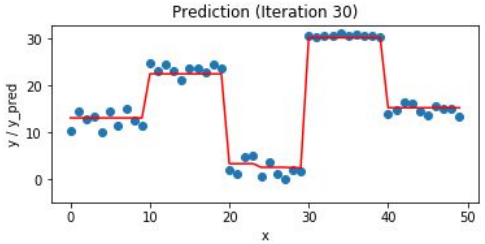
UNIVERSITY OF WESTMINSTER

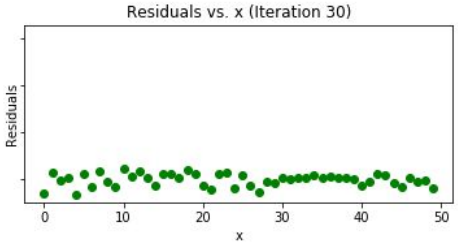








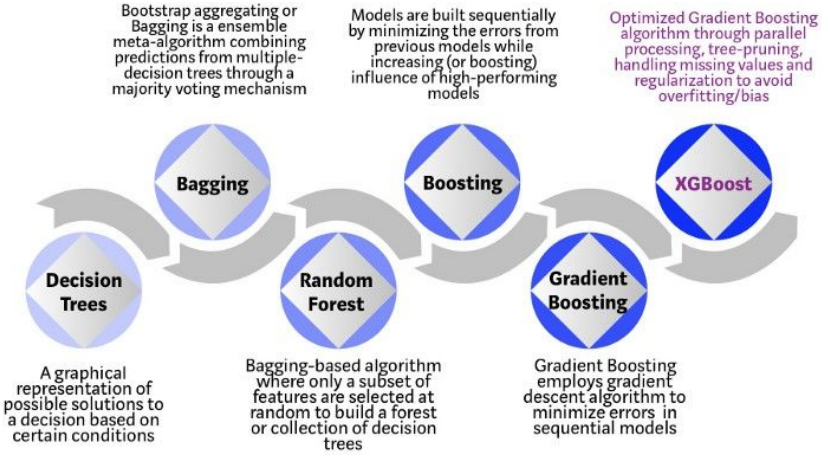


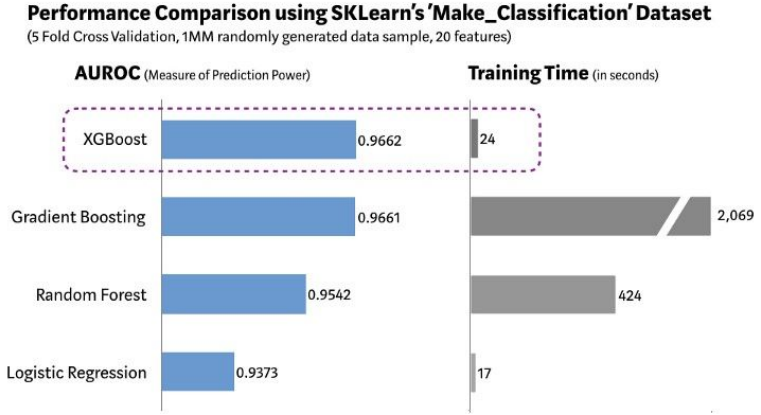


XG-Boost

UNIVERSITY OF WESTMINSTER

XGBoost and Gradient Boosting Machines (GBMs) are both ensemble tree methods that apply the principle of boosting weak learners using the gradient descent architecture. However, XGBoost improves upon the base GBM framework through systems optimization and algorithmic enhancements.

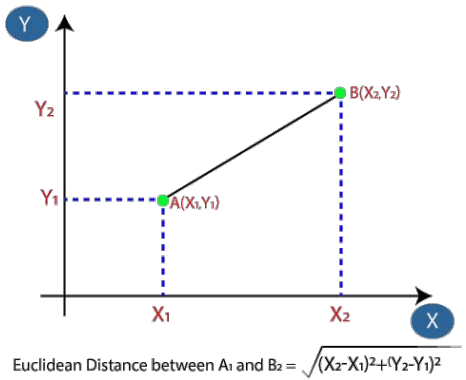




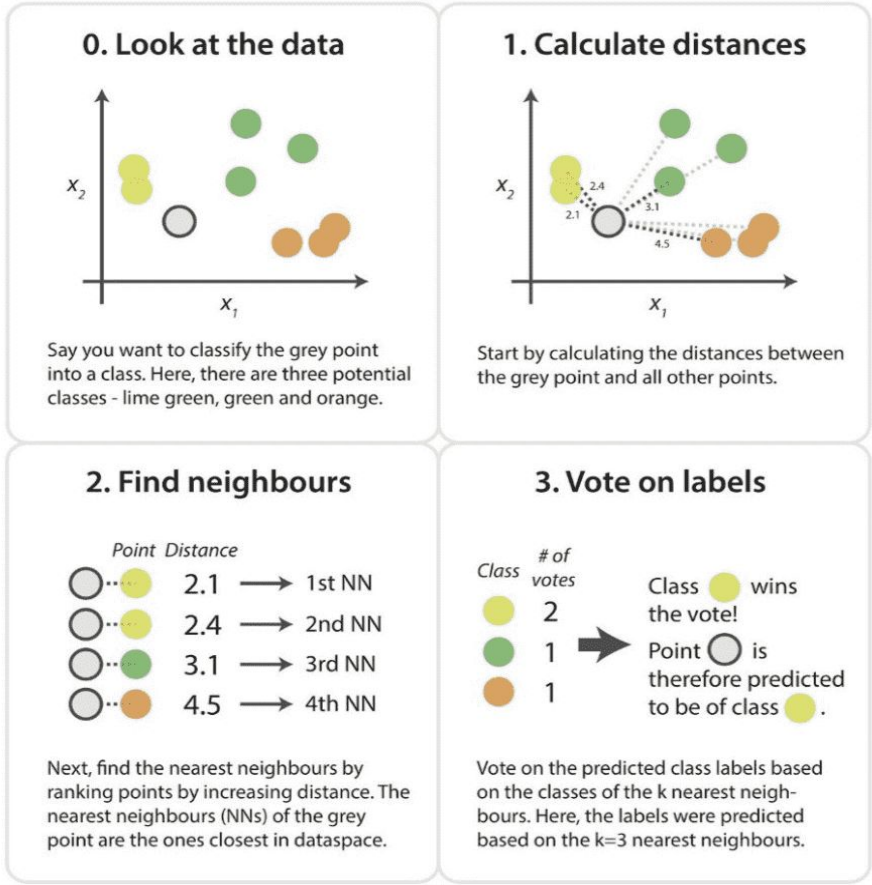
K-Nearest Neighbors (KNN)

UNIVERSITY OF WESTMINSTER

- K-nearest neighbors (KNN) is a type of supervised learning algorithm used for both regression and classification.
- KNN tries to predict the correct class for the test data by calculating the distance between the test data and all the training points.
- Then select the K number of points which is closest to the test data. The KNN algorithm calculates the probability of the test data belonging to the classes of 'K' training data and class holds the highest probability will be selected.
- In order to find the closest points, we need to apply Euclidean distance between the data points.
- Applications
  - Recommendation systems
  - Facial recognitions
  - Finance
  - Text mining



