Machine Learning Notes

1. What is ML?
   - Machine learning is growing technology which enables computers to learn automatically from past data.
   - Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information.
   - It is subset of AI.
   - Types of ML :
     - i.   Supervised
     - ii.   Unsupervised
     - iii. Reinforcement
     - iv.   Semi-supervised

2. What are different types of ML?
   - i. Supervised learning :
     - a) As per name supervised learning is based on supervision.
     - b) It means in supervised learning, we train the machines using "labelled" dataset and based on training, machine predicts output.
     - c) Main goal of supervised learning technique is to map the input variable (x) with output variable (y).
     - d) Applications :
       Risk assessment
       Fraud detection
     - e) Types :
       Classification
       Regression
   - ii. Unsupervised learning :
     - a) Unsupervised learning is different from supervised learning, as its name suggests, there is no need for supervision.
     - b) In unsupervised learning, machine is trained using unlabeled data & predicts output without supervision.
     - c) Types :
       Clustering
       Association
     - d) Applications :
       Network analysis
       Anomaly detection
   - iii. Semi-supervised learning :
     - a) Semi-supervised learning is type of machine learning algorithm that lies between supervised & unsupervised learning.
     - b) To overcome the drawbacks of supervised and unsupervised learning, the concept of semi-supervised is introduced.
     - c) Advantages :
       Simple & easy
       Highly efficient
     - d) Disadvantages :
       Accuracy low
       Results may not be stable

iv. Reinforcement learning :
   a) Reinforcement learning works on feedback process, in which an AI agent automatically explore its surrounding, taking action, learning from experiences, improving its performance.
   b) Types :
      Positive reinforcement
      Negative reinforcement
   c) Applications :
      Video games
      Robotics

3. What are the applications of ML?
   i. Image recognition :
      a) Image recognition is one of the most common application of machine learning.
      b) It is used to identify objects, persons, places, digital images, etc.
   ii. Speech recognition :
      a) While using google, we get on option of "search by voice", it comes under speech recognition and it is popular application of ML.
      b) Speech recognition is process of converting voice into text and it is also known as "Speech to text".
   iii. Traffic prediction :
      a) If we want to visit new place, we take help of google maps which shows correct map with shortest route & predicts traffic conditions.
   iv. Product recommendation :
      a) Machine learning is widely used by various e-commerce & entertainment companies such as Amazon, Netflix, etc. for product recommendation.
      b) When we search some product on amazon, then we start getting advertise for same product on same browser & this is because of ML.
   v. Stock market trading :
      a) ML is widely used in stock market trading.
      b) In stock market, there is always risk of ups & downs in shares, so for this machine learning long short term memory neural network is used for prediction.

4. Difference between Overfitting and Underfitting.

	Parameters	Overfitting	Underfitting
1	Model	More complex model	More simple model
2	Regularization	Less	More
3	Quantity of features	Larger	Smaller
4	Complexity	Need to reduce complexity	Need to increase complexity
5	Training set	Accurate for training set	Not accurate for training set

5. Explain different issues in ML.
   a) Inadequate training data :
      i) The major issue that comes while using ML algorithms is the lack of quality as well as quantity of data.
      ii) Data quality can be affected by some factors as follows :
         ➤ Noisy data
         ➤ Incorrect data
         ➤ Generalizing the output data.
   b) Poor quality of data :
      ➤ Data plays an important role in ML and it must be of good quality as well.
      ➤ Noisy data, incomplete data, inaccurate data and uncleaned data lead to less accurate results.

c) Overfitting :
   ➢ It is one of the most common issues faced by ML engineers and data scientists.
   ➢ When ML model is trained with a huge amount of data, it starts capturing noise and inaccurate data.
d) Underfitting :
   ➢ It is just opposite of overfitting.
   ➢ When ML model is trained with **forever** amount of data, then it starts capturing noise & inaccurate data.
e) Slow implementations & results :
   ➢ This is also very commonly seen in machine learning models.
   ➢ ML models are highly efficient in producing accurate results but are time consuming.
f) Lack of skilled resources :
   Although AI & ML are continuously growing in market, still these industries are freshers in comparison to others.

6. How to choose right ML algorithm?
   There are different algorithms available in ML. For choosing correct first you need to focus on your good and second you have to consider data.
   Goal :
   a) If you have chosen surprised learning, then next you need to focus on what is your target value ?
      ➢ If target value is discrete, then use classification.
      ➢ If target value is continuous, then use regression.
   b) If you have chosen unsupervised learning, then next you need to focus of what is your aim ?
      ➢ If you want to fit your data into discrete groups, then use clustering.
      ➢ If you want to find numerical estimate, then use density estimation algorithm.

