Assignment -3.1

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 is generally considered a better measure of goodness of fit in regression compared to Residual Sum of Squares . 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. Higher R-squared values indicate better fit. On the other hand, RSS measures the discrepancy between the observed values and the values predicted by the model, but it don't provide a standardized measure of goodness 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.

Answer:-

In regression analysis:

- TSS (Total Sum of Squares) measures the total variability in the dependent variable.
- ESS (Explained Sum of Squares) measures the variability in the dependent variable that is explained by the regression model.
- RSS (Residual Sum of Squares) measures the discrepancy between the observed values and the values predicted by the model. The relationship between these metrics is given by: TSS = ESS + RSS.

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

Answer:-

Regularization in machine learning is needed to prevent overfitting, where a model learns the training data too well and performs poorly on unseen data. Regularization techniques introduce a penalty term to the model's objective function, discouraging overly complex models and promoting simpler ones that generalize better to new data.

4. What is Gini–impurity index?

Answer:-

Gini impurity index is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. It is

commonly used as a criterion in decision tree algorithms for constructing decision trees by selecting the attribute that minimizes the impurity at each node.

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

Answer:-

Yes, unregularized decision trees can be prone to overfitting, especially when they are allowed to grow deep and complex. They can memorize the training data instead of learning general patterns, leading to 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. Examples include bagging, boosting, and stacking.

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

Answer:-
The main difference between bagging and boosting techniques lies in how they combine multiple models:

- Bagging (Bootstrap Aggregating) trains multiple models independently on different subsets of the training data and combines their predictions by averaging (for regression) or voting (for classification).
- Boosting trains models sequentially, where each subsequent model tries to correct the errors made by the previous models. Boosting pays more attention to examples that previous models misclassified.

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

Answer:-

Out-of-bag (OOB) error in random forests is the error rate of the model on the training data that was not used during the construction of a particular decision tree in the forest. It provides an estimate of the model's performance on unseen data 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. The training set is divided into k subsets, and the model is trained k times, each time using k-1 subsets for training and the remaining subset for validation. The performance metrics are then averaged over the k iterations.

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 model to maximize its performance on unseen data. Hyperparameters are parameters that are not learned from the data but are set before the learning process begins, such as the learning rate, regularization parameter, and the number of hidden layers in a neural network.

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

Answer:-

Large learning rates in gradient descent can lead to several issues, including:

- Overshooting the minimum: Large steps can cause the algorithm to overshoot the minimum of the cost function, leading to instability and slow convergence.
- Divergence: If the learning rate is too large, the algorithm may fail to converge to the minimum and diverge.
- Poor generalization: Large learning rates can cause the model to learn the training data too well, leading to poor performance on unseen data.

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

Answer:-

Logistic Regression is a linear model and is inherently limited to linear decision boundaries. It cannot capture complex nonlinear relationships between features. Therefore, it may not perform well on nonlinear data unless feature transformations are applied beforehand.

13. Differentiate between Adaboost and Gradient Boosting.

Answer:-

Adaboost and Gradient Boosting are both boosting algorithms, but they differ in how they weight the training examples and update the model:

- Adaboost adjusts the weights of training examples based on the error rate of the previous model, focusing more on misclassified examples in subsequent iterations.
- Gradient Boosting fits the new model to the residual errors of the previous model, minimizing the loss function in the gradient direction. It typically uses gradient descent to update the model.

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

Answer:-

Bias-variance tradeoff refers to the balance between the bias and variance of a model. A high-bias model is too simplistic and tends to underfit the training data, while a high-variance model is too complex and tends to overfit the training data. Finding the right balance between bias and variance is crucial for building models that generalize well to unseen data.

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

Answer:-

- Linear Kernel: The linear kernel is the simplest kernel function, representing the dot product of the input vectors. It works well when the data is linearly separable.
- RBF (Radial Basis Function) Kernel: The RBF kernel maps the data into a higher-dimensional space where data points are more easily separable. It is versatile and can capture complex nonlinear relationships between features.
- Polynomial Kernel: The polynomial kernel computes the dot product of the input vectors raised to a specified power, allowing it to capture polynomial relationships between features. It is useful for data that exhibits polynomial patterns.