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 p

goodness of fit model in regression and why?

Ans- for better measure of the goodness of fit in regression models R-squared is best as compared residual sum of squares.

Because R-squared provides a measure of the proportion of the variance in the dependent variable which is predictable from the independent variables . also gives an indication of how well the model fits the data, with values closer to 1 indicating a better fit. On the other hand, RSS simply measures the sum of squared differences between the observed and predicted values. While RSS is useful for understanding the magnitude of errors in the model, R-squared provides a more comprehensive assessment of the model's overall performance in explaining the variance in the data.

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- In regression analysis, TSS (Total Sum of Squares) represents the total variance of the dependent variable, and ESS (Explained Sum of Squares) measures the variance explained by the regression model, and RSS (Residual Sum of Squares) captures the unexplained variance. The relationship between these metrics can be summarized by the equation: TSS = ESS + RSS, showing that the total variance is the sum of the variance explained by the model and the unexplained variance (residuals)

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

Ans- Regularization in machine learning is used to prevent overfitting by adding a penalty term to the model's loss function that discourages overly complex models. This helps to improve the generalization capability of the model on unseen data.

4. What is Gini–impurity index?

Ans- Gini impurity is a measure of impurity or randomness within a dataset used in decision tree algorithms for classification. It quantifies the likelihood of incorrectly labeling a random data point based on the distribution of labels in a set. A Gini impurity score of 0 indicates a perfectly pure dataset with only one class present, while a score of 0.5 indicates maximum impurity with an equal distribution of classes. Decision trees split nodes based on reducing Gini impurity to create more homogeneous subsets in terms of class labels.

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

Ans- Yes, unregularized decision trees are prone to overfitting because they have high variance and can capture noise in the training data, leading to poor generalization on unseen data.

6. What is an ensemble technique in machine learning?

Ans- Ensemble techniques in machine learning involve combining multiple models to improve prediction accuracy. By leveraging diverse individual models, such as decision trees or neural networks, and combining their results, ensembles can reduce overfitting and capture a wider range of patterns in the data. Popular ensemble methods include Random Forest, Gradient Boosting, and Bagging.

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

Ans- the main difference is in how models are combined: bagging uses parallel training and averaging, while boosting uses sequential training and adaptive learning.

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

Ans- out-of-bag (OOB) error is a method used to estimate the performance of the model without the need for a separate validation set. It is calculated by measuring the prediction error on observations not used during the training (out-of-bag samples).

9. What is K-fold cross-validation?

Ans- K-fold cross-validation is a technique used in machine learning to evaluate the performance of a model by dividing data into k subsets. The model is trained on k-1 subsets and tested on the remaining subset k times.

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

Ans- Hyperparameter tuning in machine learning refers to the process of selecting the best set of hyperparameters for a machine learning algorithm. Hyperparameters are configuration settings that determine the behavior of the learning algorithm. It is done to improve the model's performance by finding the optimal hyperparameters that lead to better accuracy, generalization, and efficiency.

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

Ans- If we have a large learning rate in gradient descent, it can lead to the following issues:

a. Overshooting the minimum: Steps taken may be too large, causing the algorithm to overshoot the minimum of the loss function.

b. Divergence: The algorithm may fail to converge and instead diverge, leading to instability and inaccurate results.

c. Instability: Oscillations or erratic behavior in the optimization process can occur, making it challenging to find the optimal solution.

d. Missing the minimum: Large learning rates may cause the algorithm to miss the minimum or settle in a suboptimal solution.

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

Ans-No, logistic regression is a linear classification algorithm and is not suitable for modeling non-linear relationships in data.

13. Differentiate between Adaboost and Gradient Boosting.

Ans- The main difference between AdaBoost and Gradient Boosting is the way they create ensembles of weak learners. AdaBoost adjusts the weights of incorrectly classified observations at each iteration, whereas Gradient Boosting fits each weak learner in sequence to minimize the loss function based on the residuals from the previous learner.

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

Ans- The bias-variance trade-off in machine learning refers to the balance between a model's ability to capture the underlying patterns in the data (low bias) and its sensitivity to noise or fluctuations in the training data (variance). A model with high bias may oversimplify the data (underfitting), while a model with high variance may fit too closely to the training data but fail to generalize well to unseen data (overfitting). Achieving an optimal balance between bias and variance is crucial for building predictive models that perform well on new, unseen data.

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

Ans- a. Linear Kernel: The simplest SVM kernel, linear separates data using a straight line in the feature space.

b. RBF Kernel (Radial Basis Function): A flexible kernel that maps data to infinite-dimensional space using Gaussian functions. It is versatile for capturing complex relationships between data points.

c. Polynomial Kernel: Introduces non-linearity by mapping data into a higher-dimensional space using polynomial functions. It can handle non-linear data by finding curved decision boundaries.