   Data :

	Supervised learning	Unsupervised learning
Discrete	Classification	Clustering
Continuous	Regression	Density estimation

7. Explain the following terms : Regression line, scatter plot, error in prediction, best fitting line.
   a) Regression line :
      ➢ A regression line is straight line that represents the relationship between two variables in statistical model.
      ➢ In simple linear regression, there are two variables, independent variable (x), dependent variable (y).
   b) Scatter plot :
      ➢ A scatter plot is graphical representation of relationship between two variables.
      ➢ Each point on the plot represents a pair of values for two variables.
      ➢ Scatter plots are useful for visualizing pattern, direction, strength of relationship between variables.
   c) Predication error :
      ➢ The difference between observed values of dependent variable and values predicted by regression line is called as prediction error.
      ➢ Also known as residuals.

   d) Best fitting line :
      ➢ The best fitting line is regression line that minimizes the sum of squared prediction errors.
      ➢ This approach is known as method of least squares.

   Summary:

      ➢ Regression lines, scatter plots, errors in prediction, best fitting lines are essential concepts in regression analysis.
      ➢ This facilitates understanding & prediction of relationship between two variable

8. Step in developing a machine learning application :
   a) Collection of data :
      You can collect samples from website & extracting data.
      - From publically available data.
      - From RSS feed or API
   b) Preparation of input data :
      - Once you have input data, you need to check whether it is useable or not.
      - Some algorithm can accept target variable as string, some need them to be integers.
   c) Analyze the input data :
      - Looking at data you have passed in text editor to check collection and preparation of input data.
      - Distil multiple dimensions down to 2/3 so that you can visualize data.
   d) Train algorithm :
      - Good clean data from first two steps is given as input to algorithm.
      - The algorithm extracts information or knowledge.
      - The knowledge is mostly stored in format that is useable.
   e) Test the algorithm :
      - In this step information learned in previous step is used.
      - When you are checking algorithm, you will test it to find out whether it works properly or not.
   f) Use it :
      - In this step real program is developed to do some task & once again it is checked if all the previous steps worked as expected.

9. Explain Logistic regression.
   - Logistic regression is popular machine learning algorithm used for binary classification problems.
   - It is widely used in various fields like healthcare finance marketing & more.
   - Despite its name, logistic regression is used for classification not regression.
   1. Binary classification :
      - Logistic regression is primarily designed for binary classification problems. Where goal is to predict whether an instance belongs to one of two classes.
      - Ex. spam or not spam, fraudulent or not fraudulent.
   2. Sigmoid function :
      The logistic function also known as sigmoid function is crucial component of logistic regression.
      It is defined as : $6(z) = \dfrac{1}{1 + e^{-z}}$
   3. Hypothesis function :
      Hypothesis function in logistic regression is formulated using sigmoid function : $6\,(\theta^T\,x)$
   4. Decision boundary :
      Decision boundary is critical concept in logistic regression.
      It separates instances of different classes based on learned model.
   5. Training process :
      The model is trained using optimization algorithms like gradient descent or advanced optimization method.

      Steps :

      1) Data collection
      2) Data preprocessing
      3) Model initialization
      4) Hypothesis calculation
      5) Cost calculation
      6) Training iterations
      7) Model evaluation
      8) Prediction

Types :

1) Binomial
2) Multinomial
3) Ordinal

10. Explain concept of margin and support vector in ML.
   - The concepts of margin and support vector are fundamental to understanding support vector are fundamental to understanding support vector machines (SVM).
   - SVM is popular machine learning algorithm used for classification & regression tasks.

Margin:

   - Margin in SVM refers to distance between the decision boundary and nearest data point.
   - In binary classification the margin is width of "corridor" that algorithm is typing to maximize.
   - A larger margin generally leads to better performance as it provides better separation between classes.

Support vectors:

Support vectors are the data points that lie closest to decision boundary and play crucial role in defining margin.

These are instances that are most challenging to classify correctly.

Support vectors effectively "support" the structure of SVM model.

11. Explain Linear regression.
   - Linear regression is supervised learning algorithm used for predicting a continuous target based on one or more independent features.
   - It assumes the relationship between input features & target variable.
   - The goal is to final best-fitting line that minimizes the differences between predicted & actual various.

A) Simple linear regression :

Relationship in simple linear regression is represented as,

$Y = \beta 0 + \beta 1\, X + \epsilon$

Where, Y is target variable

X is independent variable

$\beta 0$ is Y intercept

$\beta 1$ is slope

ϵ is error term

B) Multiple linear regression :

Relation in multiple linear regression is re   as,

$Y = \beta 0 + \beta 1\, X1 + \beta 2\, X2 + ….. + \beta n\, Xn + \epsilon$

Where X1, X2, X3 are independent variable.

