Machine Learning Assignment

Machine Learning

Answer

1. R-squared is generally a better measure of goodness of fit in regression as it represents the proportion of variance explained by the model relative to the total variance, whereas RSS only measures the unexplained variance.
2. TSS (Total Sum of Squares) measures the total variance in the dependent variable, ESS (Explained Sum of Squares) measures the variance explained by the regression model, and RSS (Residual Sum of Squares) measures the unexplained variance. The equation relating these three metrics is: TSS = ESS + RSS
3. Regularization in machine learning is needed to prevent overfitting by adding a penalty term to the cost function that discourages overly complex models.
4. Gini impurity index is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.
5. Yes, unregularized decision trees are prone to overfitting because they can keep splitting the data until each data point has its own leaf node, resulting in high variance.
6. Ensemble techniques combine multiple models to improve performance or robustness, often by averaging or combining their predictions.
7. Bagging builds multiple models independently and combines their predictions, while boosting sequentially builds models, each one focusing on the mistakes of the previous ones.
8. Out-of-bag error in random forests is the error rate of the model on the samples not included in the bootstrap sample used to train each tree
9. .
   K-fold cross-validation divides the dataset into k subsets and iteratively uses each subset as a validation set while the rest are used for training, allowing for more reliable model evaluation

10.

Hyperparameter tuning involves optimizing the parameters of a machine learning model that are not learned during training, typically through techniques like grid search or random search, to improve model performance.

11

. Large learning rates in Gradient Descent can lead to overshooting the minimum of the cost function, causing the algorithm to diverge or take longer to converge.

12. Logistic Regression can be used for classification of non-linear data by applying transformations or by using non-linear features, but it may not capture complex non-linear relationships as effectively as other algorithms like kernel SVM or neural networks.

Machine Learning Assignment

13. Adaboost focuses on adjusting the weights of observations based on their classification accuracy, while Gradient Boosting builds trees sequentially, each one correcting the errors of the previous tree.

14. Bias-variance tradeoff refers to the balance between a model's ability to capture the underlying patterns in the data (bias) and its sensitivity to fluctuations in the training data (variance).

15.
- Linear Kernel: Used in SVM for linearly separable data, it calculates the dot product of the input vectors.
- RBF (Radial Basis Function) Kernel: A non-linear kernel that maps the input space into high-dimensional feature spaces, allowing SVM to handle non-linearly separable data.
- Polynomial Kernel: Another non-linear kernel that calculates the dot product of two vectors raised to a power, allowing SVM to separate data that's not linearly separable in the original space.