**GSOC Notes**

**Hypothesis**

1. Hypothesis Testing is basically drawing conclusions about a population based on sample data
2. There are 2 type of hypothesis testing:
   1. Null hypothesis:
      1. It is also called as ground truth
      2. The null hypothesis is the default assumption or the claim that there is no significant difference or relationship between variables in the population.
   2. Alternative hypothesis:
      1. The alternative hypothesis, on the other hand, suggests that there is a significant difference or relationship.
      2. It basically opposes the null hypothesis
3. In Hypothesis testing we determine the significance level (alpha) for the test, which is the probability of rejecting the null hypothesis when it is true

**P – Value**

1. The p-value is a measure of the evidence against the null hypothesis
2. It ranges between 0 and 1
3. If the p-value is less than equal to the alpha value that is 0.005 then we reject the null hypothesis
4. If the p-value is greater than equal to the alpha value that is 0.005 then we accept the null hypothesis

**Different techniques**

1. **T – Test:**
   1. The t-test is a statistical test used to compare the means of two groups and determine if they are significantly different from each other.
   2. Types of t-test:
      1. **One sample t-test:**
         1. It is a statistical test used to determine if the mean of a single sample significantly differs from a known population mean.
         2. ttest\_1samp() is the function required to perform one sample test
         3. Eg: Average age is 30
      2. **Independent T-Test:**
         1. An independent t-test, also known as a two-sample t-test, is used to compare the means of two independent groups and determine if they are significantly different from each other.
         2. Eg: Comparing the test scores of two different groups of students.
      3. **Paired Samples T-Test:** 
         1. This test is used when you have paired or matched observations within the same group.
         2. It compares the means of the differences between paired observations
         3. Eg: Bp measured before medication and measured after medication
2. **Z – Test:**
   1. The z-test is a statistical test that is used to determine whether a sample mean significantly differs from a known population mean when the population standard deviation is known.
   2. It is based on the standard normal distribution (z-distribution).
   3. The z-test is appropriate when the sample size is large (typically greater than 30) or when the population standard deviation is known.
   4. If the sample size is small or the population standard deviation is unknown, the t-test is usually preferred.
3. **F- Test:**
   1. The F-test, also known as Fisher's F-test, is a statistical test used to compare the variances of two or more groups or populations.
   2. The F-test is commonly used in analysis of variance (ANOVA) to determine if there are significant differences between the means of multiple groups.
   3. It compares the ratio of the mean square variation between groups to the mean square variation within groups.
   4. In 2 way F – Test we just use array

**Skewness**

1. Skewness is a statistical measure that describes the asymmetry or lack of symmetry in a probability distribution.
2. It indicates the deviation of a distribution from the normal distribution.
3. Types:
   1. **Positive Skewness:**
      1. A positive skewness indicates a longer or fatter tail on the right side of the distribution.
      2. The mean is usually greater than the median, and the distribution is said to be right-skewed.
   2. **Negative Skewness:**
      1. A negative skewness indicates a longer or fatter tail on the left side of the distribution.
      2. The mean is usually less than the median, and the distribution is said to be left-skewed.
   3. **Zero Skewness:**
      1. A skewness value of zero indicates that the distribution is perfectly symmetric

**Chi Square Test**

1. The chi-square test is a statistical test used to determine if there is a significant association between two categorical variables.
2. It assesses whether the observed frequencies of a contingency table differ significantly from the expected frequencies

**Sub Plot**

1. Subplots method provides a way to plot multiple plots on a single figure. Given the number of rows and columns , it returns a tuple ( fig , ax ), giving a single figure fig with an array of axes ax .

**Exploratory Data Analysis (EDA):**

1. It is an important step in the data analysis process.
2. It involves visualizing the data to understand its main characteristics, identify patterns and get insights for analysis
3. Python provides various libraries, such as NumPy, Pandas, and Matplotlib, which are commonly used for performing EDA.

**Machine Learning**

1) Machine Learning is a subset of AI.

2)It is the practice of getting machines to make decisions without being programmed.

Applications:-

1)Banking and Financial Sector

2)Online Fraud Detection

**Learning Algorithms:**

1. There are 3 types of learning techniques:
   1. Supervised learning
   2. Unsupervised learning
   3. Reinforcement learning

|  |  |  |
| --- | --- | --- |
| Supervised learning | Unsupervised learning | Reinforcement learning |
| 1)In Supervised learning the data is already labeled, so we already know the target variable . | 1)It predicts patterns based on the past data. | 1)The main goal of reinforcement learning is to train the agent to complete a task within uncertain environment. |
| 2)Input data is labeled | 2)Input data is not labeled | 2)Not predefined |
| 3)Deals with linear, logistic regression problems | 3)Deals with k-means, hierarchical clustering problems | 3)Deals with Q-learning, Deep Q-learning Neural networks |
| 4)Number of classes are known | 4)Number of classes are not known | 4)No classes |
| 5)Very complex | 5)Less complex | 5)Works within the user environment |
| 6)Provided with both Input and Output variables | 6)Provided with only Input variables |  |
| 7)Eg:- Image detection, Population growth rate prediction | 7)Eg:- Targeted marketing, Customer segmentation, Recommendation system, etc | 7)Eg:- Driver-less cars, gaming, health care,etc |

**1) Supervised Learning:**

1) In this type of learning, the data is already labeled, which means you know the target variable

2)Number of classes are known

3)Using this method systems can predict the future outcomes based on past fata

4)Deals with decision tree, logistic, linear regression, Naïve Baye’s , etc

5) Eg: Image detection, growth rate prediction, etc

**2) Unsupervised Learning:**

1) It predicts patterns based on the past data.

