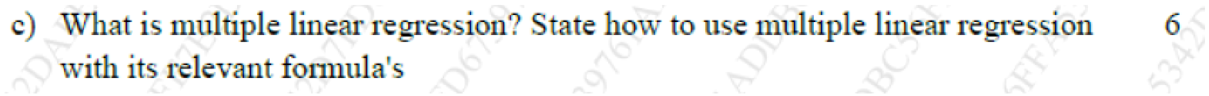


c) What is multiple linear regression? State how to use multiple linear regression with its relevant formula's 6

Solu->

What is multiple linear regression? State how to use multiple regression with formula

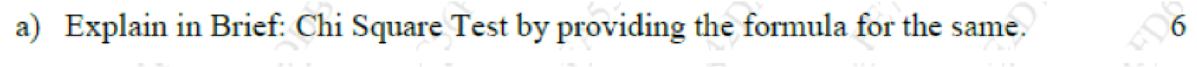
Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable.

Multiple linear regression formula

The formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

- y = the predicted value of the dependent variable
- B_0 = the y-intercept (value of y when all other parameters are set to 0)
- $B_1 X_1$ = the regression coefficient (B_1) of the first independent variable (X_1) (a.k.a. the effect that increasing the value of the independent variable has on the predicted y value)
- … = do the same for however many independent variables you are testing
- $B_n X_n$ = the regression coefficient of the last independent variable
- ϵ = model error (a.k.a. how much variation there is in our estimate of y)

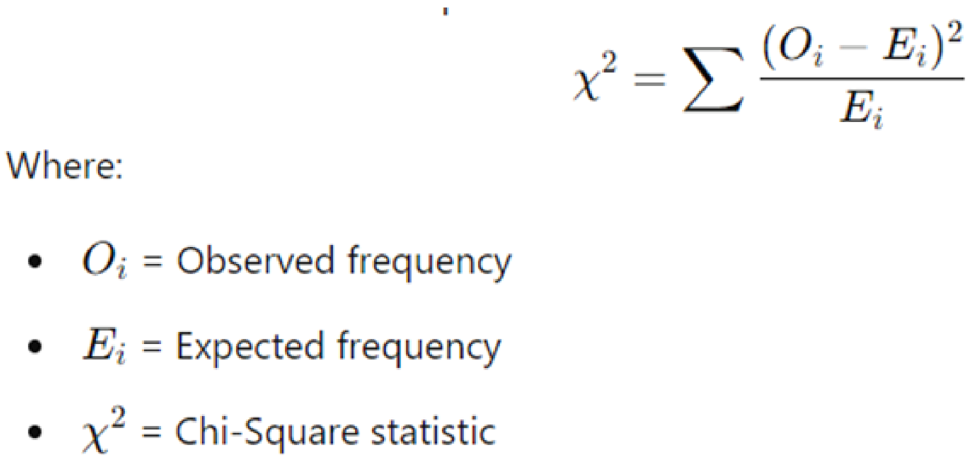


a) Explain in Brief: Chi Square Test by providing the formula for the same. 6

Solu->

The **Chi-Square Test** is a non-parametric statistical test used in hypothesis testing to assess relationships between categorical variables. Since it is non-parametric, it does not require the data to follow a specific distribution. The Chi-Square Test is versatile and can be applied in two primary ways: as a **test of goodness of fit** and as a **test of independence**.

- **Purpose:** This test evaluates how well the observed data matches the expected data under a specific theoretical distribution. It is used to determine if the differences between observed frequencies and expected frequencies are due to chance or if they are statistically significant.
- **How It Works:**
  - The test compares the observed frequencies (the actual data) with the expected frequencies (data that would be expected if the null hypothesis were true).



$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where:

- O_i = Observed frequency
- E_i = Expected frequency
- χ^2 = Chi-Square statistic

a) Write a short note on T- Test. 6

The **T-test** is a parametric statistical test used in hypothesis testing, primarily for comparing the means of a small sample (typically less than 30) to determine if there is a significant difference between them. The test is based on the **Student's T-distribution** and was developed by William Sealy Gosset, who published his work under the pseudonym "Student."

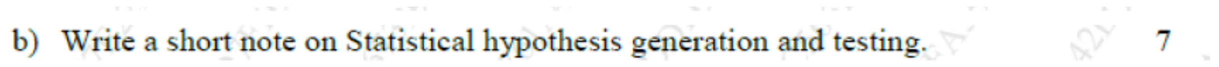
Key Features of the T-Test:

1. **Purpose:**
   - The T-test is used to assess whether the mean of a sample significantly differs from a known value (e.g., population mean) or whether the means of two samples are significantly different from each other. It is particularly useful when the sample size is small and the population standard deviation is unknown.
2. **Types of T-Tests:**
   - **One-Sample T-Test:** Compares the mean of a single sample to a known population mean.

