Machine Learning

Assignment – 5

1. R squared is the better measures of goodness of fit in the regression model . Because R square shows how well the data fit the regression model .
2. TSS measures how much variation there is in the observed data , while RSS measures the variation in the error between the observed data and modelled values . The ESS is the sum of squares of the deviations of the predicted values from the mean value of a response variable. The equation relating of these 3 metrics is TSS = ESS+RSS.
3. Regularization refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting . Using regularization we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.
4. The Gini impurity is a measurement used to build decision trees to determine how the features of a dataset should split nodes to from the tree.
5. Yes . Because overfitting happens when any learning processing overly optimizes training set error at the cost test error . While it's possible for training and testing to perform quality well in cross validation . In the case of decision tree's they can learn a training set to a point of high granularity that makes them easily overfit .
6. Ensemble techniques are that aim at improving the accuracy of results in models by combining multiple models instead of using a single model . The combined model increase the accuracy of the results significantly . This has boosted the popularity of ensemble techniques in machine learning .
7. (a) Bagging is a method of merging the same type of predictions . Boosting is a method of merging different types of predictions. (b) Bagging decreases variance , not bias and solves over fitting issues in a model. Boosting decreases bias ,not variance .
8. The out of bag 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 random forest classifier to be fit and validated whilst being trained .
9. 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. In machine learning , hyperparameter tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm . A hyperparameter is a parameter whose value is used to control the learning process . By contrast , the values of other parameters are learned.
11. Gradient Descent is too sensitive to the learning rate . If it is too large , the algorithm may bypass the local minimum and overshoot.
12. No we can't use logistic regression for classification of non linear data . The reason is that the target label has no linear correlation with the features . In such cases , logistic regression can't predict targets with good accuracy .
13. 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. In machine learning bias variance trade off 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. Linear in SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line , then such data is termed as linearly separable data and classifier is used called as Linear SVM classifier.
    RBF is a popular kernel function used in various kernelized learning algorithms . In particular it is commonly used in SVM. RBF is popular in SVM because of it’s similarity to K nearest Neighbourhood Algorithm.
    Polynomial kernel is a kernel function commonly used with SVM that represents the similarity of vectors in a feature space over polynomials of the original variables allowing learning of non linear models.

Statistics worksheet – 5

1.(c)

2.(c)

3.(c)

4.(b)

5.(c)

6.(b)

7.(a)

8.(a)

9.(b)

10.(a)