B0, $\beta 1$, βn are corresponding coefficients.

Steps:

1) Data collection :
   Gather data set containing target variable & independent features.
2) Data preprocessing :
   Handle missing values, scale or normalize features if needed.
3) Model initialization :
   Initialize linear regression model with initial values for coefficients.
4) Model training :
   Use training data to fit model & find optimal values for coefficients.

5) Model evaluation :
   Assess model's performance on testing set to evaluate its generalization to new unseen data.
6) Prediction :
   Once trained model can make predictions on new data by applying learned coefficients.

12. Explain Multiclass classification.
   ➢ In machine learning, multiclass or multinomial classification is problem of classifying instances into one of the three or more classes.
   ➢ Many classification algorithms, especially multiclass logistic regression, permit the use of more than two classes.
   ➢ Popular algorithm for multiclass classification :
      i)   K-nearest neighbors
      ii)   Decision trees
      iii) Naïve bayes
      iv)   Random forest
   ➢ Method of multiclass classification :
      i)   Load dataset from source
      ii)   Load dataset into 'training' and 'test' data
      iii) Train decision tree, SVM & KNN classifiers on training data
      iv)   Use above classifiers to predict labels for test data
      v)   Measure accuracy & visualize classification
   ➢ Types of multiclass classification :
      Transformation to binary :
      Reducing problem of multiclass classification to multiple binary classification can be categorized into :
      i)   One – vs – all :
            One – vs – all strategy involves involves training single classifier per class, with samples as that class as possible and all other samples as negatives.
      ii)   One – vs – one :
            In one –vs – one reduction, one trains $\frac{K(K-1)}{2}$ binary classifiers for k-way Multiclass problem, each receives samples of pair classes from original training set & must learn to distinguish these two classes.

13. Explain random forest algorithm.
   ➢ Random forest is a popular machine learning algorithm that belongs to supervised learning technique.
   ➢ It can be used for both classification and regression problems in ML.
   ➢ It is based on concept of ensemble learning, which is process of combining multiple classifiers to solve a complex problem & improve performance of model.
   ➢ As per name "random forest is classifiers that contains number of decision tree on various subsets of given dataset.
   ➢ Instead of relying on one decision tree, random forest takes prediction from each tree & then predicts the final output.
   ➢ The greater number of trees in forest leads to higher accuracy & prevents proble of over fitting.
   ➢ Working :
      Step 1 : Select random K data points from training set.
      Step 2 : Build decision trees   with selected data points.
      Step 3 : Choose number N for decision trees that you want to build
      Step 4 : Repeat step 1 & 2
      Step 5 : for new data points find predictions for each decision tree & assign new data points to category that wins majority votes.
   ➢ Applications :
      i)   Banking
      ii)   Medicine

      iii)   Marketing

      iv)   Land use

- ➢ Advantages :
  - i)   Robustness – less prone to over fitting
  - ii)   Handles missing values – can handle missing values
  - iii)   Parallelization
- ➢ Disadvantages :
  - i)   Complexity – challenging, complex, hard
  - ii)   Computational cost – expensive

14. Explain EM algorithm.
- ➢ EM algorithm stands for expectation – maximization algorithm.
- ➢ The EM algorithm is considered latent variable model to find local maximum likelihood parameters of statistical model.
- ➢ The EM algorithm is defined as combination of various unsupervised machine learning algorithm, which is used to determine local maximum likelihood estimates (MLE)
- ➢ EM algorithm consist of two models : i) E – step   ii) M – step
- ➢ In first mode, use estimate the latent variables. Hence it is referred as expectation / estimation step (E – step)
- ➢ Second mode is used to optimize parameters of models so that it can explain data more clearly. Second mode is known as maximization step (M – step)
- ➢ Steps :
  Step 1 : Initialize parameter values.
  Step 2 : This step is known as E – step which is used to estimate values of missing or incomplete data using observed data.
- ➢ Step 3 : This step is known as M – step where we use complete data obtained from step 2 to update parameter values.
- ➢ Step 4 : Last step is to check whether values of latent variable are converging or not. if yes then stop and if no then repeat process from step 2.
- ➢ Applications :
  - i)   NLP (natural language processing)
  - ii)   Gaussian mixture model (GMM)
  - iii)   Data clustering in ML
  - iv)   Medical & healthcare
- ➢ Advantages :
  - i)   Easy
  - ii)   Guaranteed results
- ➢ Disadvantages :
  - i)   Convergence is slow
  - ii)   Make convergence for local optima only

