MACHINE LEARNING

Assignment 5

Q1 to Q15 are subjective answer type questions, Answer them briefly.

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?

Answer-R-squared, also known as the coefficient of determination, is generally considered a better measure of goodness of fit in regression compared to Residual Sum of Squares (RSS). R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It provides an indication of how well the independent variables explain the variability of the dependent variable. On the other hand, RSS measures the unexplained variation or the residuals of the model. While RSS is informative about the amount of error in the model, R-squared provides a standardized measure of goodness of fit, making it easier to compare models and interpret the overall fit.

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.

Answer-In regression analysis, Total Sum of Squares (TSS) represents the total variability in the dependent variable, Explained Sum of Squares (ESS) represents the variability in the dependent variable explained by the regression model, and Residual Sum of Squares (RSS) represents the unexplained variability or the residuals of the model. The equation relating these three metrics is: TSS = ESS + RSS.

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

Regularization in machine learning is needed to prevent overfitting, which occurs when a model learns the training data too well, capturing noise and random fluctuations instead of the underlying patterns. Regularization techniques introduce a penalty term to the loss function, discouraging the model from fitting the training data too closely and promoting generalization to unseen data.

4. What is Gini–impurity index?

Answer-The Gini impurity index is a measure used in decision tree algorithms to evaluate the impurity or disorder of a set of data points. It quantifies the likelihood of incorrectly classifying a randomly chosen element if it were randomly labeled according to the distribution of labels in the set. A lower Gini impurity indicates a more homogeneous set of data points, while a higher Gini impurity indicates greater diversity among the labels

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

Answer-Yes, unregularized decision trees are prone to overfitting. Decision trees have high variance, meaning they tend to capture noise and outliers in the training data. Without regularization, decision trees can become very complex, fitting the training data closely and resulting in poor performance on unseen data

6. What is an ensemble technique in machine learning?

Answer-Ensemble techniques in machine learning combine multiple models to improve predictive performance. Instead of relying on a single model, ensemble methods aggregate the predictions of several base models to make more accurate and robust predictions

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

Answer-Bagging involves training multiple base models independently on random subsets of the training data and then aggregating their predictions through averaging or voting

Boosting sequentially trains base models, where each subsequent model focuses on the errors made by the previous ones, thus improving overall predictive performance

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

Answer-Out-of-bag error in random forests refers to the error rate calculated on the data points that were not included in the bootstrap sample used to train a particular decision tree. In random forests, each tree is trained on a bootstrap sample, and the remaining data points (out-of-bag samples) are used to estimate the model's performance without the need for a separate validation set

9. What is K-fold cross-validation?

Answer-K-fold cross-validation is a technique used to assess the performance of a machine learning model. It involves splitting the dataset into K equal-sized folds, using K-1 folds for training the model and the remaining fold for validation. This process is repeated K times, with each fold used once as the validation set. The performance metrics are then averaged across all folds to provide an estimate of the model's generalization performance.

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

Answer-Hyperparameter tuning in machine learning involves selecting the optimal hyperparameters for a given model. Hyperparameters are parameters that are not learned directly from the data but rather set before the learning process begins. Hyperparameter tuning is done to improve the model's performance, increase its generalization ability, and avoid overfitting

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

Answer-Overshooting: Large learning rates can cause the algorithm to overshoot the minimum of the loss function, leading to oscillations or instability during training.

Divergence: If the learning rate is too large, the algorithm may fail to converge to a minimum and diverge away from the optimal solution

Poor convergence: Large learning rates may result in the algorithm converging too quickly to a suboptimal solution, preventing it from exploring the parameter space effectively.

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

Answer-Logistic Regression is a linear classification algorithm and is inherently limited to linear decision boundaries. Therefore, it may not be suitable for classifying non-linear data efficiently. While it can still be applied to non-linear data, it may not capture complex relationships effectively, leading to suboptimal performance compared to non-linear classifiers such as support vector machines or decision trees.

13. Differentiate between Adaboost and Gradient Boosting.

Answer-Adaboost (Adaptive Boosting) focuses on adjusting the weights of incorrectly classified instances, with subsequent weak learners giving more weight to difficult cases

Gradient Boosting builds trees sequentially, with each new tree fitting the residual errors of the previous ones, thereby reducing the error at each step.

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

Answer-Bias refers to the error introduced by approximating a real-world problem with a simplified model. High bias models may underfit the data and fail to capture the underlying patterns

Variance refers to the model's sensitivity to fluctuations in the training data. High variance models may overfit the data and perform well on the training set but poorly on unseen data

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

Answer-Linear, RBF (Radial Basis Function), and Polynomial kernels are used in Support Vector Machines (SVM) for mapping input data into higher-dimensional feature spaces

Linear kernel: It represents a linear decision boundary and is suitable for linearly separable data

RBF kernel: It uses a Gaussian-like function to map data into a higher-dimensional space, allowing for non-linear decision boundaries and capturing complex relationships in the data

Polynomial kernel: It maps data into a higher-dimensional space using polynomial functions, enabling SVM to model non-linear relationships between features

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

Answer - Expected

2. Chisquare is used to analyse

a) Score

b) Rank

c) Frequencies

d) All of these

Answer- 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

Answer- 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

Answer- b) Chisqared distribution

5. Which of the following distributions is Continuous

a) Binomial Distribution

b) Hypergeometric Distribution

c) F Distribution

d) Poisson Distribution

Answer- 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

Answer- 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

Answer- 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

Answer- a) Two tailed

9. Alternative Hypothesis is also called as?

a) Composite hypothesis

b) Research Hypothesis

c) Simple Hypothesis

d) Null Hypothesis

Answer- 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

Answer- a) np