ASSIGNMENT - 5

Machine Learning

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 or R2 is better** as R-Squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variable(s) in a regression model while RSS measures the level of variance in the error term, or residuals, of a regression model. The smaller the residual sum of squares, the better your model fits your data; the greater the residual sum of squares, the poorer your model fits your data.

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: Residual sum of sqaures (RSS): The residual sum of squares (RSS) is the sum of the squared distances between actual versus predicted values:
Explained Sum of Squares(ESS) : The explained sum of squares (ESS) is the sum of the squares of the deviations of the predicted values from the mean value of a response variable, in a standard regression model.
Total Sum of Squares(TSS): The total sum if sqaures is equal to summation of residual and explained sum of squares.
The relation between all the above mentioned is :
TSS = RSS + ESS

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

Ans: Regularization is used to prevent model from overfitting. The common methods used are either LassoCV, RidgeCv and Cross validation in all models. Regularization helps to reduce the variance of the model, without a substantial increase in the bias and vice versa. This is made sure by selection of good value of cross validation which provided highest score usually very close or same as testing score.

4. What is Gini–impurity index?

Ans:It is a measure of how often a randomly chosen elemeent from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset.It is calculated by mutiplying the probability that a given observation is classified into the correct class and sum of all the probabilities when that particular observation is classfied into wrong class. Gini impurity values lies between 0 (impurity) and 1(random distribution). The node where Gini Impurity is at least is selected as root node for split.

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

Ans: Yes, decision trees are prone to overfitting. This usually happens when tree grows very deep. This is handled using either pre- pruning or post-pruning in which one stops the non significant branches from generating while in latters case, the decision trees is generated first and then the non- significant branches are removed.

6. What is an ensemble technique in machine learning?

Ans: Ensemble technique aggregate predictions from a group of predictors, which may be classifiers or regressors and most of the time prediction is better than the one obtained using a single predictor. Such algorithms are called ensemble technique in machine learning. Ensemble methods takes multiple small models and combine their predictions to obtain a more powerful predictive power.

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

Ans: Bagging: The data will be divided into numerous models such as decision trees and each tree will predict model and mean/average of this output will be final output for bagging model for each data.

Boosting: In this case it starts from a weaaker decision and keeps on building the models such that the final prediction is the weighted sum of all the weaker decision makers. The weights are assigned based on the performance of individual tree.

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

Ans: in Random Forests, when different samples are collected no sample contains all the data but a fraction of the original dataset. There might be some data which are never sampled at all. These data which is not sampled is called out of bag instances. Since the model never trains over these data, they are used for evaluating the accuracy of the model.

9. What is K-fold cross-validation?

Ans: K- Cross validation is kind of resampling technique to check whether model is overfitting or not. In this case the whole dataset is divided into k sets of equal sizes. Then the 1^{st} set is selected as the test set and is trained on remaining k-1 sets and its accuracy is calculated. Next 2^{nd} k set is taken and this continues until k times. The mean of all the accuracy is taken to find mean of accuracy of model from dataset.

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

Ans: Hyperparameter tuning done to increase accuracy of the model after the model is originally trained. In this case certain parameters are selected based on the type of models which will then increase the accuracy. This hyperparameter tuning is done using 2 methods which are GridSearchCV and RandomizedSearch CV. The range of all the parameter are fed into above mentioned algorithms and this gives the best parameters which will help in provoding best/highest accuracy for the model. This parameters are then taken and based on them model is trained again.

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 increase.

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

Ans: No, logistic regression only forms linear decision surface. Logistic Regression has traditionally been used as a linear classifier, i.e., when the classes can be separated in the feature space by linear boundaries.

13. Differentiate between Adaboost and Gradient Boosting.

Ans: While Gradinet Boosting works with different loss functions for ex. Regression, classification,modelling, etc. Adaboost on other hand works with only exponential work function. Also,AdaBoost can be interpreted from a much more intuitive perspective and can

be implemented without the reference to gradients by reweighting the training samples based on classifications from previous learners.

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

Ans: While building the machine learning model, it is really important to take care of bias and variance in order to avoid overfitting and underfitting in the model. If the model is very simple with fewer parameters, it may have low variance and high bias. Whereas, if the model has a large number of parameters, it will have high variance and low bias. So, it is required to make a balance between bias and variance errors, and this balance between the bias error and variance error is known as the Bias-Variance trade-off.

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

Ans:Linear kernels - Linear Kernel is used when the data is Linearly separable, that is, it can be separated using a single Line. It is one of the most common kernels to be used. It is mostly used when there are a Large number of Features in a particular Data Set.

Polynomial kernels- It represents the similarity of vectors in the training set of data in a feature space over polynomials of the original variables used in the kernel.

RBF kernel - When the data set is linearly inseparable or in other words, the data set is non-linear, it is recommended to use kernel functions such as RBF.