15. Explain DBSCAN.
- ➢ DBSCAN stands for density based spatial clustering of applications with noise.
- ➢ DBSCAN is a clustering algorithm used in machine learning for identifying dense regions of data points in dataset.
- ➢ Like k-means, DBSCAN does not require predefined number of clusters.
- ➢ Key idea behind DBSCAN is to group together data points that are closely packed.
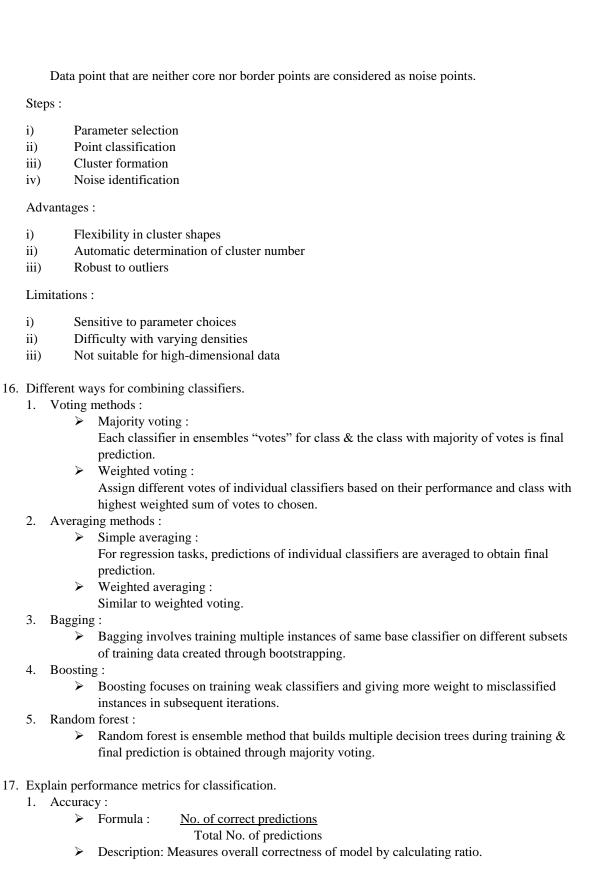
Key concepts :

1. Care points :
   A data point is core point if it has at least a specified number of data points (MinPts) within specified radius (EPS)
2. Border points :
   A data point is border point if it is within the specified radius (EPS) but does not   enough neighbors to be considered as core point.
3. Noise points :

Data point that are neither core nor border points are considered as noise points.

Steps :

i)   Parameter selection
ii)   Point classification
iii)   Cluster formation
iv)   Noise identification

Advantages :

i)   Flexibility in cluster shapes
ii)   Automatic determination of cluster number
iii)   Robust to outliers

Limitations :

i)   Sensitive to parameter choices
ii)   Difficulty with varying densities
iii)   Not suitable for high-dimensional data

16. Different ways for combining classifiers.
   1. Voting methods :
      ➢ Majority voting :
        Each classifier in ensembles "votes" for class & the class with majority of votes is final prediction.
      ➢ Weighted voting :
        Assign different votes of individual classifiers based on their performance and class with highest weighted sum of votes to chosen.
   2. Averaging methods :
      ➢ Simple averaging :
        For regression tasks, predictions of individual classifiers are averaged to obtain final prediction.
      ➢ Weighted averaging :
        Similar to weighted voting.
   3. Bagging :
      ➢ Bagging involves training multiple instances of same base classifier on different subsets of training data created through bootstrapping.
   4. Boosting :
      ➢ Boosting focuses on training weak classifiers and giving more weight to misclassified instances in subsequent iterations.
   5. Random forest :
      ➢ Random forest is ensemble method that builds multiple decision trees during training & final prediction is obtained through majority voting.

17. Explain performance metrics for classification.
   1. Accuracy :
      ➢ Formula : $\dfrac{\text{No. of correct predictions}}{\text{Total No. of predictions}}$
      ➢ Description: Measures overall correctness of model by calculating ratio.

   2. Precision :
      ➢ Formula: $\dfrac{\text{True positives}}{\text{True positives + False positives}}$
      ➢ Description: Measures accuracy of positive predictions.
   3. Recall :
      ➢ Formula: $\dfrac{\text{True positives}}{\text{True positives + False negatives}}$

➤ Description: Measure ability of model to capture all positive instances.
4. Specificity :
➤ Formula: True Negatives   .
              True Negatives + False positives
➤ Description: Measures ability of model to correctly identity negative instances.
5. False positive rate (FPR)
➤ Formula: False positives   .
              False positive + true negative
➤ Description: Measure proportion of actual negative instances that were incorrect as positive.
6. False negative rate (FNU) :
➤ Formula: False negative   .
              False negative + True positive
➤ Description: Measures proportion of actual positive instances that were incorrect as negative.

