MACHINE LEARNING ASSIGNMENT

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?

-Both R-squared and Residual Sum of Squares (RSS) are better
-R-squared is easy to interpret, but it can be misleading for small datasets or non-linear models, and it doesn't penalize complexity.

-RSS will scale-independently, it encourages simpler models, and emphasize on error minimization. But it cannot directly interpretable as a percentage, larger models naturally have higher RSS.

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.

-Total Sum of Squares (TSS): Measures the total variance in the dependent variable.

-Explained Sum of Squares (ESS): Measures the variance explained by the model.

-Residual Sum of Squares (RSS): Measures the unexplained variance (error).

-TSS = ESS + RSS

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

-In machine learning, regularization combats overfitting. Complex models can memorize training data perfectly, but fail on unseen examples. Regularization techniques penalize overly complex models, encouraging simpler ones that generalize better, capturing the overall trend instead of noise in the data.

4) What is Gini–impurity index?

- Gini-impurity measures how often a random observation would be misclassified in a decision tree node. Lower values indicate better separation between classes, making it a popular choice for building effective and interpretable decision trees.

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

-Yes, unregularized decision trees are prone to overfitting. Their ability to split on any feature at any level allows them to perfectly fit the training data, creating complex rules that capture even noise, leading to poor performance on unseen data.

6) What is an ensemble technique in machine learning?

-Combine multiple models to improve predictive power and stability. Eg:- Bagging & Boosting

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

-Bagging: Creates diverse models by training on different data subsets, then averages their predictions for better stability.

-Boosting: Builds models sequentially, each focusing on correcting errors made by the previous model, leading to a more powerful ensemble.

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

-Out-of-bag error in random forests estimates the model's performance on unseen data. It utilizes data points not used to train each individual tree, providing an unbiased gauge of how well the forest generalizes.

9) What is K-fold cross-validation?

- It will splits data into K folds, trains on K-1 and tests on 1, repeats for all K, giving a reliable estimate of model performance on unseen data.

10) What is hyperparameter tuning in machine learning and why it is done?

-Hyperparameter tuning in machine learning is a fine-tuning knobs like learning rate. It is crucial for optimal learning and preventing overfitting, leading to better model performance.

-Optimizes model hyperparameters (e.g., learning rate, tree depth) that control model behaviour. Grid search or random search with cross-validation to find the best hyperparameter combination.

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

-If we have a large learning rate in Gradient Descent, then gradient descent may overshoot the minimum, lead to oscillations, or not converge at all. Tune the learning rate carefully for efficient and stable convergence.

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

-Yes, we cannot because while logistic regression is powerful for linear relationships, it can't model non-linear relationships directly. This is because it assumes a straight line separates the classes, which wouldn't work for curves or bends in the data.

13) Differentiate between Adaboost and Gradient Boosting

- AdaBoost emphasizes sample weights

- AdaBoost is more interpretable due to weaker individual models

-Gradient Boosting emphasizes gradients of the loss function

- Gradient Boosting is less interpretable due to complex combinations of weaker models.

14) What is bias-variance trade off in machine learning

-The bias-variance trade-off in machine learning is a balancing act between two sources of error:

-Bias: How well your model captures the overall trend in the data

-Variance: How sensitive your model is to specific training data

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

-Linear will project data into a higher-dimensional space

-RBF also known as Radial Basis Function will project data into an infinite-dimensional space using a Gaussian function.

-Polynomial will project data into a higher-dimensional space based on polynomial combinations of features.