  - **Independent (Two-Sample) T-Test:** Compares the means of two independent samples to see if they are significantly different.

3. **Application and Importance:**

  - The T-test is essential for testing the significance of mean values in situations where sample sizes are small. It allows researchers to make inferences about a population based on a sample, even when limited data is available.
  - It is widely used in various fields such as medicine, social sciences, and economics, where sample sizes often cannot be large due to practical constraints.



b) Write a short note on Statistical hypothesis generation and testing. 7

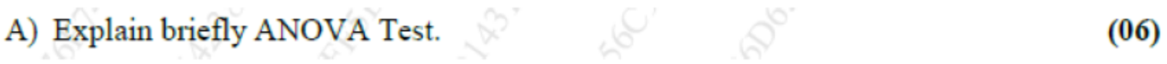
Statistical Hypothesis Generation and Testing is a critical process in inferential statistics used to make decisions or draw conclusions about a population based on sample data.

1. Hypothesis Generation:

- **Definition:** Hypothesis generation involves creating assumptions or educated guesses about a population parameter. These assumptions are based on existing knowledge, theories, or preliminary data.
- **Types of Hypotheses:**
  - **Null Hypothesis (H_0):** A statement that assumes no effect, no difference, or no relationship in the population. It represents the status quo or the claim to be tested.
  - **Alternative Hypothesis (H_1 or Ha):** A statement that contradicts the null hypothesis. It represents the outcome the researcher aims to support, suggesting that there is an effect, difference, or relationship.
- **Example:** In testing a new drug, the null hypothesis might state that the drug has no effect on patients, while the alternative hypothesis would state that the drug does have an effect.

2. Hypothesis Testing:

- **Definition:** Hypothesis testing is a formal statistical procedure used to decide whether to reject or fail to reject the null hypothesis based on sample data.



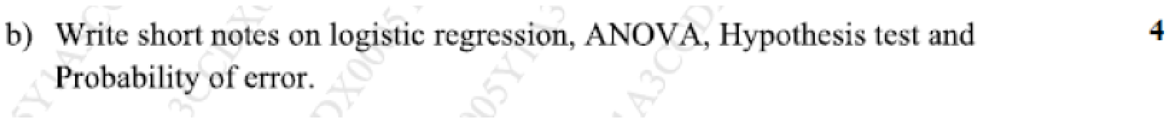
A) Explain briefly ANOVA Test. **(06)**

Solu->**Analysis of Variance (ANOVA)** is a powerful parametric statistical test used in hypothesis testing to compare the means of three or more groups

The test is particularly useful when determining whether there are any statistically significant differences between the means of multiple independent groups.

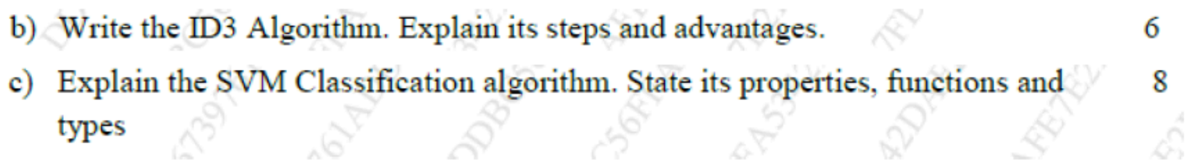
Key Features of the ANOVA Test:

1. **Purpose:**
    - ANOVA is used to test the significance of differences in mean values across more than two groups. Unlike the T-test, which is limited to comparing the means of two groups, ANOVA can handle multiple groups simultaneously. This makes it a vital tool for experiments or studies where multiple treatments, conditions, or groups are being compared.
2. **F-Test:**
    - The ANOVA test utilizes the F-test to compare the variance between the groups to the variance within the groups. The F-statistic calculated during the test helps determine whether the observed differences in means are greater than what could be expected by random chance alone.
3. **Application and Importance:**
    - ANOVA is extensively used in fields such as psychology, agriculture, biology, and marketing, where researchers need to compare the effects of different treatments or conditions.



b) Write short notes on logistic regression, ANOVA, Hypothesis test and Probability of error. 4

Solu->



b) Write the ID3 Algorithm. Explain its steps and advantages. 6

c) Explain the SVM Classification algorithm. State its properties, functions and types 8

Support Vector Machine or SVM is one of the most popular supervised learning algorithm which is used for classification as well as regression problems.

