Machine learning

Algorithms →
- Linear Regression
- logistic Regression
- K-Means clustering
- Decision Tree
- Random Forest
- KNN Algorithm
- Support Vector Machine
- Naive Bayes Classifier
- Hierarchical Clustering

Learnings ↓
- Supervised → Labelled data
- Unsupervised → Unlabelled data (Recognize hidden patterns)
- Reinforcement → Reward system.

Input → model → Output

Applications :- Healthcare, Sentiment Analysis, Fraud Detection, E-commerce

Definition :- Machine learning is the science of making computers learn and act like humans by feeding data and information without being explicitly programmed.

Steps Involved :- Define Objective → Collect Data → Prepare Data → Select Algorithm → Train Model → Test Model → Predict → Deploy

Regression → Predict Quantity

Classification → To predict category

Anamoly Detctn → Detect Anamoly

Clustering → Discover structure in unexplored data

*Linear Regression

Independent Variable → Is not affected by others (Rain)
Dependent Variable → Is affected by others. (Crops)

Application → Economic growths, Product Pricing, Housing Sales, Score Predn.

Defn :- Linear Regression is a statistical model used to predict the relationship between independent and dependent variables.

Which variables in particular are significant predictors of the outcome variables.

How significant is the Regression line to make predictions with highest possible accuracy

Eqn :- $y = mx + c$

y :- Dependent
x :- Independent

m :- Slope $= \frac{y_2 - y_1}{x_2 - x_1}$

$$m = \frac{((n * \Sigma(x*y)) - (\Sigma(x) * \Sigma(y))}{((n * \Sigma(x^2)) - (\Sigma(x)^2)}$$

$$c = \frac{((\Sigma(y) * \Sigma(x^2)) - (\Sigma(x) * \Sigma(x*y))}{((n * \Sigma(x^2)) - (\Sigma(x)^2)}$$

For Errors To Minimize the Distance of line & Data Points / Finding the Best fit line

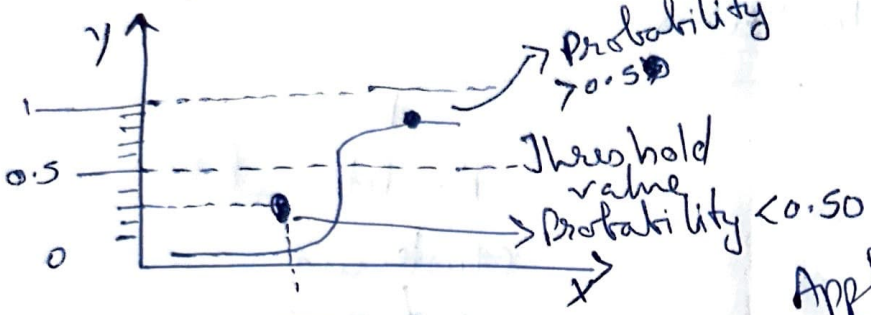
↳ Use → Sum of Sq errors, Sum of Absolute errors, Root Mean Sq errors, etc

Multiple Linear Regression

$$Y = m_1 * x_1 + m_2 * x_2 + m_3 * x_3 + \ldots + m_n * x_n + c$$

* Logistic Regression

Defn: - It is a classification algorithm, used to predict binary outcomes for a given set of independent variables. The dependent variable's output is discrete.



Eqn of sigmoid func

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

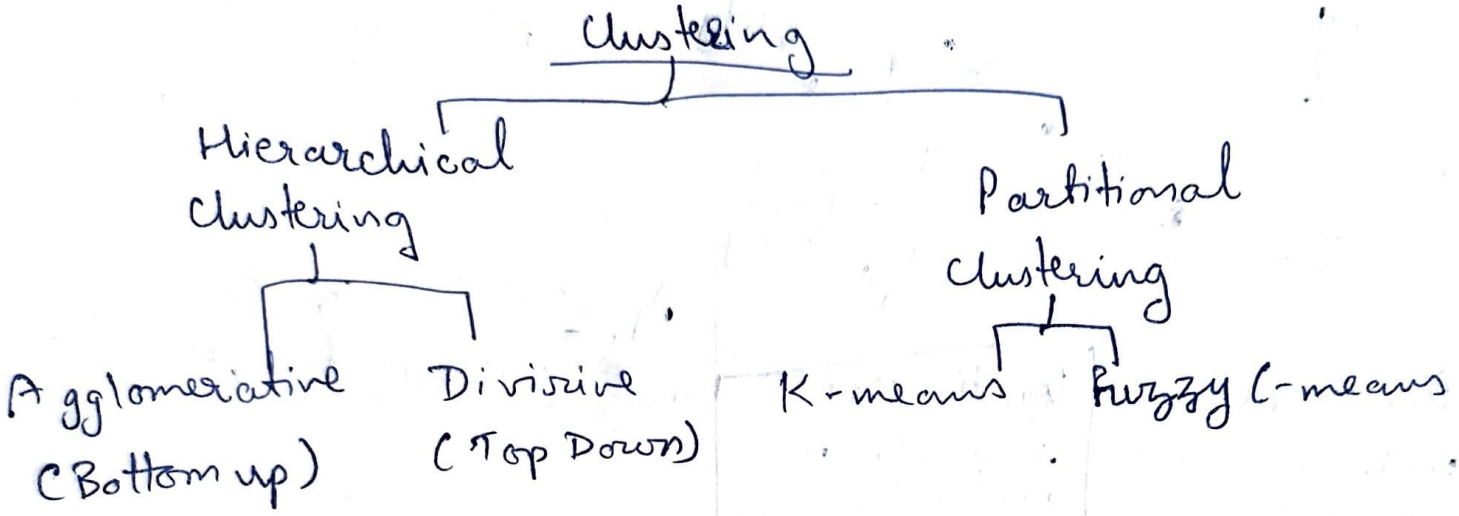
$$\hookrightarrow \log\left(\frac{p(x)}{1-p(x)}\right)$$

Linear Regn	Logistic Regn
• Solves Regression problems	Solves Classification Problems
• Respons variable continuous in nature	Response variable categorical in nature
• Estimates the dependent variable in the independent variable	• Helps calculating the possibility of an event happening
• Straight line	• S-curve.