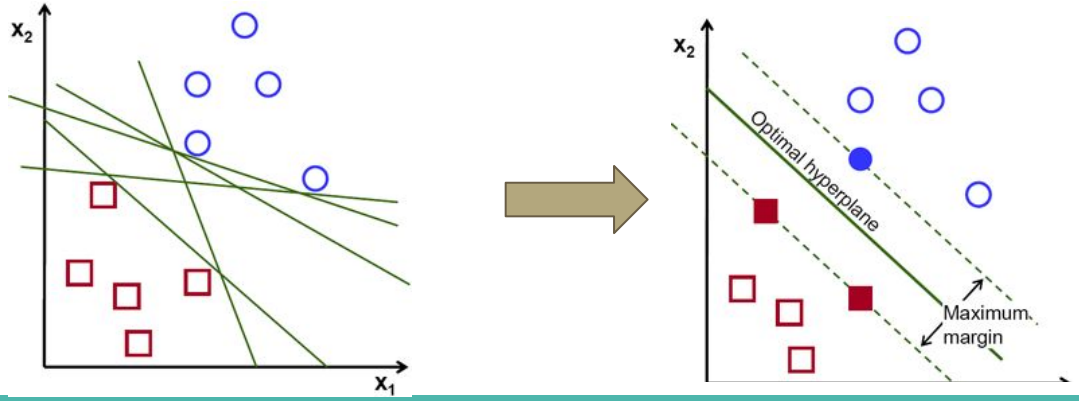
Euclidean Distance between A_1 and $B_2 = \sqrt{(X_2-X_1)^2+(Y_2-Y_1)^2}$



Support Vector Machines

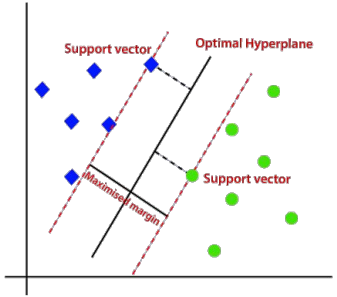
UNIVERSITY OF WESTMINSTER

- Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.
- Applications:
  - Image Classification
  - Prediction of common diseases
  - Face detection
- The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a *hyperplane*.
- We always create a hyperplane that has a *maximum margin*, which means the maximum distance between the data points.

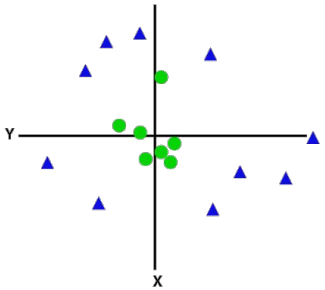


Support Vector Machines

- Linear SVM : Linear SVM is used for data that are linearly separable i.e. for a dataset that can be categorized into two categories by utilizing a single straight line. Such data points are termed as linearly separable data, and the classifier is used described as a Linear SVM classifier.



- Non-linear SVM: is used for data that are non-linearly separable data i.e. a straight line cannot be used to classify the dataset. Such data points are termed as non-linear data, and the classifier used is termed as a Non-linear SVM classifier.



Support Vector Machines

UNIVERSITY OF WESTMINSTER

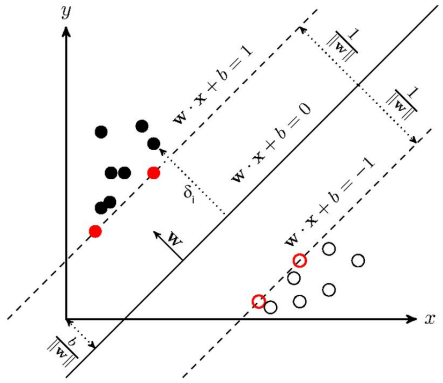
How to find the hyperplane?

- The hyperplane is defined by finding the optimal values w or weights and b or intercept which. And these optimal values are found by minimizing the cost function.
- We are looking for the hyperplane parameters (w, b), so that the distance between the hyperplane and the observations is maximized. We shall call the Euclidean distance between the point xi and the hyperplane as geometric margin.

$$\delta_i = \frac{y_i\left(\omega \cdot x_i + b\right)}{\|\omega\|}$$

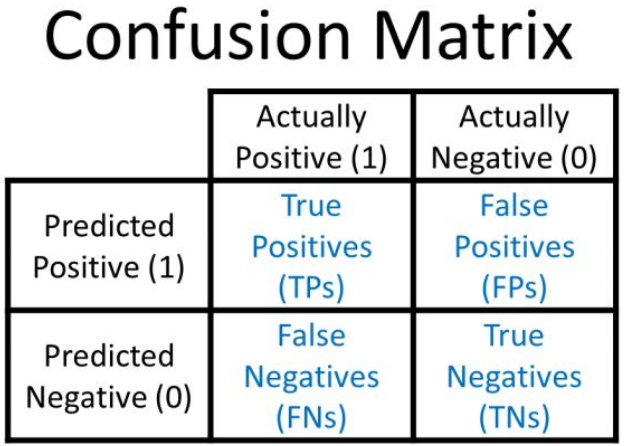
- The optimal hyperplane has equation $w.x+b = 0$. The left support vector has equation $w.x+b=-1$ and the right support vector has $w.x+b=1$.
- The Cost Function is used to train the SVM. By minimizing the value of $J(w)$, we can ensure that the SVM is as accurate as possible.

$$J(\omega) = \frac{1}{2}\|\omega\|^2 + C\left[\frac{1}{N}\sum_i^n max\left(0, 1 - y_i \cdot \left(\omega \cdot x_i + b\right)\right)\right]$$



Confusion Matrix

- A confusion matrix is a technique for summarizing the performance of a classification algorithm. It consists of a two dimensional table.
- **True Positive:** The number of times our actual positive values are equal to the predicted positive. You predicted a positive value, and it is correct.
- **False Positive:** The number of times our model wrongly predicts negative values as positives. You predicted a negative value, and it is actually positive.
- **True Negative:** The number of times our actual negative values are equal to predicted negative values. You predicted a negative value, and it is actually negative.
- **False Negative:** The number of times our model wrongly predicts negative values as positives. You predicted a negative value, and it is actually positive.



Classification report

UNIVERSITY OF WESTMINSTER

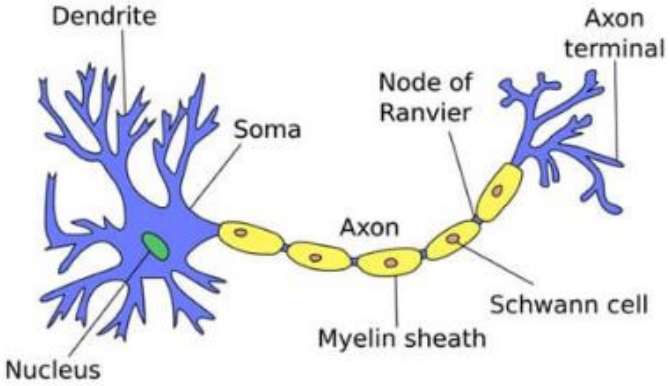
It is one of the performance evaluation metrics of a classification-based machine learning model. It displays your model's precision, recall, F1 score and support. It provides a better understanding of the overall performance of our trained model.

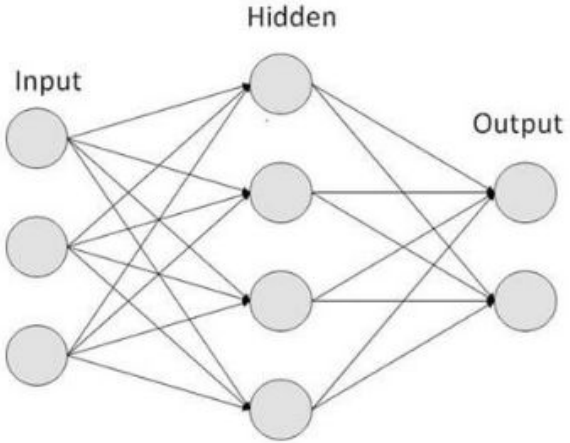
Metrics	Definition
Precision	Precision is defined as the ratio of true positives to the sum of true and false positives.
Recall	Recall is defined as the ratio of true positives to the sum of true positives and false negatives.
F1 Score	The F1 is the weighted harmonic mean of precision and recall. The closer the value of the F1 score is to 1.0, the better the expected performance of the model is.
Support	Support is the number of actual occurrences of the class in the dataset. It doesn't vary between models, it just diagnoses the performance evaluation process.