The goal of SVM algorithm is to create best line or decision boundary that can segregate n-dimensional spaces into classes so that we can easily put the new data point into correct category in future.

SVM algorithm can be used for face detection, image classification, text categorization etc.

SVM Algorithm Steps

1. Select two hyperplanes (2D) which separates the data with no points between them (red lines)

2. Maximize their distance (the margin)

3. The average line (here the line half way between the two red lines) will be the decision boundary.

Key concepts of SVM

Support vectors: Data points that are closet to the hyperplane is called support vectors.

Seperating line: will be defined with the help of these points.

Hyperplane- It is a decision plane or space which is divided between set of objects having different classes.

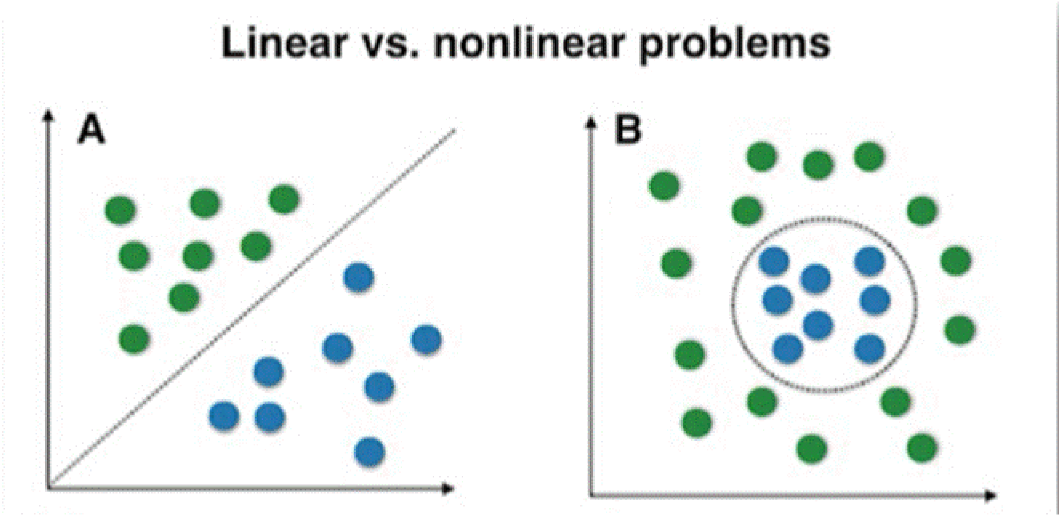
Margin: It is defined as gap between two lines on the closet data points of different classes.It can be calculated as perpendicular distance from the line to the support vectors.Large margin is considered as good margin and small margin is considered as bad margin.

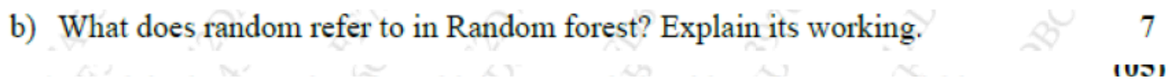
Types of SVM

Linear SVM: It is used for linearly seperable data. Which means if a dataset can be classified into two classes by using single straight line such data is termed as linearly seperable data.

And classified used is called as linear SVM classifier.

Non-linear SVM: Non-linear SVM is used for non-linearly seperable data, which means if a dataset cannot be classified using straight line, then such data is termed as non-linear data and classifier used is called as Non-linear classifier.




b) What does random refer to in Random forest? Explain its working. 7

Solu->

Random Forest (Simplified Explanation)

Random Forest is an improvement over **bagging (bootstrap aggregation)**. The main idea is to build many decision trees that are **less correlated** with each other, and then **average** their predictions to reduce variance and improve accuracy.

Core Idea

Bagging reduces variance by averaging many trees, but those trees can still be similar to each other. Random Forest fixes this by **forcing the trees to be different** using random feature selection. This makes the trees *decorrelated*, so the final averaged result becomes more stable.

2. Working of the Random Forest Algorithm

(i) Bootstrap Sampling

- For each tree, a bootstrap sample (sample drawn with replacement) is created from the training dataset.
- Each tree is grown on its own bootstrap sample.

(ii) Tree Construction with Random Feature Selection

- At every node/split, only **m random features** (out of total p features) are considered.
- The best split is chosen from this subset.
- Trees are generally grown to full depth (no pruning).

