MACHINE LEARNING WORKSHEET 5

1. RSS is the sum of the squares of the errors made by the model on each data point. So, RSS depend upon the number of data-points in the data. If dataset is large, then naturally it will have large RSS because RSS is the sum of squares of the errors made by all the data points. While on the other hand R-squared is given as follows: Where, y i = True value of the dependent variable y ˆ = predicted value of the dependent variable y = mean value of dependent variable So, R-squared does not depend upon the number of data-points in the data, rather it depends only on the quality of the fit of the curve on the data, while RSS depends on both the quality of fit and also the number of data points in the 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. • TSS is total sum of squares. It is equal to the variance of the data. • ESS is called the explained sum of squares. It is the variance of the data which has been explained by the model. The model is able to explain this much variance of the data. To refer the formula please look above in the image. • RSS is called the residual sum of squares. It is equal to the sum of squares of all the errors or residuals made by the model on the data. The relationship between tss, rss and ess is given as: TSS = ESS + RSS

3. We need to regularize the models in machine learning so that the model does not become overly complex and overfit the training data. We want our model to capture the patterns from the training data which can be generalized to unseen data. We have applied regularization so that the model does not overfit the training data.

4. Gini - impurity index is the measure of how heterogeneous a dataset is. If a dataset has data-points belonging to more than one labels it becomes heterogeneous (Gini Index>0) and if the dataset has data-points belonging to only one class, it becomes homogenous (Gini Index=0). The formula of Gini index is given by: Where, i p = Probability of ith class

5. Unregularized decision-trees are highly prone to overfitting. If we do not restrict the depth up to which a tree can be grown or control it in any other way, the decision tree will most likely learn each and every data point in the training dataset. So, it will learn the training data patterns too closely and when it will be tested on unseen data, it will most likely perform poorly. So, in order to solve this problem, we regularize decision trees by a number of ways, either by controlling the depth of the tree, or controlling the maximum number of leaves tree can have etc.

6. In ensemble technique we combine together a number of models to get a better performing model on the dataset. The individual models in the ensemble technique act as complementary to each other, so if one model work poorly on some area, then there is some other model in the ensemble which takes up on this weakness of the previous model. So, in this way the models act as complementary to each other and overall after ensemble them we get a better model.

7. Bagging is the ensemble technique in which N decision trees are trained in parallel on N Randomly Generated datasets from the training data. The final result of the ensemble is produced by taking average of results of all the member decision trees for regression and for classification we take mode of the classes predicted by the member trees. Example-Random Forest. Boosting is the ensemble technique in which trees are trained in series rather than parallelly. In boosting, each tree works on the errors of the previous class until errors are minimized to the level we want. examples are Gradient boosting, Ada boost.

8. Out of bag error in the random forests is used to evaluate the random forest model. As we know in random forest a number of decision trees are trained in parallel on bootstrapped samples. while training a particular tree the data points which are not used for training of that tree act as unseen data. For each data point predictions are made by those trees in whose training these points were not included and error on these predictions are called out of bag error.

9. K-fold cross validation is the model evaluation technique which is generally used when we have limited data. In this technique we create k groups on the training data, train the model on k-1 groups and test the resultant model on the left-out group. In this we do it on every possible combinations of groups and then take average of the evaluation metric.

10. Hyper parameter tuning is the technique of finding best possible values of a set of hyper parameters used in model on which the model gives best performance. So, hyper parameter tuning is like fine tuning your model on the basis of hyper parameter values used in the model so that we get the best possible version of that model.

11. The two main issues which can occur if we perform gradient descent with large learning rate are: • The gradient descent algorithm can diverge from the optimal solution if we try out a very large learning rate. The algorithm can go away from the optimal solution if we have a very large learning rate. • The gradient may simply keep oscillating around the optimal solution if the learning rate is high, and it will not settle at the optimal solution.

12. We cannot use Logistic Regression for classification of Non-Linear Data because the decision boundary produced by logistic regression is linear and if we have nonlinear data where we have nonlinear decision boundaries then if we try to use the logistic regression it will perform poor on the data, as the decision boundary in the data is nonlinear.

13. Adaboost and Gradient Boosting are ensemble techniques, in which the trees are trained in series. The major difference between Adaboost and Gradient Boosting is that in Adaboost we assign weights to each of the data points of the training data and the weights changes according to the errors made by the previous tree in the series. So basically a tree in Adaboost puts more emphasis on the datapoints on which the previous tree did not perform well. While in Gradient Boosting, each tree is trained on the errors made by the previous tree, so in this way the error made on the training data keeps on decreasing as we keep increasing the trees in the series.

14. The bias variance tradeoff is the tradeoff which happens between bias (error made by the model) and variance (how much the model changes with change in training data) when the model complexity changes. If the model is too simple the model makes too much errors and its predictions becomes inaccurate, so when we decrease the model complexity the bias (or errors) increases but the variance decreases (that is the model does not change much with change in training data). On the other hand, if model is too complex the bias although becomes low but the variance increases. So, there is tradeoff between bias and variance. So, we always have to find that sweet spot where the model does not have much bias neither much high variance.

15. The SVM uses kernel functions to transform the data from one set of dimensions to another set of dimensions so that the decision boundary in the resultant space is simpler than the decision boundary in original space. Now, the kernel to be used depend upon the nature of the data we have. • Linear kernel will be used if the original data is linearly separable. • Polynomial kernel is used when the data is the form of polynomial of some degree n. • RBF is used when the data follows some complex pattern which is neither linear nor polynomial