Neural Networks

UNIVERSITY OF WESTMINSTER

- Neural Networks are a subset of Machine Learning techniques which learn the data and patterns in a different way utilizing Neurons and Hidden layers.
- Neural Networks are way more powerful due to their complex structure and can be used in applications where traditional Machine Learning algorithms just cannot suffice.
- A neural network works similarly to the human brain's neural network.
- A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates.
- Applications:
    - Classification
    - Face Recognition
    - Pattern Recognition
    - Speech Recognition





Lasso can be interpreted as least-squares linear regression where:

Weights are regularized with the L1 norm	The weights have a Gaussian prior	Weights are regularized with the L2 norm	The solution algorithm is simpler

Suppose we train several classifiers to learn f : X → Y , where X is the feature vector X =< X1, X2, X3 >. Which classifier contains sufficient information to allow calculating P(X1, X2, X3, Y)?

Linear Regression	Naive Bayes Classifier	Logistic Regression	K-means Clustering

In which of the following algorithms, the subsequent predictors learn from the mistakes of the previous predictors?

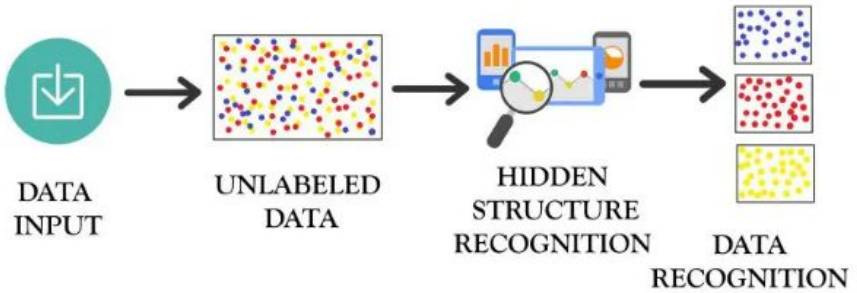
SVM	Linear regression	Gradient Boosting	Random Forest	None of the above

Unsupervised Learning Algorithms

Unsupervised Learning

UNIVERSITY OF WESTMINSTER

- Unsupervised machine learning is the training of models on raw and unlabelled training data. It is often used to identify patterns and trends in raw datasets, or to cluster similar data into a specific number of groups.
- The data given to unsupervised algorithms is not labelled, which means only the input variables (x) are given with no corresponding output variables.
- This algorithm can easily analyze massive amounts of data. Moreover, the speed of providing results is also faster. And even though the speed is fast, it is able to provide accurate data recognition results.
- Unsupervised machine learning is mainly used to:
    - Cluster datasets on similarities between features or segment data
    - Understand relationship between different data point such as automated music recommendations
    - Anomaly detection



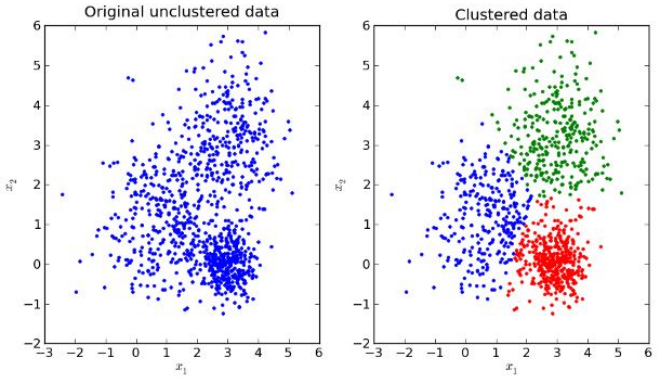
Unsupervised Learning

UNIVERSITY OF WESTMINSTER

- Clustering is basically a technique that groups similar data points such that the points in the same group are more similar to each other than the points in the other groups. The group of similar data points is called a Cluster.
- Let's consider a dataset of m-dimensional samples: $X = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n\} \;\; where \;\; \bar{x}_i \in \mathbb{R}^m$
- Let's assume that it's possible to find a criterion (not a unique) so that each sample can be associated with a specific group according to its peculiar features and the overall structure of the dataset:

$$g_k = G(\bar{x}_i) \;\; where \;\; k = \{0, 1, 2, \ldots, t\}$$

- Conventionally, each group is called a cluster, and the process of finding the function, G, is called clustering.



Unsupervised Learning

UNIVERSITY OF WESTMINSTER

Types of Clustering Algorithms:

- **Centroid Algorithm**: This is an iterative approach of finding cluster centroid and deciding cluster based on the point's distance to each cluster center. This process is repeated until the centroid movement converges. (Ex:- k-Means)
- **Density Algorithm**: Density-Based Clustering refers to unsupervised learning methods that identify distinctive groups/clusters in the data, based on the idea that a cluster in a data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density. This algorithm requires only one-pass over the data space. (Ex:- DBSCAN)
- **Distribution Algorithm**: This algorithm is based on the idea that clusters can easily be defined as objects belonging most likely to the same distribution. (Ex:- Gaussian Mixture Models)
- **Connectivity Algorithm**: The core idea behind this algorithm is that data points closer to one another in N-dimensional space tend to have similar properties than data points much farther away. (Ex:- Hierarchical Clustering).

K-means Clustering

UNIVERSITY OF WESTMINSTER

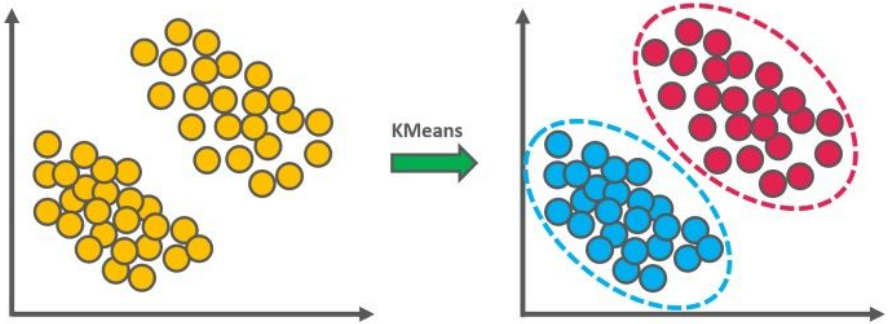
- K-Means clustering is an unsupervised learning algorithm that learns properties of a set of data points and forms partitions called clusters, that represent data with similar properties.
- For continuous data, each cluster is represented by the centroid which is the mean of cluster members. K-Means uses squared Euclidean distance as the similarity measure for cluster membership:

$$d_{sq} = \sum_{i=1}^{D} (x_i - y_i)^2$$

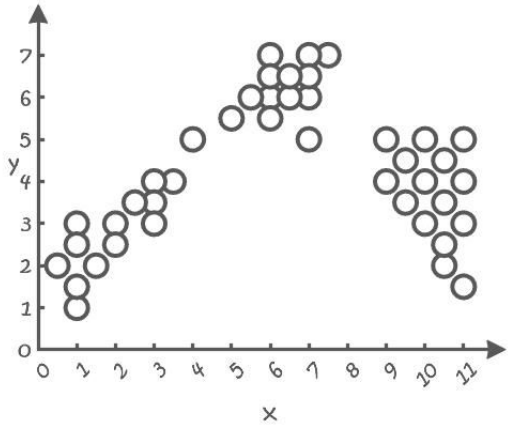
- Here x and y are points in the D-dimensional space. Number of clusters k is determined by minimizing the Sum of Squared Errors (SSE), which is the sum of squared error of each data point and its nearest centroid. It is given by