(iii) Ensemble Prediction

- **Regression:** Average of all tree predictions.
- **Classification:** Majority vote among all trees.

---

3. Key Features

(i) Decorrelation of Trees

- Randomly selecting features at each split ensures that no two trees are too similar.
- Reduces correlation (ρ) between trees → significantly lowers variance of the final model.

(ii) Bias–Variance Behaviour

- Individual trees: low bias, high variance.
- Random Forest: keeps bias low while reducing variance through averaging.

(iii) Out-of-Bag (OOB) Error Estimation

- Around one-third of training samples are left out of each bootstrap sample.
- These OOB samples act as internal validation data.
- OOB error closely approximates test error and removes the need for cross-validation.

---

4. Advantages

- High accuracy with minimal parameter tuning.

- Robust to noise and overfitting.
- Works well with large datasets and high-dimensional feature spaces.
- Provides feature importance estimates.

C) Explain the following. (04)

i) Boosting

ii) Bagging

Bagging:

- Bagging creates multiple versions of the training set by randomly sampling with replacement. This means that some data points may appear multiple times, while others might be left out.
- Training Different Learners: Each learner (model) is trained on a different version of the training set, and in the end, we combine their results, usually by voting.
- Bagging helps to reduce this instability by combining multiple models trained on different samples of the data.
- Example: If you train a decision tree on slightly different subsets of data, you might get very different results. By using bagging, we average out these results, making the model more robust.
- For Regression and Classification: Bagging can be used for both classification problems

(e.g., deciding if an email is spam or not) and regression problems (e.g., predicting house prices). The final prediction is the average or majority vote of the models.

Boosting:

- Boosting is an iterative process where we train weak models (models that aren't very strong) sequentially. Each model tries to correct the mistakes made by the previous ones.
- Focus on Misclassified Data: After each model is trained, Boosting gives more importance to the data points that were classified incorrectly by earlier models, forcing the new model to focus on these harder cases.
- Example: Imagine you're training a model to classify animals in pictures. If your first model misclassifies some cats as dogs, the next model will focus more on these specific pictures to get them right.
- Boosting Procedure:

Step 1: Start by training a model on the entire dataset with equal weights for all observations.

Step 2: Identify where the model makes mistakes and increase the weight

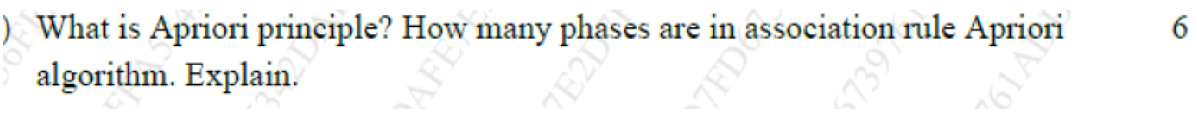
(importance) of these misclassified data points.

Step 3: Train a new model that focuses on correcting these mistakes.

Step 4: Continue this process for several iterations until the model improves its

accuracy.

Aspect	**Bagging**	**Boosting**
Data Partition	Data is randomly split into subsets, and each subset is used to train a different model.	Misclassified data points are given higher importance after each round of training.
Goal	Reduce the variance of the model (i.e., reduce overfitting).	Increase the accuracy of the model by focusing on misclassified data.
Method	Uses random sampling.	Uses methods like gradient descent to improve accuracy.
Function	Combines model outputs by voting or averaging.	Uses a weighted voting method, giving more importance to models that performed better.



) What is Apriori principle? How many phases are in association rule Apriori algorithm. Explain. 6

Solu->

Apriori Algo:

Association Mining rule

- Apriori algorithm uses frequent item sets to generate association rule and it is designed to work on databases that contain transactions.
- With the help of these association rule, it determines how strongly or how weakly two objects are connected.
- It is mainly used for market basket analysis and helps to find out those products that can be bought together.

Steps for Apriori algorithm

1. Determine the support of item sets in the transactional database, and select the minimum support and confidence.

2. Take all support in the transaction with higher support value than minimum or selected support value.

3. Find all the rules of these subsets that have higher confidence value than the threshold or minimum confidence.

4. Sort the rules as the decreasing order of lift.

Advantages:

- Easy to understand algorithm
- Easily implemented on large datasets

Disadvantages:

- Slow compared to other algorithms
- Overall performance is reduced
- Time complexity and space complexity is high

Applications:

- Market Basket Analysis
- Medical Diagonosis: Helps in identifying probability of illness for particular disease