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?
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.
3. What is the need of regularization in machine learning?
4. What is Gini–impurity index?
5. Are unregularized decision-trees prone to overfitting? If yes, why?
6. What is an ensemble technique in machine learning?
7. What is the difference between Bagging and Boosting techniques?
8. What is out-of-bag error in random forests?
9. What is K-fold cross-validation?
10. What is hyper parameter tuning in machine learning and why it is done?
11. What issues can occur if we have a large learning rate in Gradient Descent?
12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?
13. Differentiate between Adaboost and Gradient Boosting.
14. What is bias-variance trade off in machine learning?
15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.

Ans-9 To achieve K-Fold Cross Validation, we have to split the data set into three sets, Training, Testing and Validation, with the challenge of the volume of the data.

Ans-8 The out-of-bag error is the average error for each predicted outcome calculated using predictions from the trees,that do not contain that data point in their respective bootstrap sample.

Ans-7 If the classifier is unstable (high variance), then we need to apply bagging.If the classifier is straightforward (high bias), then we need to apply boosting.Bagging attempts to tackle the over-fitting issue.Boosting tries to reduce bias.

Ans-10 If you tune your hypermeters on the training set, you will protect model to overfit and suffer a performance hit on the test set. we used to tune with different parameters to find the best parameter out off for better performance of model.

Ans-4 Gini impurity is a function that determines how well a decision tree was split.Gini impurity ranges values from 0 to 0.5

Ans-13 In AdaBoost shortcomings are identified by high weighted data points, In GradientBoost shortcomings are identified by gradients.

Ans-14 Basically Bias variance Trade off is tradeoff between bias and variance it differe as if simple model then may be high bias low variance and if large parameter model then may be high variance and low bias.

Ans-15 Plynomial Kernel is best in image processing, RBF is a general-purpose kernel; used when there is no prior knowledge about the data. Linear Kernel is useful when dealing with large sparse data

Ans-3 Regularization is a technique to prevent the model from overfitting. Sometimes model performs well with training data but not with test data. It comes in overfitting.

Ans-6 Ensemble techniquese we use when we have very huge datset in terms of rosw and columns. Then we use some Ensemble techniques like Random Forest , AdaBoost, GradientBoost.