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{i,j} \left\| x_i - c_j \right\|^2$$

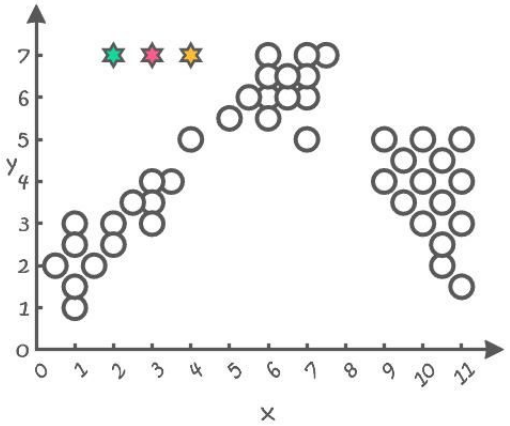
- Applications:
  - Identifying fraudulent or criminal activity.
  - Customer segmentation
  - Insurance Fraud Detection



K-means clustering

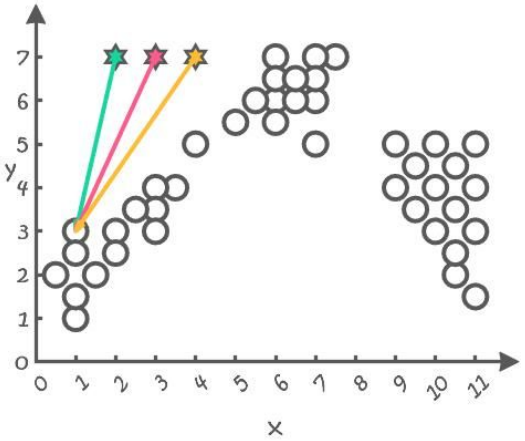


(1) Data points



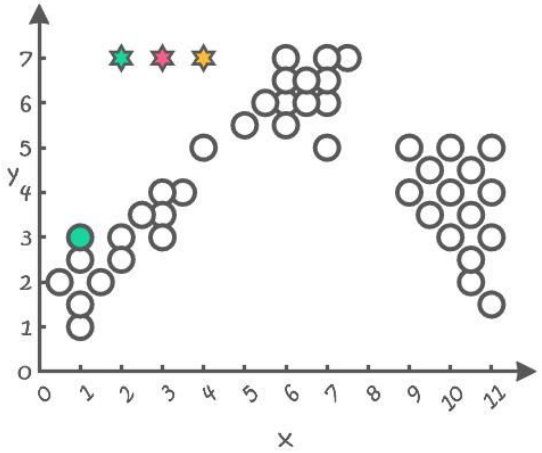
(2) Specifying K.

Here the k =3 so we have 3 random data points which will be centroids

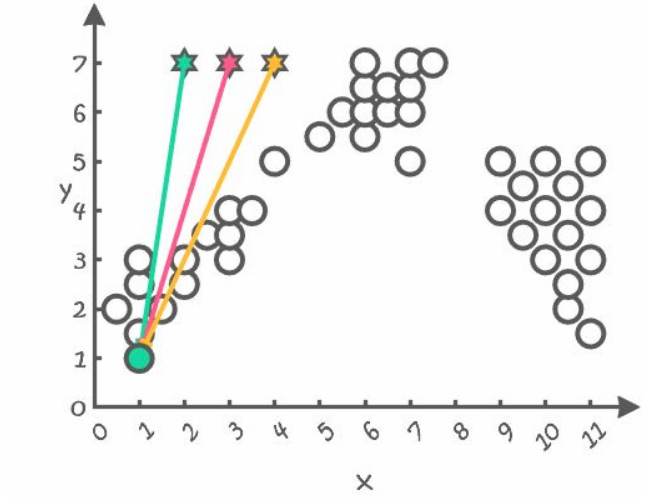


(3) Measure the distance between each data point with each of randomly selected points (centroids).

K-means Clustering



By applying the distance calculation for each data point, we have



- After calculating the distance between the data points and each centroid, now the first phase has been completed!
- Now, each data point finds its cluster. But it is not completed yet. We need to place each of those three centroid at the center of each cluster!

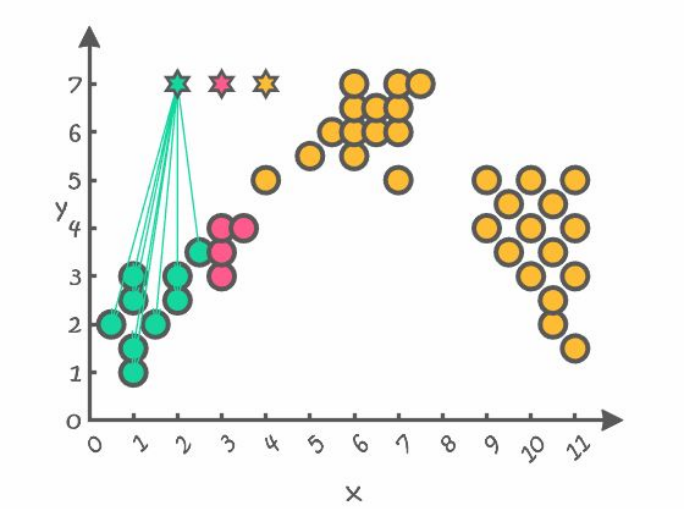
K-Means Clustering

UNIVERSITY OF WESTMINSTER

- If we calculate the average of data points in each cluster, we will have a new cluster center. By having this, we will be able to assign the centroids at the center of their clusters.

$$c_k = \frac{1}{|s_k|} \sum_{p_i \in s_k} p_i$$

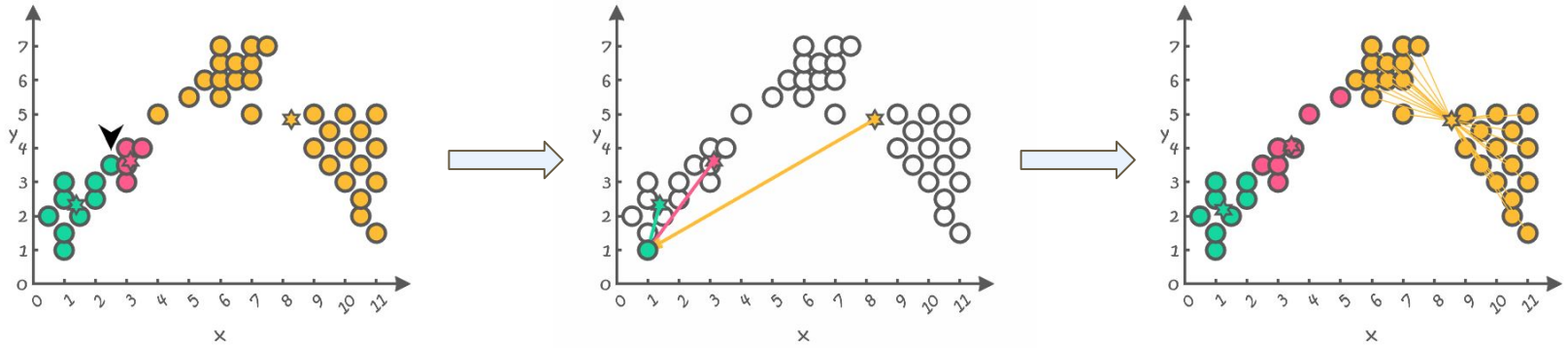
$$c_1 = \frac{(0.5,2) + (1,1) + (1,1.5) + (1,2.5) + (1,3) + (1.5,2) + (2,2.5) + (2,3) + (2.5,3.5)}{9} = \frac{(12.5,21)}{9} = (1.39,2.33)$$



K-means Clustering

UNIVERSITY OF WESTMINSTER

- Now, we have assigned each centroid to the center of its corresponding cluster.
- The next step is to repeat step 3, which is calculating again the distance of each data point and each centroid. This is done as we want to make sure that each data point is in the right cluster as in the previous step, the locations of the centroids were changed.



