Q1 to Q15 are subjective answer type questions, Answer them briefly.

Q1. 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 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. RSS only measures unexplained variance.

Q2. 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:
TSS (Total Sum of Squares) is the total variance, ESS (Explained Sum of Squares) is the variance explained by the model, and RSS (Residual Sum of Squares) is the unexplained variance. The equation is TSS = ESS + RSS.

3. What is the need of regularization in machine learning?
Answer:
Regularization in machine learning is needed to prevent overfitting by adding a penalty term to the model's complexity, discouraging overly complex models.

4. What is Gini–impurity index?
Answer:
Gini impurity is a measure of how often a randomly chosen element from a set would be incorrectly labeled based on the distribution of labels in the set.

5. Are unregularized decision-trees prone to overfitting? If yes, why?
Answer:
Yes, unregularized decision-trees are prone to overfitting because they can become too complex, fitting noise in the training data that doesn't generalize well.

6. What is an ensemble technique in machine learning?
Answer:
An ensemble technique in machine learning combines predictions from multiple models to improve overall performance and robustness.

7. What is the difference between Bagging and Boosting techniques?
Answer:
Bagging builds models independently and combines them, while Boosting builds models sequentially, giving more weight to misclassified instances.

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

Answer:

Out-of-bag error in random forests is the error on instances not used in the training of a particular tree, serving as an unbiased estimate of the model's performance.

9. What is K-fold cross-validation?

Answer:

K-fold cross-validation involves partitioning the dataset into K subsets, using K-1 for training and the remaining subset for testing, repeated K times.

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

Answer:

Hyperparameter tuning adjusts model settings to optimize performance on unseen data, done to find the best configuration for the model.

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 overshooting the minimum, causing oscillation or divergence rather than convergence.

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

Answer:

Logistic Regression is linear, so it may not perform well on non-linear data without appropriate feature engineering or transformations.

13. Differentiate between Adaboost and Gradient Boosting.

Answer:

Adaboost adjusts instance weights, while Gradient Boosting builds trees sequentially, correcting errors with each iteration.

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

Answer:

Bias-variance trade-off in machine learning balances bias (error from simplistic models) and variance (error from complex models), aiming for an optimal model complexity.

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

Answer:

Linear Kernel: Suitable for linearly separable data, projects data linearly.

RBF Kernel: Useful for non-linear data, computes similarity in a high-dimensional space.

Polynomial Kernel: Extends linear kernel with polynomial terms, handling more complex relationships.