Worksheet Set 5

Machine Learning

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 side 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. TSS or Total sum of squares tells us how much variation there is in the dependent variable. Explained sum of squares or ESS tells us how much of the variation in the dependent variable your model explained. RSS or residual sum of squares tells us how much of the dependent variable's variation your model did not explain. Explained Sum of Squares, which marks the variation in the data explained by the regression model. On the other hand, Residual Sum of Squares (RSS) defines the variations marked by the discrepancies in the dataset not explained by the estimation model. The Total Sum of Squares (TSS) defines the variations in the observed values or datasets from the mean. In contrast, the Residual Sum of Squares (RSS) assesses the errors or discrepancies in the observed data and the modeled data.
   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. Gini Index or Gini Impurity Index is a powerful measure of the randomness or the impurity or entropy in the values of a dataset. Gini Index aims to decrease the impurities from the root nodes (at the top of decision tree) to the leaf nodes (vertical branches down the decision tree) of a decision tree model.
5. Unregularized decision-trees are prone to overfitting. When we create decision tree, we need to regularize it. Regularization in terms of decision tree means to control the growth of the tree. When decision tree becomes too large, they tend to over-fit. To avoid over-fitting, we regularize the tree.
6. Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model. The combined models increase the accuracy of the results significantly.
7. Bagging and Boosting are two types of Ensemble Learning. These two decrease the variance of a single estimate as they combine several estimates from different models. So the result may be a model with higher stability. Bagging: is a homogeneous weak learners' model that learns from each other independently in parallel and combines them for determining the model average. Boosting: is also a homogeneous weak learners' model but works differently from Bagging. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm.
8. Out of bag error also called out of bag estimate is a method of measuring the prediction error of random forests, boosted decision trees and other machine learning models. Out of bag error is frequently used for error estimation within random forests. Its advantage is that it requires less computation and allows one to test the model as it is 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. K refers to the number of groups the data sample is split into. For example, if you see that the k-value is 5, we can call this a 5-fold cross-validation.
10. In machine learning, hyperparameter optimization or 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 (typically node weights) are learned. Combination of hyperparameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors.

11. If learning rate is too large, gradient descent can overshoot the minimum. It may fail to converge and even diverge.
12. Non-linear problems can't be solved with logistic regression because it has a linear decision surface. Linearly separable data is rarely found in real-world scenarios.
13. Adaboost increases the performance of all the available machine learning algorithms and it is used to deal with weak learners. It gains accuracy just above the arbitrary chances of classifying the problem. The adaptable and most used algorithm in AdaBoost is decision trees with a single level. The gradient boosting depends on the intuition which is the next suitable possible model, when get combined with prior models that minimize the cumulative predicted errors. The crucial idea of gradient boosting is to fix the targeted outcomes for the next model to reduce the error.
14. If the algorithm is too simple (hypothesis with linear eq.) then it may be on high bias and low variance condition and thus is error-prone. If algorithms fit too complex (hypothesis with high degree eq.) then it may be on high variance and low bias. In the latter condition, the new entries will not perform well. Well, there is something between both of these conditions, known as Trade-off or Bias Variance Trade-off. This trade off in complexity is why there is a trade-off between bias and variance. An algorithm can't be more complex and less complex at the same time. The 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 Kernel: It is the most basic type of kernel, usually one dimensional in nature. It proves to be the best function when there are lots of features. The linear kernel is mostly preferred for text-classification problems as most of these kinds of classification problems can be linearly separated. These are faster than other functions.
    Polynomial Kernel: It is a more generalized representation of the linear kernel. It is not as preferred as other kernel functions as it is less efficient and accurate.
    RBF: RBF or Gaussian Radial Basis Function is one of the most preferred and used kernel functions in svm. It is usually chosen for non-linear data. It helps to make proper separation when there is no prior knowledge of data.

STATISTICS WORKSHEET-5

1. d) Expected
2. c) Frequencies
3. c) 6
4. b) Chisqared distribution
5. c) F Distribution
6. b) Hypothesis
7. a) Null Hypothesis
8. a) Two tailed
9. b) Research Hypothesis
10. a) np