Centroids are placed at the center of each cluster

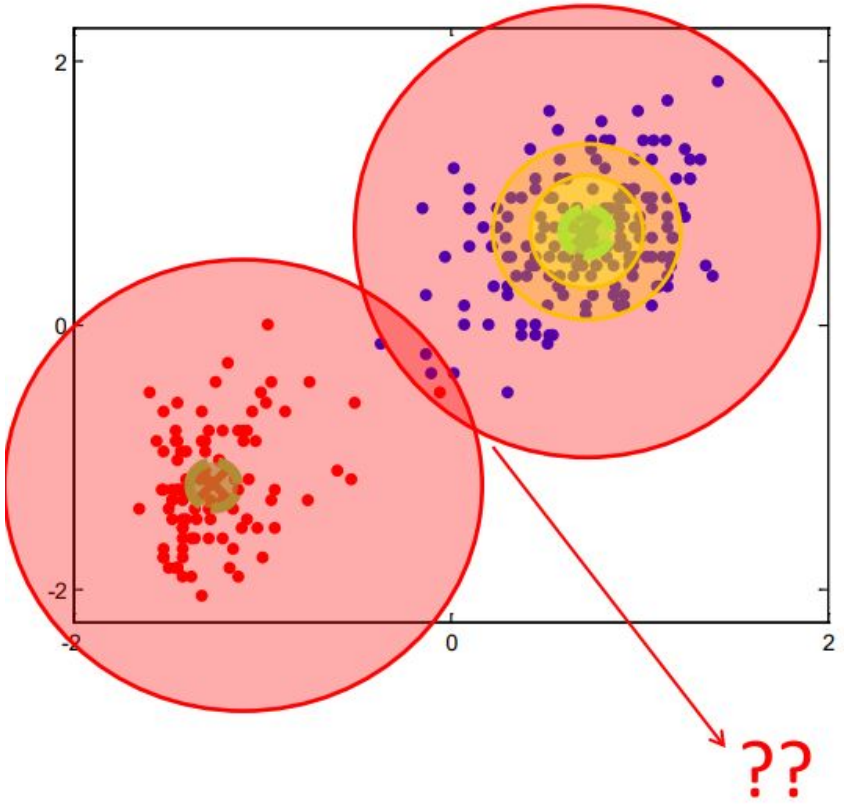
Calculating the distance of each data point and the centroids at the center of each cluster

Final Clustering phase

Problems with K-means Clustering?

- K-Means has no mechanism to handle the uncertainty when a data point is close to more than one cluster centroid.
- K-Means fails to produce optimal clusters for complex, non-linear decision boundaries.
- It is sensitive to initial guess of centroids. Different initializations may lead to different clusters.

Solution?



Gaussian Mixture Model (GMM)

UNIVERSITY OF WESTMINSTER

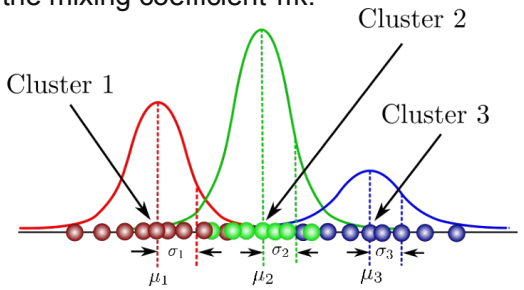
- Gaussian distribution, also called the Normal distribution, is a continuous probability distribution, given by:

$$\mathcal{N}(X|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}\sqrt{|\Sigma|}} exp\left\{-\frac{(X-\mu)^T\Sigma^{-1}(X-\mu)}{2}\right\}$$

- A GMM is an unsupervised clustering technique that forms ellipsoidal shaped clusters based on probability density estimations using the Expectation-Maximization. Each cluster is modelled as a Gaussian distribution. The mean and the covariance rather than only the mean as in K-Means, give GMMs the ability to provide a better quantitative measure of fitness per number of clusters.
- A GMM is represented as a linear combination of the basic Gaussian probability distribution and is expressed as:

$$p(X) = \sum_{k=1}^{K} \pi_k \mathcal{N}(X|\mu_k,\Sigma_k)$$

- where, K is the number of components in the mixture model and πk is called the mixing coefficient, which gives an estimate of the density of each Gaussian component. The Gaussian density given by $N(X|\mu k, \Sigma k)$, is called a component of the mixture model. Each component k is described by a Gaussian distribution with mean μk, covariance Σk and the mixing coefficient πk.



Density-Based Spatial Clustering of Applications with Noise(DBSCAN)

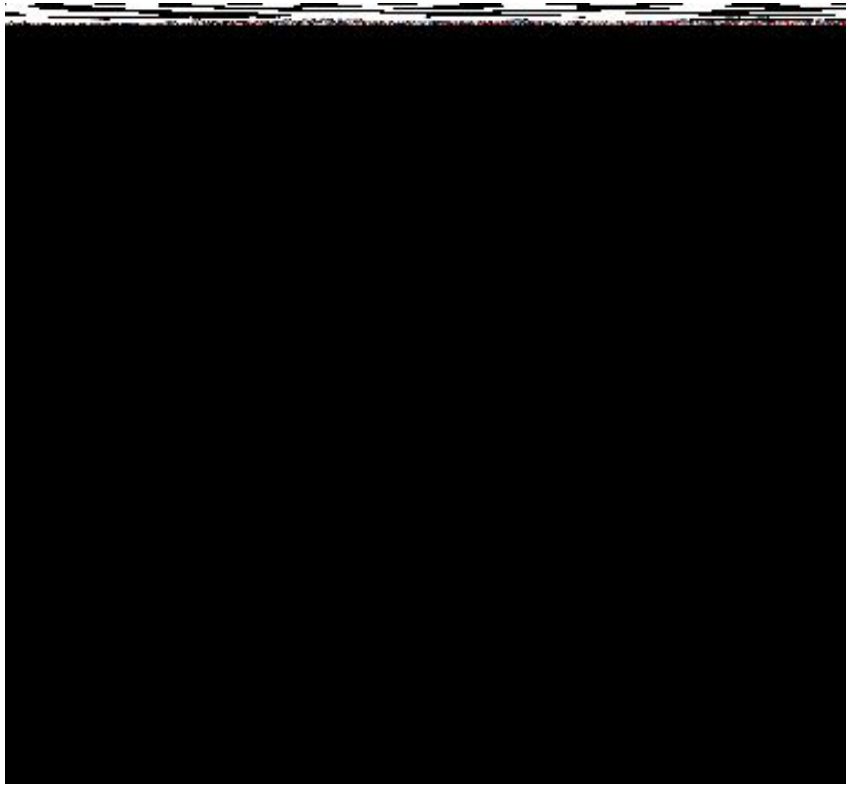
UNIVERSITY OF WESTMINSTER

- Density-Based **Clustering** refers to unsupervised learning methods that identify distinctive groups/clusters in the data, based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density.
- DBSCAN is a base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers.
- The DBSCAN algorithm uses two parameters:
    - **minPts:** The minimum number of points (a threshold) clustered together for a region to be considered dense.
    - **eps (ε):** A distance measure that will be used to locate the points in the neighborhood of any point.

Density-Based Spatial Clustering of Applications with Noise(DBSCAN)

Algorithmic steps for DBSCAN clustering:

1. The algorithm proceeds by arbitrarily picking up a point in the dataset (until all points have been visited).
2. If there are at least ‘minPoint’ points within a radius of ‘ε’ to the point then we consider all these points to be part of the same cluster.
3. The clusters are then expanded by recursively repeating the neighborhood calculation for each neighboring point.