2) Number of classes are not known

3)The system are able to identify hidden features from the input data provided

4)Deals with K – means clustering , hierarchial clustering , apriori algoriyhm

5) Eg: Targeted marketing, Customer segmentation, etc

**3) Reinforcement learning:**

1. The main goal of reinforcement learning is to train the agent to complete a task within uncertain environment.
2. The agent receives observations and a reward from the environment and sends actions to the environment
3. The reward measures how successful action is with respect to completing the task goal
4. Deals with Q-learning, Deep Q-learning Neural networks
5. Eg: Driver-less cars, gaming, health care,etc
6. 1)The goal of reinforcement learning is to train an agent to complete a task within an uncertain environment.
7. 2)The agent receives observations and a reward from the environment and sends actions to the environment.
8. 3)The reward measures how successful action is with respect to completing the task goal.
9. 4)Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer
10. itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task.
11. 5)In the absence of a training dataset, it is bound to learn from its experience.
12. 6)Eg:-Q-learning and Deep Q-learning Neural Networks.

**Linear Regression Model**

1)It is used to estimate real values based on continuous variables.

2)Here, we establish a relationship between the independent and dependent variables by fitting the best line.

3)This best fit line is known as the regression line and represented by a linear equation Y= aX + b.

4)In this equation:-

Y – Dependent Variable

a – Slope

X – Independent variable

b – Intercept

5)These coefficients a and b are derived based on minimizing the ‘sum of squared differences’ of distance between data points and regression line.

6)Eg:-

1)The weight of the person is linearly related to their height,this shows a linear relationship between the height and weight of the person.

According to this, as we increase the height, the weight of the person will also increase.

7)Applications of Liner Regression Model:-

1)Cost of houses

2)Total number of calls

3)Total sales

**Naïve Bayes algorithm**

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

2)It is mainly used in text classification that includes a high-dimensional training dataset.

3)It predicts on the basis of the probability of an object.

4) The Naïve Bayes algorithm is comprised of two words Naive and Bayes, which can be described as:-

Naive: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features.

Bayes: It is called Bayes because it depends on the principle of Baye's Theorem.

5)Bayes' Theorem:-

1)Baye's theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge.

2)It depends on the conditional probability.

3) The formula for Bayes' theorem is given as:-

P(A|B)= P(B|A)P(A)/ P(B)

Where,

P(A|B) is Posterior probability

P(BIA) is Likelihood probability

P(A) is Prior Probability

P(B) is Marginal Probability

6) Types of Naive Bayes's Model:-

1)Gaussian

2) Multinomial

3) Bernoulli

7) Applications:-

1)Spam filtration

2)Sentimental analysis

3) Classifying articles

**KNN Algorithm**

1)K-nearest neighbors (KNN) is a type of supervised learning algorithm used for both regression and classification.

2)KNN tries to predict the correct class for the test data.

3)KNN is a "Non-parametric algorithm", which means it does not make any assumption on underlying data.

4)It is also called a "lazy learner algorithm" because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an

action on the dataset.

5)The K-NN working can be explained on the basis of the below algorithm:-

Step-1: Select K number of the neighbors

Step-2: Calculate the Euclidean distance of K number of neighbors

Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.

Step-4: Among these K neighbors, count the number of the data points in each category.

Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

Step-6: Our model is ready.

6)Applications:-

1)Banking System

2)Calculating Credit Ratings

3)Politics

**Decision Tree Algorithm**

1)Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems.

2)It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

3)In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node.

4)Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

5)The decisions or the test are performed on the basis of features of the given dataset.

6)It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

7)In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.

8)A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

9)Decision Tree Terminologies:-

1)Root Node:-Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.

2)Leaf Node:-Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.

3)Splitting:-Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.

4)Branch/Sub Tree:-A tree formed by splitting the tree.

5)Pruning:-Pruning is the process of removing the unwanted branches from the tree.

6)Parent/Child node:-The root node of the tree is called the parent node, and other nodes are called the child nodes.

**Support Vector Machine Algorithm :**

1) Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems.

2)Primarily, it is used for Classification problems in Machine Learning.

3)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.

4)This best decision boundary is called a hyperplane.

5)SVM chooses the extreme points/vectors that help in creating the hyperplane.

6)These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

7)Types of SVM:-

a. Linear SVM

b. Non-Linear SVM

8)Applications:-

1)Face Observation

2)Handwriting Remembrance

**Elbow method for K means clustering:**

1. The Elbow method is a technique to determine the optimal number of clusters for KMeans clustering. It involves fitting the KMeans algorithm with different values of k (number of clusters) and recording the within-cluster sum of squares (inertia) for each k. The inertia measures how spread out the data points are within their respective clusters. The goal is to find the k value where adding more clusters does not significantly reduce inertia, leading to an "elbow" in the plot.

**Principal Component Analysis(PCA):**

1. It is a popular dimensionality reduction technique used in machine learning and data analysis.
2. The primary goal of PCA is to reduce the number of features in a dataset while preserving the most important patterns and relationships among the data points.
3. Steps involved are:
   1. Standardization
   2. Covariance matrix
   3. Eigenvectors and Eigenvalues
   4. Selection of the principal components
   5. Projection
4. Reducing the dimensionality of the data, PCA allows for easier visualization and faster computation in machine learning tasks.
5. PCA is widely used in various fields, including image processing, feature extraction, data compression, and data visualization.