Applications: - Weather Predcn
Img categorization, Healthcare

* K-Means Clustering

Defn: K-means performs division of objects into clusters which are "similar" between them and are "dissimilar" to the objects belonging to another cluster.



Applns:- Academic Performance, Diagnostic Systems, Search Engines, Wireless Sensors Networks, Colour compression

Euclidean Distance Measure:- $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Working:

start
↓
Elbow Point (k)
↓
→ measure the Distance
↓
⊖ Grouping based on minimum distance
↓
Reposition the centroids
↑ IF clusters are stable
Convergence
⊕
IF clusters are unstable

Algorithm

Assuming we have inputs x1, x2, x3, ... and value of k,

Step1: Pick k random points as cluster centers called centroids.

Step2: Assigns each xi to nearest cluster by calculating its distance to each centroid

Step3: Find new cluster center by taking the average of the assigned points.

Step4: Repeat steps 2 & 3 until no of the clusters arrignments change.

* Decision Tree

Problems that Decision Tree can Solve :-

1) Classification.

A classification tree will determine a set of logical if-then cond's to classify problems

2) Regression

Regression tree is used when the target variable is numerical or continuous in nature.

Advantages

1) Simple to Understand, interpret and visualize
2) Little effort required for data preparation
3) Nonlinear parameters don't effect its performance.

Disadvantages

1) Overfitting occurs when the algorithm captures noise in the data.
2) The model can get unstable due to small variation in data
3) A highly complicated Decision tree tends to have a low bias which makes it difficult for the model to work with new data

IMP Terms

Entropy - It is the measure of ~~decrease in entropy after~~ randomness or unpredictability in the dataset

Information Gain - It is the measure of decrease in entropy after the dataset is split

Leaf Node - Carries the classification or the Decision

* Random Forest

Applⁿ - Remote Sensing, Object Detection, Kinect (Gaming)

Why Random Forest?

1) No overfitting
2) High Accuracy
3) Estimates Missing Data

What is Random Forest?

Random Forest or Random Decision Forest is a method that operates by constructing multiple Decision Trees during training phase. The decision of majority of the trees is choosen by the random forest as the final decision

---

* KNN

Defⁿ :- K Nearest Neighbors, is one of the simplest Supervised Machine Learning algorithm mostly used for classification

How do we choose the factor 'k'?

KNN Algorithm is based on feature similarity: Choosing the right value of K is a process called parameter tuning, and it is important for better accuracy

- To choose value of k:-

→ Sqrt(n), where n is the total number of data points
→ odd value of k is selected to avoid confusion between two classes of data

When do we use KNN?

1) Data is labelled
2) Data is noise free
3) Dataset is small.

---

* Support Vector Machine

Appln: Face Detection, Text & Hyper Text Categorization, Classification of images, Bioinformatics.

Why SVM?

SVM is a supervised learning method that looks at data and sorts it into one of the two categories

Advantages

1) High Dimensional Space
2) Sparse documents vectors.
3) Regularization parameters.

---

* Naive Bayes Classifier

* Bayes Thm - $P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$

Where is Naive Bayes Used?

1) Face Recognition
2) Weather Prediction
3) Medical Diagnosis
4) News Classification

Advantages

1) Very simple & Easy to implement
2) Needs less training
3) Handles both continuous & discrete data
4) Highly scalable with numbers of predictors & data points
5) As it is fast, it can be used in real time predictions
6) Not sensitive to irrelevant features.

Applications of Machine Learning

- Virtual Assistant
- Traffic Prediction
- Social media Personalization
- Email Spam Filtering
- Online Fraud Detection
- Assistive Medical Tech
- Automatic Translation

What does ML Engineer do?

- Create & maintain ML solutions to solve business problems
- Optimizes these solutions for performance & scalability
- Solves business problems like reducing customer churn, improving targeted marketing campaigns and improving product experiences
- Contribute to cutting edge research in AI & ML.

Math Required for ML (Important Topics)

- Probability & Statistics - Bayes Theorem, Probability distribution, Sampling, Hypothesis testing
- Linear Algebra - Matrices, Vectors
- Calculus - Differential calculus, Integral Calculus

* Applications of Machine Learning

1) Virtual Assistans
2) Traffic Predictions
3) Social Media Personalization
4) Email Spam Filtering
5) Online Fraud Detection
6) Assistive Medical Tech
7) Automatic Translation

* What does ML Engineer Do?

- Creates & maintains ml solutions to solve buriness problems
- Optimizes there solutions for performance & scalability
- Solves buriness problems like reducing customers churn, running targeted marketing campaigns and improving product experience.
- Contributes to cutting edge rereards in ~~ai~~ AI & ML

* Math Required for ML (IMP Topics)

* Probability & Statistics - Bayes Theorem, Probability Distribution, Sampling, Hypothesis testing
* Linear Algebra - Matrices, Vectors
* Calculus - Differential calculus, Integral Calculus.