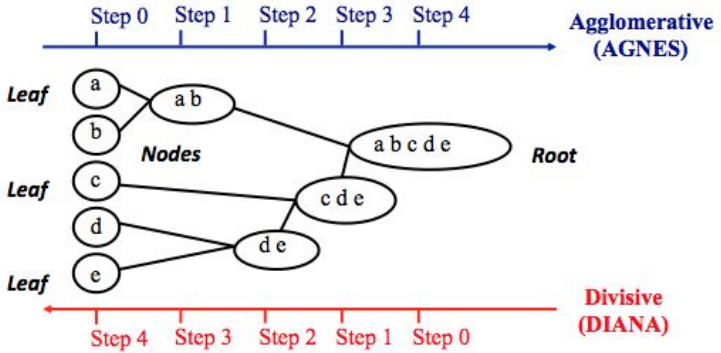


Hierarchical Clustering

UNIVERSITY OF WESTMINSTER

Hierarchical Clustering is based on the general concept of finding a hierarchy of partial clusters, built using either a bottom-up or a top-down approach. More formally, they are split into two categories:

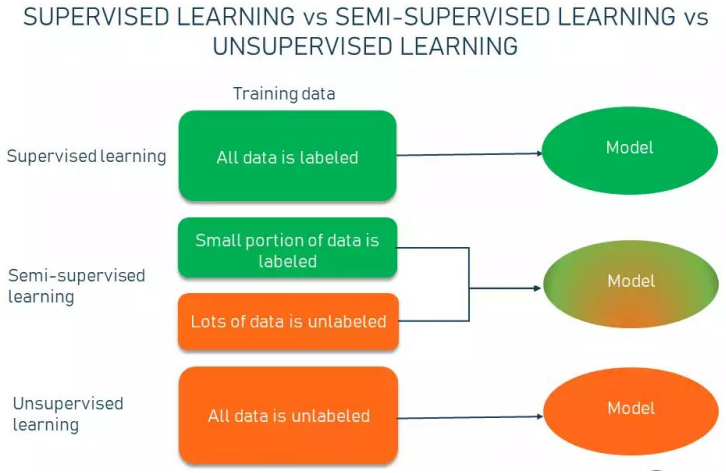
- **Agglomerative Clustering:** The process starts from the bottom (each initial cluster is made up of a single element) and proceeds by merging the clusters until a stop criterion is reached. In general, the target has a sufficiently small number of clusters at the end of the process.
- **Divisive Clustering:** In this case, the initial state is a single cluster with all samples, and the process proceeds by splitting the intermediate cluster until all the elements are separated. At this point, the process continues with an aggregation criterion based on dissimilarity between elements.



Semi-Supervised Learning

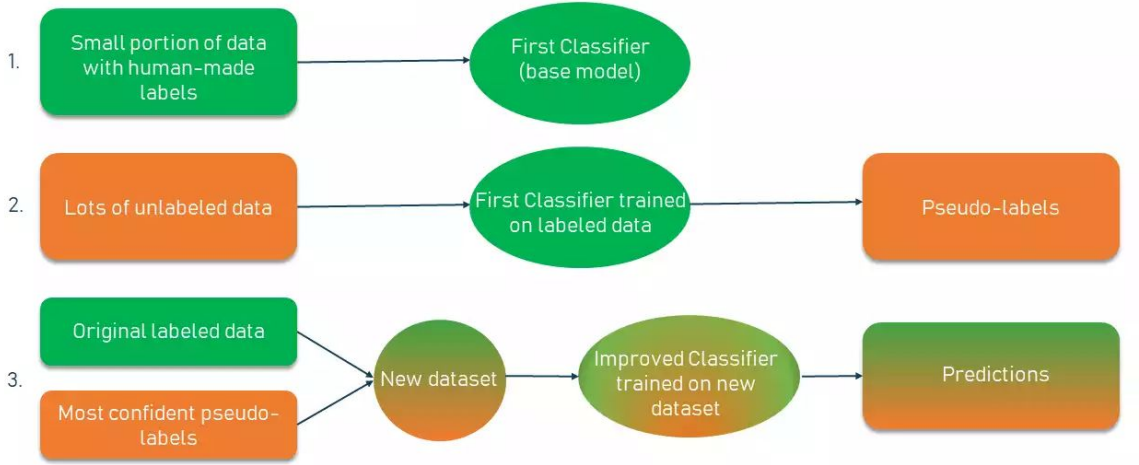
UNIVERSITY OF WESTMINSTER

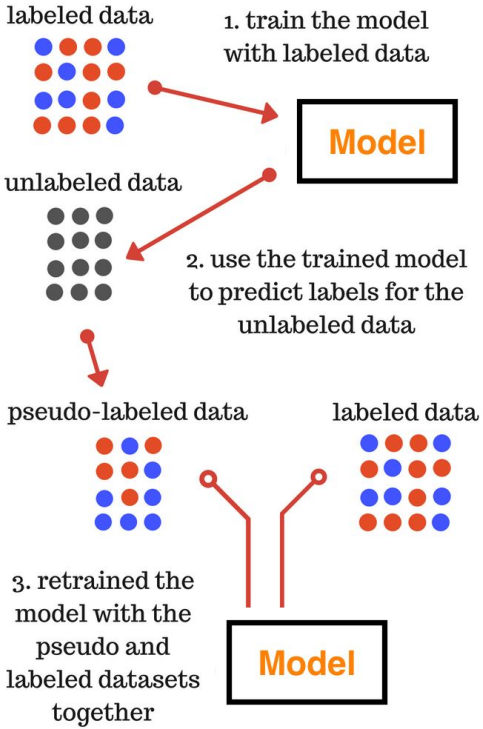
- **Semi-supervised learning (SSL)** is a machine learning technique that uses a small portion of labeled data and lots of unlabeled data to train a predictive model.
- It bridges supervised learning and unsupervised learning techniques to solve their key challenges. With it, you train an initial model on a few labeled samples and then iteratively apply it to the greater number of unlabeled data.
- Unlike unsupervised learning, SSL works for a variety of problems from classification and regression to clustering and association.
- Unlike supervised learning, the method uses small amounts of labeled data and also large amounts of unlabeled data, which reduces expenses on manual annotation and cuts data preparation time.



Semi-supervised Learning

UNIVERSITY OF WESTMINSTER





This unsupervised clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iteration.

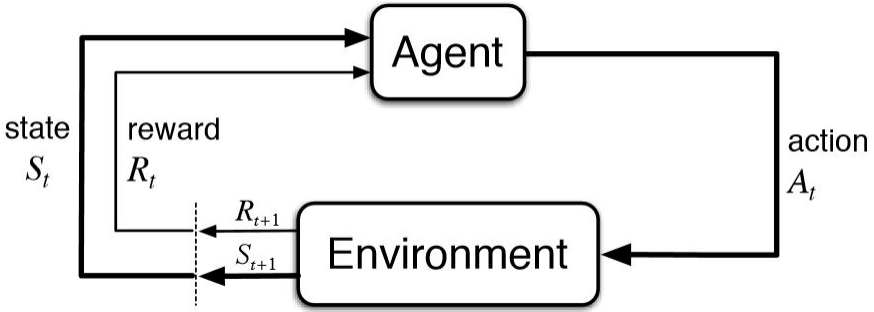
- Agglomerative clustering
- Hierarchical clustering
- K-means clustering

