MACHINE LEARNING ASSIGNMENT - 5

Q1. R-squared or RSS, which is better and why?
R-squared is generally considered a better measure of goodness of fit in regression models because it represents the proportion of variance in the dependent variable that can be explained by the independent variables. In contrast, Residual Sum of Squares (RSS) merely quantifies the total deviation of the predicted values from the actual values without providing insight into how well the model explains the variability of the data.

Q2. What are TSS, ESS, and RSS in regression?

- **TSS (Total Sum of Squares):** This measures the total variance in the response variable, capturing how much the observed values deviate from the mean.
- **ESS (Explained Sum of Squares):** This represents the portion of variance explained by the regression model, showing how much better the model predicts the outcome compared to the mean.
- **RSS (Residual Sum of Squares):** This quantifies the variance not explained by the model, indicating the discrepancies between the observed and predicted values.
- **Relation:** The relationship among these metrics is given by the equation: TSS=ESS+RSS$TSS=ESS+RSS$.

Q3. What is the need for regularization in machine learning?
Regularization is essential in machine learning to prevent overfitting, which occurs when a model becomes too complex and learns noise in the training data instead of the underlying pattern. By adding a penalty term to the loss function for large coefficients, regularization encourages simpler models that generalize better to unseen data.

Q4. What is Gini impurity index?
The Gini impurity index is a metric used to measure the impurity or disorder in a dataset. It calculates the probability of a randomly chosen element being incorrectly labeled if it was labeled according to the distribution of labels in the subset. A lower Gini impurity indicates a purer subset.

Q5. Are unregularized decision trees prone to overfitting?
Yes, unregularized decision trees are indeed prone to overfitting. They can create very complex models that fit the training data too closely, capturing noise rather than the true signal. This results in poor performance on unseen data, as the model fails to generalize.

Q6. What is an ensemble technique in machine learning?
An ensemble technique involves combining multiple machine learning models to improve overall predictive performance. The idea is that by leveraging the strengths of various models, the ensemble

can achieve better accuracy and robustness than individual models, as errors from one model can be compensated by others.

Q7. What is the difference between Bagging and Boosting techniques?

- **Bagging (Bootstrap Aggregating):** This technique reduces variance by training multiple models on different subsets of the training data, then averaging their predictions. It helps stabilize the model's predictions and is effective for high-variance models like decision trees.
- **Boosting:** This is a sequential technique that combines weak learners to create a strong model. Each new model focuses on correcting the errors made by the previous models, thereby improving overall accuracy. Boosting aims to reduce both bias and variance.

Q8. What is out-of-bag error in random forests?
Out-of-bag error is an estimate of the model's performance that is calculated using the data samples not included in the training set for each individual tree in a random forest. This method provides a built-in validation mechanism, allowing the model to be evaluated without requiring a separate validation dataset.

Q9. What is K-fold cross-validation?
K-fold cross-validation is a technique used to assess the performance of a model. The dataset is divided into K equally sized subsets, or folds. The model is trained on K-1 folds and tested on the remaining fold. This process is repeated K times, with each fold serving as the test set once. The final performance metric is usually the average of all K tests, providing a more reliable estimate.

Q10. What is hyperparameter tuning in machine learning, and why is it done?
Hyperparameter tuning involves optimizing the parameters that are not learned during training but are set before the training process begins, such as learning rates and regularization strengths. This process is crucial for enhancing model performance, ensuring the model generalizes well to new data and avoids overfitting.

Q11. What issues can occur if we have a large learning rate in Gradient Descent?
A large learning rate in gradient descent can cause significant issues, such as overshooting the minimum of the loss function. This can lead to divergence, where the model fails to converge to an optimal solution, or oscillation, where the model keeps bouncing around the minimum instead of settling down.

Q12. Can we use Logistic Regression for the classification of non-linear data? If not, why?
Logistic regression is not suitable for directly classifying non-linear data because it assumes a linear relationship between the independent variables and the log-odds of the dependent variable. To handle non-linear data, transformations or different algorithms that can capture non-linear relationships are necessary.

Q13. Differentiate between Adaboost and Gradient Boosting.

- **Adaboost (Adaptive Boosting):** This technique focuses on misclassified instances by assigning them higher weights in subsequent iterations. It combines multiple weak classifiers to create a strong classifier, emphasizing accuracy on hard-to-classify examples.
- **Gradient Boosting:** This method builds models sequentially, where each new model aims to minimize the residual errors of the previous models using gradient descent. It can handle complex relationships and often achieves high accuracy.

Q14. What is bias-variance trade-off in machine learning?

The bias-variance trade-off is a fundamental concept in machine learning that describes the trade-off between two types of error: bias (error due to overly simplistic assumptions in the learning algorithm) and variance (error due to excessive sensitivity to fluctuations in the training data). A well-performing model seeks to minimize both types of error to improve generalization.

Q15. Give a short description of Linear, RBF, and Polynomial kernels used in SVM.

- **Linear Kernel:** This kernel represents a linear separation of data and is effective when the data is linearly separable. It is computationally efficient and simple.
- **RBF Kernel (Radial Basis Function):** This kernel maps data into a higher-dimensional space, making it effective for non-linear data. It captures complex relationships and is widely used in practice.
- **Polynomial Kernel:** This kernel allows for polynomial relationships between features, capturing interactions of different degrees. It can be useful when the relationship between classes is non-linear but can still be approximated by polynomial functions.