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?

Ans- R-squared (R^2) is a better measure of goodness of fit for regression models because it indicates the proportion of variability in the dependent variable that is explained by the independent variables. Residual Sum of Squares (RSS) measures the unexplained variability in the dependent variable, providing insight into prediction errors, but it doesn't offer a standardized measure of fit like R-squared does.

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:

- Total Sum of Squares (TSS) measures the total variability of the dependent variable.

- Explained Sum of Squares (ESS) quantifies the variability explained by the regression model.

- Residual Sum of Squares (RSS)measures the unexplained variability left after fitting the regression model.

The relationship between these is summarized as:

TSS = ESS + RSS

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

Ans- Regularization in machine learning is essential to prevent overfitting, where a model learns the training data too well but fails to generalize to unseen data. It introduces constraints on model parameters during training to control complexity, ensuring better generalization to new data.

4. What is Gini–impurity index?

Ans- The Gini impurity index is a measure used in decision tree algorithms to evaluate the impurity or disorder of a dataset. It quantifies the probability of incorrectly classifying a randomly chosen element if it were labeled according to the distribution of labels in the dataset. In short, it assesses how often a randomly chosen label from the dataset would be incorrect. A lower Gini impurity indicates a purer dataset with less disorder, making it a desirable split criterion in decision tree algorithms.

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

Ans- Yes, unregularized decision trees are prone to overfitting. Decision trees have a tendency to grow deep and complex, capturing noise and outliers in the training data. Without any constraints on their growth, they can perfectly fit the training data, resulting in high variance and poor generalization to unseen data. Regularization techniques like pruning or limiting the tree depth are necessary to prevent overfitting by controlling the complexity of the tree.

6- What is an ensemble technique in machine learning?

Ans- Ensemble techniques in machine learning involve combining multiple models to improve predictive performance. Instead of relying on a single model, ensembles aggregate the predictions of multiple models, often resulting in better accuracy and robustness. Popular ensemble methods include bagging, boosting, and stacking.

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

Ans- Bagging trains multiple models independently on different subsets of the data and averages their predictions to reduce variance. Boosting trains models sequentially, each focusing on correcting the errors of its predecessor by giving more weight to misclassified instances. Bagging aims to reduce variance, while boosting aims to reduce bias.

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

Ans- The out-of-bag (OOB) error in random forests is an estimate of the model's performance on unseen data. In random forests, each tree is trained on a bootstrap sample (a subset of the original data created by sampling with replacement). The OOB error is calculated by evaluating each observation using only the trees for which it was not

included in the bootstrap sample. This allows for an unbiased estimate of the model's accuracy without the need for a separate validation set.

9- What is K-fold cross-validation?

Ans- K-fold cross-validation splits the dataset into K equally sized folds, trains the model on K-1 folds, and tests on the remaining fold. This process is repeated K times, and the performance is averaged for an overall estimation. It's a robust technique for evaluating model performance and reducing variance.

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

Ans- Hyperparameter tuning in machine learning involves selecting the best hyperparameters for a model to optimize its performance on unseen data. It's done to improve accuracy, prevent overfitting, adapt to different datasets, and enhance efficiency.

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

Ans- A large learning rate in Gradient Descent can lead to overshooting the minimum, instability, oscillations, failure to converge, and difficulty in generalization.

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

Ans- No, Logistic Regression is not suitable for classifying non-linear data because it assumes a linear relationship between features and outcomes.

13- Differentiate between Adaboost and Gradient Boosting.

Ans- Adaboost focuses on misclassified instances by adjusting instance weights, while Gradient Boosting minimizes the loss function of residuals to improve prediction accuracy.

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 simplicity (bias) and its ability to capture the variability in the data (variance). Achieving low bias usually increases variance, and vice versa. Striking the right balance is crucial for building models that generalize well to new data.

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

Ans-1. Linear Kernel: Computes a linear decision boundary by calculating the dot product between feature vectors. Suitable for linearly separable data.

2. RBF (Radial Basis Function) Kernel: Measures similarity between data points in a high-dimensional space. Effective for capturing non-linear relationships in data.

3. Polynomial Kernel: Computes similarity using a polynomial function. Useful for capturing non-linear decision boundaries and polynomial relationships in data.

Thank You