Reinforcement Learning

Reinforcement Learning

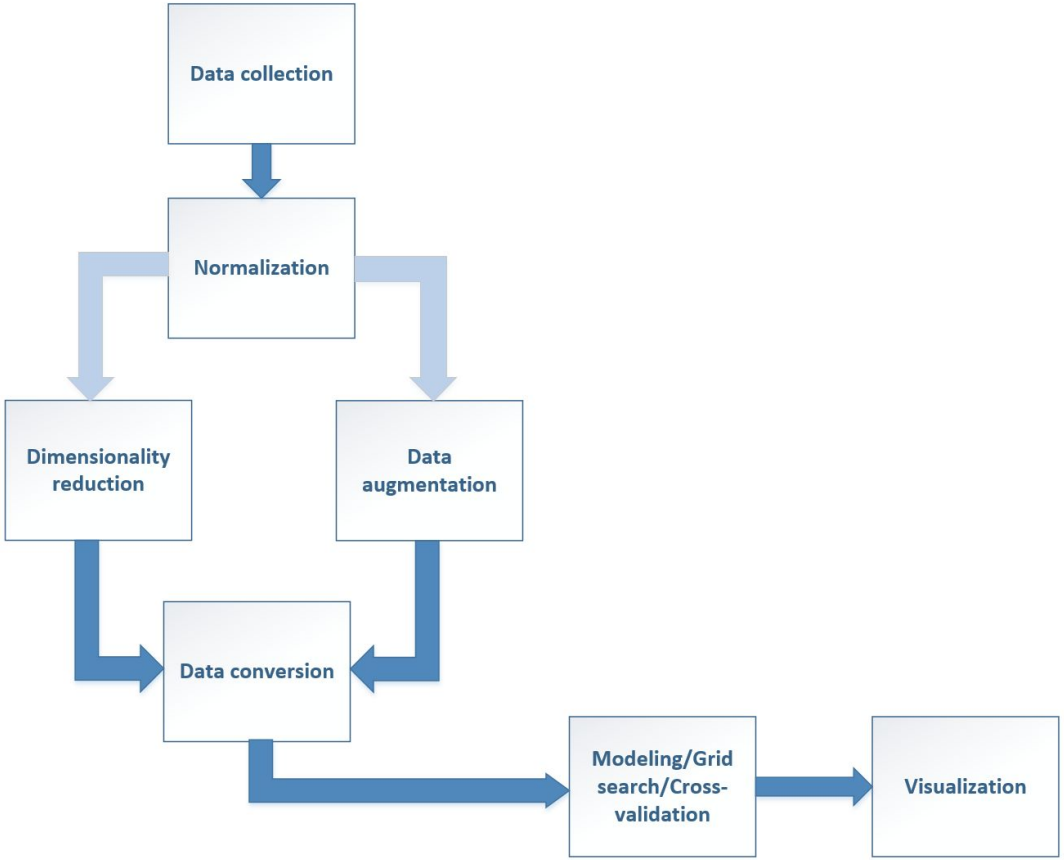
UNIVERSITY OF WESTMINSTER

- Reinforcement Learning(RL) is a type of machine learning technique that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences.
- As compared to unsupervised learning, reinforcement learning is different in terms of goals. While the goal in unsupervised learning is to find similarities and differences between data points, in the case of reinforcement learning the goal is to find a suitable action model that would maximize the total cumulative reward of the agent.



Building Machine Learning Architecture

UNIVERSITY OF WESTMINSTER



Building Machine Learning Architecture

UNIVERSITY OF WESTMINSTER

- **Data collection:**
  - The ideal situation is to have a Comma-Separated Values (CSV). A very famous and powerful Python library for data manipulation is pandas (part of SciPy).
  - It's mainly based on the concept of DataFrame (an abstraction of SQL table) and implements many methods that allow the selection, joining, grouping, and statistical processing of datasets that can fit in memory.
  - Data can be collected from different sources:
    - Dataset Search by Google
    - Kaggle
    - Stanford University
    - Visual Data Discovery
    - Amazon Datasets

Building Machine Learning Architecture

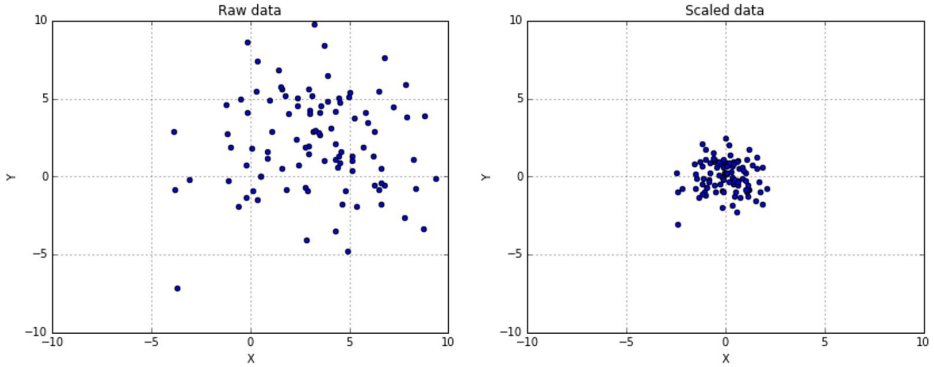
UNIVERSITY OF WESTMINSTER

- **Feature engineering/selection** is the first step in a machine learning pipeline and involves all the techniques adopted to clean existing datasets, increase their signal-noise ratio, and reduce their dimensionality.
    - **Filter Method:** In this method, features are dropped based on their relation to the output, or how they are correlating to the output. We use correlation to check if the features are positively or negatively correlated to the output labels and drop features accordingly.
    - **Wrapper Method:** We split our data into subsets and train a model using this. Based on the output of the model, we add and subtract features and train the model again. It forms the subsets using a greedy approach and evaluates the accuracy of all the possible combinations of features.
    - **Intrinsic Method:** This method combines the qualities of both the Filter and Wrapper method to create the best subset.

Building Machine Learning Architecture

UNIVERSITY OF WESTMINSTER

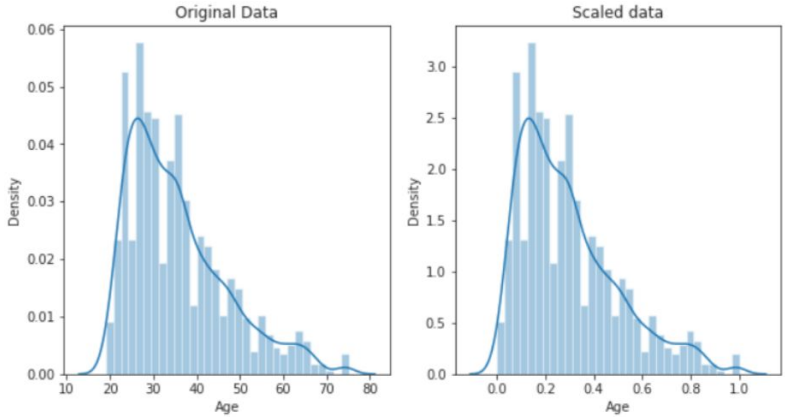
- **Managing missing features:** Sometimes, a dataset can contain missing features, so there are a few options that can be taken into account:
  - Removing the whole line (Suitable for very large datasets)
  - Creating a submodel to predict those features (It is complex as we need to create a new model to supervise each feature)
  - Using an automatic strategy to input them according to the other known values



Building Machine Learning Architecture

UNIVERSITY OF WESTMINSTER

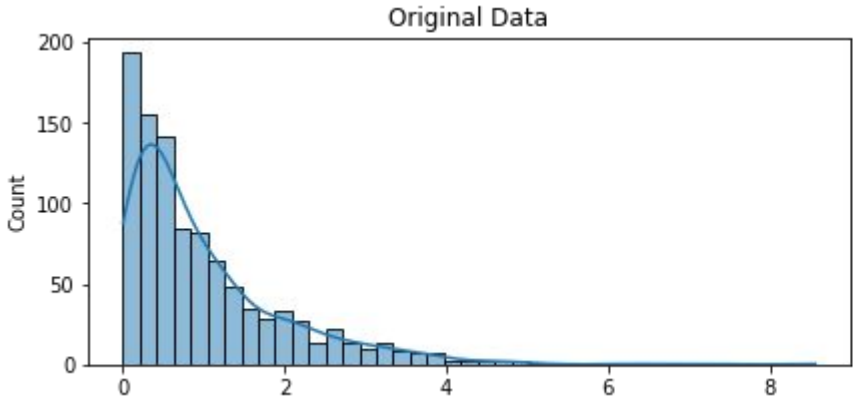
- **Data Conversion (Managing categorical values):** In many classification problems, the target dataset is made up of categorical labels that cannot immediately be processed by every algorithm. Therefore, it is very important to convert the categorical features of a dataset into numeric values. This is one of the primary reasons we need to pre-process the categorical data.
- **Data Scaling:** in scaling, we're transforming the data so that it fits within a specific scale, like 0-100 or 0-1. Usually 0-1. You want to scale data especially when you're using methods based on measures of how far apart data points are. Such as SVM, and K-means clustering.