18. What is dimensionality reduction.
➤ Dimensionality reduction is a technique used to reduce the number of features in dataset while retaining as much as important information as possible.
➤ It is process of transforming high-dimensional data into lower-dimensional space that still preserves essence of original data.
➤ High dimensional data refers to data with large number of features or variables.
➤ Dimensionality reduction can help to mitigate these problems by reducing complexity of model.
➤ There are two components of dimensional reduction :
  i)   Feature selection :
        It involves 3 ways :
        a) Filter
        b) Wrapper
        c) Embedded
  ii)   Feature extraction :
        This reduces data in high dimensional space to lower dimensional space.
➤ Methods of dimensionality reduction :
  i)   Principal component analysis (PCA)
  ii)   Linear discriminant analysis (LDA)
  iii) Generalized discriminant analysis (GDA)
➤ Advantages :
  i)   Helps in data compression
  ii)   Reduces computation time
➤ Disadvantages :
  i)   May lead to data loss
  ii)   Overfitting

19. Explain PCA algorithm.
➤ Principal component analysis is an unsupervised learning algorithm that is used for dimensionality reduction in ML.
➤ It is statistical process that converts observations of correlated features into set of linearly uncorrelated features.
➤ This newly transformed features are called as principal components.
➤ PCA is based on following mathematic components :
  i)   Variance & covariance
  ii)   Eigenvalues & eigenvectors
➤ Applications of PCA
  i)   Image processing
  ii)   Movie recommendation system
  iii) Computer vision

- ➤ Steps for PCA also
    - i) Getting dataset
    - ii) Representing data into structure
    - iii) Standardizing data
    - iv) Calculating
    - v) Calculating eigen values & eigen vectors
    - vi) Sorting eigen vectors
    - vii) Calculating new features principal components
    - viii) Remove less or unimportant features from new dataset.
- ➤ Common terms in PCA :
    - i) Dimensionality : It is number of features or variables present in given dataset. More easily no of columns present in dataset.
    - ii) Correlation : It signifies how strongly two variables are related to each other.
    - iii) Orthogonal : It defines that variables are not correlated to each other hence correlation between variables is O.
- ➤ Covariance matrix : It is matrix containing covariance between pair of variables.

20. Explain LDA.
- ➤ Linear discriminant analysis is one of the commonly used dimensionality reduction techniques in machine learning to solve classification problems.
- ➤ It is also known as normal discriminant analysis (NDA)
- ➤ It is used to project the features in higher dimension space into lower dimension space.

Advantages of LDA :

i) Simple & efficient
ii) Can handle multicollinearity

Disadvantages of LDA :

i) Sensitive to outliers
ii) Assumes normality and equal covariance matrices, which may not hold all dataset.

Applications :

i) Face recognition
ii) Medical
iii) Customer identification
iv) Predictions

21. Explain bagging.
- i) Bootstrap aggregating, also known as bagging is machine learning ensemble meta-algorithm designed to improve, stability and accuracy of machine learning algorithms used in statistical classification & regression.
- ii) It is usually applied to decision tree methods.
- iii) Bagging is special case of model averaging approach.

Implementation steps of bagging :

i) Multiple subsets are created from original data set with equal tuples selecting observations with replacement.
ii) A base model is created on each of these subsets.
iii) Each model is learned in parallel with each training set and independent of each other.
iv) The final predictions are determined by combining predictions from all models.

Example :

The random forest model uses bagging where decision tree models with higher variance are present.

Several random trees make a random forest.

22. Explain boosting.
    i) Boosting is an ensemble modelling technique that attempts to build strong classifier from the number of weak classifiers.
    ii) It is done by building a model by using weak models in series.
    iii) Firstly a model is built from the training data. Then second model is built which tries to correct errors present in first model.
    iv) This procedure is continued & model are added until either the complete training dataset is predicted correctly or maximum no. of models are added.
    v) Boosting algorithm :
       ➢ Adaboost was first really successful boosting algorithm developed for purpose of binary classification.
       ➢ Adaboost is short for adaptive boosting & is very popular boosting technique that combines multiple "weak classifiers" into single "strong classifier".
    vi) Similarities between bagging & boosting :
       ➢ Both are ensemble methods to get N learners from 1.
       ➢ Both generate several training data sets by random sampling.
       ➢ Both are good at reducing variance & provide higher stability.