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?

Ans. R-squared (R^2) and Residual Sum of Squares (RSS) are both commonly used measures to assess the goodness of fit of a regression model, but they capture different aspects of model performance, and the choice between them depends on the context and what you want to evaluate.

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.

Ans. TSS = ESS + RSS, where TSS is Total Sum of Squares, ESS is Explained Sum of Squares and RSS is Residual Sum of Suqare. The aim of Regression Analysis is explain the variation of dependent variable Y.

3. What is the need of regularization in machine learning?

Ans. we use regularization in machine learning to properly fit a model onto our test set. Regularization techniques help reduce the chance of overfitting and help us get an optimal model.

4. What is Gini–impurity index?

Ans. Gini impurity measures how often a randomly chosen element of a set would be incorrectly labeled if it were labeled randomly and independently according to the distribution of labels in the set. It reaches its minimum (zero) when all cases in the node fall into a single target category.

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

Ans.

6. What is an ensemble technique in machine learning?

Ans. Ensemble methods are techniques that create multiple models and then combine them to produce improved results.

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

Ans. The bagging technique combines multiple models trained on different subsets of data, whereas boosting trains the model sequentially, focusing on the error made by the previous model.

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

Ans. The out-of-bag (OOB) error is the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample.

9. What is K-fold cross-validation?

Ans. K-fold cross-validation is a technique for evaluating predictive models.

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

Ans. The only way to determine these is through multiple experiments, where you pick a set of hyperparameters and run them through your model. This is called hyperparameter tuning.

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

Ans. When the learning rate is too large, gradient descent can suffer from divergence. This means that weights increase exponentially, resulting in exploding gradients which can cause problems such as instabilities and overly high loss values.

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

Ans. Logistic Regression has traditionally been used as a linear classifier, i.e. when the classes can be separated in the feature space by linear boundaries. That can be remedied however if we happen to have a better idea as to the shape of the decision boundar.

13. Differentiate between Adaboost and Gradient Boosting.

Ans. In the case of AdaBoost, the shifting is done by up-weighting observations that were misclassified before, while Gradient Boosting identifies the difficult observations by large residuals computed in the previous iterations.

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

Ans. In machine learning, as you try to minimize one component of the error (e.g., bias), the other component (e.g., variance) tends to increase, and vice versa. Finding the right balance of bias and variance is key to creating an effective and accurate model. This is called the bias-variance tradeoff.

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

Ans. Linear- Linear regression uses the relationship between the data-points to draw a straight line through all them.

RBF- Radial Basis Function (RBF) Networks are a particular type of Artificial Neural Network used for function approximation problems.

Polynomial Kernels- The polynomial kernel function takes the dot product of the input data points and adds a constant to the result, which is raised to a power specified by the degree parameter of the function.

STATISTICS WORKSHEET-5

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) Expected

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

Ans. c) Frequencies

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) 6

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) Chi- squared distribution

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

Ans. c) F Distribution

6. A statement made about a population for testing purpose is called?
a) Statistic
b) Hypothesis
c) Level of Significance
d) TestStatistic

Ans. b) Hypothesis

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) Null Hypothesis

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) Two tailed

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

Ans. b) Research Hypothesis

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) np