STATISTICS

1. Using a goodness of fit,we can assess whether a set of obtained frequencies differ from a set of frequencies.
a) Mean
b) Actual
c) Predicted
d) Expected

ANS---D

2. Chisquare is used to analyse
a) Score
b) Rank
c) Frequencies
d) All of these

ANS—C

3. What is the mean of a Chi Square distribution with 6 degrees of freedom?
a) 4
b) 12
c) 6
d) 8

ANS—C

4. Which of these distributions is used for a goodness of fit testing?
a) Normal distribution
b) Chisqared distribution
c) Gamma distribution
d) Poission distribution

ANS—B

5. Which of the following distributions is Continuous
a) Binomial Distribution
b) Hypergeometric Distribution
c) F Distribution
d) Poisson Distribution

ANS—C

6. A statement made about a population for testing purpose is called?
a) Statistic
b) Hypothesis

c) Level of Significance
d) TestStatistic

ANS—B

7. If the assumed hypothesis is tested for rejection considering it to be true is called?
a) Null Hypothesis
b) Statistical Hypothesis
c) Simple Hypothesis
d) Composite Hypothesis

ANS—A

8. If the Critical region is evenly distributed then the test is referred as?
a) Two tailed
b) One tailed
c) Three tailed
d) Zero tailed

ANS—A

9. Alternative Hypothesis is also called as?
a) Composite hypothesis
b) Research Hypothesis
c) Simple Hypothesis
d) Null Hypothesis

ANS—B

10. In a Binomial Distribution, if 'n' is the number of trials and 'p' is the probability of success, then the mean value is given by
a) np
b) n

ANS—A

MACHINE LEARNING

1. R-squared or Residual Sum of Squares (RSS) which one of these two is a better measure of goodness of fit model in regression and why?

R-squared regression is the best method to measure goodness of fit model in regression.because it ranges from 0-1 higher values are good fir and its easy to interpret and compare with different models. It gives meaningful comparisons between the models. Anybody can understand and explain about it.

2. What are TSS (Total Sum of Squares), ESS (Explained Sum of Squares) and RSS (Residual Sum of Squares) in regression. Also mention the equation relating these three metrics with each other.

These 3 are important components in regression to findout the goodness of fit of a regression model.
TSS:
It represents the total variability in the dependent variable Y which doesn't consider any predictors.

$$TSS=\sum_{i=1n}(Y_i-\bar{Y})_2$$

ESS:
It represents the total variability in the dependent variable Y which is explained by the regression model
It is calculated by sum of the squared difference between the predicted values and mean of the Y

$$ESS=\sum_{i=1n}(\hat{Y}_i-\bar{Y})_2$$

RSS:
It represents the residuals in the dependent variable Y that are not accounted for by the regression model.

$$RSS=\sum_{i=1n}(Y_i-\hat{Y}_i)_2$$

The relation between these 3 isTSS=ESS+RSS

3. What is the need of regularization in machine learning?
   It is a technique used to prevent overfitting and underfitting and help us to get an optimal model. Which improves model performance, stability and interpretability. It reduces complexity in model.

4. What is Gini–impurity index?

Gini impurity index is also referred as Gini Impurity. It ranges from 0-0.5.it is used in Decision Tree algorithm particularly in binary classification tasks used to determine the best split at each node.
0 indicates subset is completely homogeneous
0.5 indicates subset is completely heterogeneous

$$\text{Gini}(S)=1-\sum_{i=1N}p_i2$$

5. Are unregularized decision-trees prone to overfitting? If yes, why?

unregularized decision trees are prone to overfitting. there are several reasons, some of them are
High variance
Sensitive to small changes
Memorization of noise
Lack of generalization

6. What is an ensemble technique in machine learning?

Ensemble methods are techniques which is used to improve the accuracy of results in models by combining multiple models instead of using a single model. The combined models increase the accuracy of the results significantly.
Bagging
Boosting
Random forest
Stacking

voting

7. What is the difference between Bagging and Boosting techniques?

Bagging and Boosting are both ensemble techniques that combine multiple models to improve predictive performance. Bagging uses parallel training of independent models to reduce variance, while Boosting uses sequential training to focus on correcting errors and improve overall performance.

8. What is out-of-bag error in random forests?

In Random Forests the out-of-bag (OOB) error is an estimate of the model's performance on unseen data. It is calculated using the training data itself, without the need for a separate validation set.

9. What is K-fold cross-validation?

Cross-validation is a statistical method used to estimate the skill of machine learning models. K-fold cross-validation is a technique for evaluating predictive models. The dataset is divided into k subsets or folds. The model is trained and evaluated k times, using a different fold as the validation set each time. Performance metrics from each fold are averaged to estimate the model's generalization performance.

10. What is hyper parameter tuning in machine learning and why it is done?

Hyperparameter tuning, also known as hyperparameter optimization, is the process of finding the optimal set of hyperparameters for a machine learning model. Hyperparameters are parameters that are set prior to the training process and control the behavior of the learning algorithm.
Its is done because of several reasons those are,
To prevent overfitting
To improve efficancy
To optimise model performance
To handle different models

11. What issues can occur if we have a large learning rate in Gradient Descent?

several issues can occur
some of them are;
instability
difficulty
overshooting the minimum
divergence

12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?

We can use logistic regression for non linear data but it will struggle to perform the action.
It cant understand the independent variable and testing values it will confuse.
Hence, out put is not efficient

13. Differentiate between Adaboost and Gradient Boosting.

Adaboost and gradient boosting are different on the following aspects those are,
Weighting of samples
Training process
Loss function
Model complexity

14. What is bias-variance trade off in machine learning?

It is the balance between the bias and variance of a predictive model and used to generalize the unseen data

15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.

Linear Kernel:

The linear kernel is the simplest kernel used in SVM.
It defines the decision boundary as a straight line in the feature space.
The linear kernel is computationally efficient and works well when the data is linearly separable.

RBF (Radial Basis Function) Kernel:

The RBF kernel is a non-linear kernel that maps the data into a higher-dimensional space.
It defines the decision boundary as a non-linear function that can capture complex patterns in the data.
The RBF kernel is characterized by a single parameter, the bandwidth (gamma), which controls the smoothness of the decision boundary.
The RBF kernel is versatile and can handle non-linearly separable data, but it may be prone to overfitting if the bandwidth parameter is not properly tuned.

Polynomial Kernel:

The polynomial kernel is another non-linear kernel used in SVM.
It defines the decision boundary as a polynomial function of the input features.
The polynomial kernel has a degree parameter that controls the degree of the polynomial, which determines the complexity of the decision boundary.
Like the RBF kernel, the polynomial kernel can capture non-linear relationships in the data, but it may also be prone to overfitting if the degree parameter is too high.