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 is better than Residual sum of square.

- ✓ R squared is the ratio and hence is independent of unit of measurement, and hence can be used for the comparison purpose.
- ✓ R squared explains the explained variation(that is the variation explained by the features used for model building.
- ✓ Where as residual sum of squares can be decreased by increasing number of features.

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:

- ✓ TSS (Total Sum of Squares) is the variation in response(label) variable without considering any feature(label). In other words it is the pure variation in data.

  So TSS=$\sum_1^n (yi - \bar{y})^2$

  (TSS is sum of squares of deviation of response variable y from its mean)

- ✓ ESS (Explained Sum of Squares) is the variation in estimated response(label) variable from its mean. In other words it is the pure variation in the estimated data.

  So ESS=$\sum_1^n (\hat{y}i - \bar{y})^2$

  (ESS is sum of squares of deviation of estimated response variable y from its mean)

- ✓ RSS (Residual Sum of Squares) is the variation in response(label) variable and its estimated value. In other words it is noise/ nuisance variation in the model.

  So RSS=$\sum_1^n (\ yi - \hat{y}i\ )^2$

  (RSS is sum of squares of deviation reponse variable from iestimated response variable)

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

Ans:

- ✓ Sometimes when some observations in (response variable (Label)) are repeated then there is a Chance of overfitting of model in such a case there is a need of regularization in Machine learning.
- ✓ Generally then accuracy score is too high or too low.

4. What is Gini–impurity index?

Ans:

- ✓ Gini index is measurement of impurity.
- ✓ It may be also called as measure of heterogeneity.
- ✓ It is used in Decision tree
- ✓ It is odds in favor of certain decision.

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

Ans:

- ✓ Over-fitting is one of the reason of unregularized decision tree.
- ✓ The other reasons may be the bias in the data.
- ✓ Or the larger variation in the data.

6. What is an ensemble technique in machine learning?

Ans: The use of multiple models in Machine learning is called as ensembled technique.

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

Ans:

- ✓ Bagging and boosting both are ensemble techniques.
- ✓ In bagging techniques single training algorithm is used for different subsets of training data by selecting samples with replacement. And then the average estimated value Is found.
- ✓ Boosting is nothing but development/growth from weaker side to stronger side. Prediction is the weighted mean of estimated values.

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

Ans: In ensembled technique When data is trained some part of data is kept aside. Such a data is out of Bag.

9. What is K-fold cross-validation?

Ans: Cross validation is a technique of validating the model.
In K fold validation every time data is divided into k parts. Data is trained on k-1 parts and tested on K th part which is kept aside.
In this way data gets trained considering all observations.

10. What is hyper parameter tuning in machine learning and why it is done?
Considering the model accuracy we like to concentrate on regression coefficients of the model fitted. Techniques of controlling the contributing features is called as hypertuning.

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?

Ans:

- ✓ Logistic regression model can be fitted when label is binary type data.
- ✓ If we fit such a label considering linear linear relationship between feature variables then it will be simple logistic regression model.
- ✓ If we build nonlinear relationship between feature variables then it will be nonlinear model.
- ✓ Logistic model itself is considered as non-linear model.
- ✓ Logistic model we estimate probability of occurrence or non-occurrence of label variable so building non-linear model may not give much more accuracy than simple relationship.

13. Differentiate between Adaboost and Gradient Boosting.

14. What is bias-variance trade off in machine learning?
When data is heterogeneous then it corresponds to higher variance.

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

FLIP ROBO