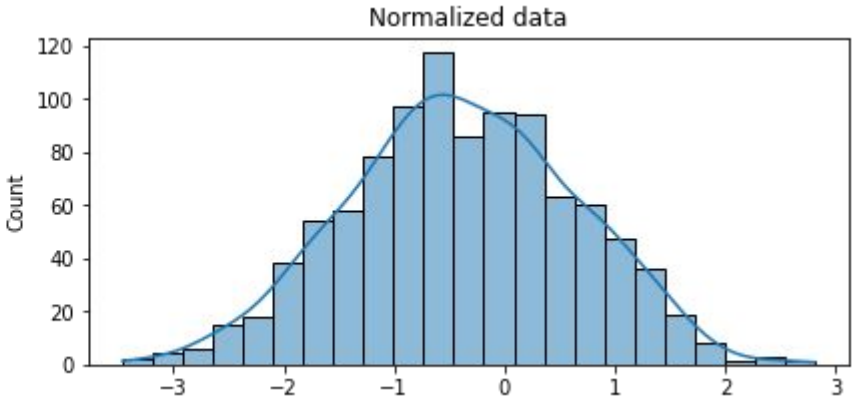


Building Machine Learning Architecture

UNIVERSITY OF WESTMINSTER

Data Normalization: Scaling just changes the range of your data. Normalization is a more radical transformation. The point of normalization is to change your observations so that they can be described as a normal distribution.





Building Machine Learning Architecture

UNIVERSITY OF WESTMINSTER

- **Dimensionality Reduction:**
  - The higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, most of these features are correlated, and hence redundant.This is where dimensionality reduction algorithms come into play.
  - Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be referred to the Feature Engineering/selection.
  - For example, six-dimensional space is difficult to visualise in a three-dimensional world, or on a two-dimensional plot.
  - Accordingly, we take advantage of a mathematical technique called Principal Component Analysis (PCA) to analyze the relationships between the features and summarize each observation as coordinates for two principal components - in other words, we'll translate the six-dimensional feature values into two-dimensional coordinates.

Building Machine Learning Architecture

UNIVERSITY OF WESTMINSTER

Data Augmentation:

- Data augmentation is a set of techniques used to increase the amount of data in a machine learning model by adding slightly modified copies of already existing data or newly created synthetic data from existing data. It helps smooth out the machine learning model and reduce the overfitting of data.
- Data augmentation can be utilized to address both requirements; the amount of data and the diversity of the training data needed to create an accurate machine learning model.

Building Machine Learning Architecture

UNIVERSITY OF WESTMINSTER

Model Optimization:

- Hyperparameter optimization in machine learning intends to find the hyperparameters of a given machine learning algorithm that deliver the best performance as measured on a validation set. Hence, for building a concrete model, we need to apply optimization methods.
- Hyperparameters, in contrast to model parameters, are set by the machine learning engineer before training
- Some examples of model hyperparameters include:
  - The C and γ hyperparameters for support vector machines.
  - The k in k-nearest neighbors.
- Optimisation methods:
  - **Manual Hyperparameter Tuning**
  - **Grid Search:** With this technique, we simply build a model for each possible combination of all of the hyperparameter values provided, evaluating each model, and selecting the architecture which produces the best results.
  - **Random Search:** Random Search sets up a grid of hyperparameter values and selects random combinations to train the model and score. This allows you to explicitly control the number of parameter combinations that are attempted.
  - **Gradient Descent**

Building Machine Learning Architecture

UNIVERSITY OF WESTMINSTER

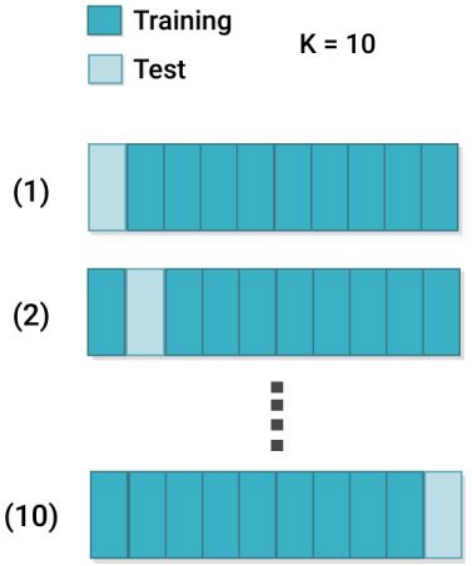
- **Cross Validation(CV)** is a technique for evaluating a machine learning model and testing its performance.
- CV is commonly used in applied ML tasks. It helps to compare and select an appropriate model for the specific predictive modeling problem.
- Generally speaking, the CV includes the following steps:
    a. Divide the dataset into two parts: one for training, other for testing
    b. Train the model on the training set
    c. Validate the model on the test set
    d. Repeat a-c steps a couple of times. This number depends on the CV method that you are using

Building Machine Learning Architecture

k-Fold cross-validation: k-Fold introduces a new way of splitting the dataset which helps to overcome the “test only once bottleneck”.

The algorithm of the k-Fold technique:

1. Pick a number of folds – k. Usually, k is 5 or 10 but you can choose any number which is less than the dataset’s length.
2. Split the dataset into k equal (if possible) parts (they are called folds)
3. Choose k – 1 folds as the training set. The remaining fold will be the test set
4. Train the model on the training set. On each iteration of cross-validation, you must train a new model independently of the model trained on the previous iteration
5. Validate on the test set
6. Save the result of the validation
7. Repeat steps 3 – 6 k times. Each time use the remaining fold as the test set. In the end, you should have validated the model on every fold that you have.
8. To get the final score average the results that you got on step 6.



Tools


Pandas
PyTorch
NumPy
plotly
matplotlib
theano
TensorFlow
scikit learn
Keras
python
Natural Language Analyses with NLTK

Conclusion

UNIVERSITY OF WESTMINSTER

- Machine Learning is a subset of AI that allows computers to learn from data and paste experiences.
- There are different types of machine learning: Supervised, unsupervised and Reinforcement Learning.
- In Supervised learning algorithms, a machine uses labeled data to build a model and predict the outcome. Those types of algorithms fall into two categories: regression and classification.
- In Unsupervised learning, we use unlabeled data and predict the output using clustering methods.
- In reinforcement learning, we enforce the machine to learn its environment using the reward and policy strategies.

References

Fundamentals of Machine Learning | Oxford Academic (oclc.org)

scikit-learn : Machine Learning Simplified (oreilly.com)

Machine Learning Algorithms (oreilly.com)

Python Data Analytics: With Pandas, NumPy, and Matplotlib (oreilly.com)

UNIVERSITY OF WESTMINSTER

Questions?