MACHINE LEARNING

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?

   Ans – R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. The residual sum of squares (RSS) is a statistical technique used to measure the amount of variance in a data set that is not explained by a regression model itself. Instead, it estimates the variance in the residuals, or error term. The residual sum of squares (RSS) is the absolute amount of explained variation, whereas R-squared is the absolute amount of variation as a proportion of total variation. Hence R-squared is the better measure of goodness of fit model in regression.

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.
   Ans –
   A. The residual sum of squares (RSS) is a statistical technique used to measure the amount of variance in a data set that is not explained by a regression model itself.
   B. The term Total Sum of Squares refers to a statistical technique used in regression analysis to determine the dispersion of data points. The sum of squares can be used to find the function that best fits by varying the least from the data.
   C. Explained Sum of Squares, alternatively known as the model sum of squares or sum of squares due to regression, is a quantity used in describing how well a model, often a regression model, represents the data being modelled.
   D. Equation – TSS = RSS + ESS

3. What is the need of regularization in machine learning?
   Ans – Regularization is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting. This reduces overfitting of the model.

4. What is Gini–impurity index?
   Ans – Gini impurity is a function that determines how well a decision tree was split. Basically, it helps us to determine which splitter is best so that we can build a pure decision tree. Gini impurity ranges values from 0 to 0.5.

5. Are unregularized decision-trees prone to overfitting? If yes, why?
   Ans – Decision trees are prone to overfitting, especially when a tree is particularly deep. Since Decision trees perform classification without requiring much computation it becomes overfitting.

6. What is an ensemble technique in machine learning?
   Ans – Ensemble methods are techniques that create multiple models and then combine them to produce improved results.

7. What is the difference between Bagging and Boosting techniques?
   Ans – Bagging and boosting are both ensembles learning methods in machine learning.
   a. Bagging helps to decrease the model's variance whereas boosting helps to decrease the model's bias.

8. What is out-of-bag error in random forests?
   Ans – The out-of-bag (OOB) error is the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample. This allows the RandomForestClassifier to be fit and validated whilst being trained.

9. What is K-fold cross-validation?
   Ans – K-fold Cross-Validation is when the dataset is split into a K number of folds and is used to evaluate the model's ability when given new data.

10. What is hyper parameter tuning in machine learning and why it is done?
    Ans – Hyperparameter tuning consists of finding a set of optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any data set. That combination of hyperparameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors.

11. What issues can occur if we have a large learning rate in Gradient Descent?
    Ans – Gradient Descent is too sensitive to the learning rate. If it is too big, the algorithm may bypass the local minimum and overshoot. If it too small, it might increase the total computation time to a very large extent.

12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?
    Ans – Logistic regression is indeed nonlinear in terms of Odds and Probability; however, it is linear in terms of Log Odds. Non-linear problems can’t be solved with logistic regression because it has a linear decision surface. Linearly separable data is rarely found in real-world scenarios.

13. Differentiate between Adaboost and Gradient Boosting.
    Ans – AdaBoost is the first designed boosting algorithm with a particular loss function. On the other hand, Gradient Boosting is a generic algorithm that assists in searching the approximate solutions to the additive modelling problem. This makes Gradient Boosting more flexible than AdaBoost.

14. What is bias-variance trade off in machine learning?
    Ans – In statistics and machine learning, the bias–variance tradeoff is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters.

15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.
    Ans – did not understand the question.