MACHINE LEARNING WORKSHEET5 SOLUTIONS

1. R-squared, It is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale.

2. TSS is the squared differences between the observed dependent variable and its mean, ESS is the sum of the differences between the predicted value and the mean of the dependent variable, RSS is the difference between the observed value and the predicted value. TSS = ESS + RSS

3. Regularisation is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting. The commonly used regularisation techniques are : L1 regularisation. L2 regularisation.

4. Gini index or Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen. The degree of Gini index varies between 0 and 1, where 0 denotes that all elements belong to a certain class or if there exists only one class, and 1 denotes that the elements are randomly distributed across various classes.

5. Over-fitting is the phenomenon in which the learning system tightly fits the given training data so much that it would be inaccurate in predicting the outcomes of the untrained data. In decision trees, over-fitting occurs when the tree is designed so as to perfectly fit all samples in the training data set. Thus it ends up with branches with strict rules of sparse data. Thus this effects the accuracy when predicting samples that are not part of the training set.

6. Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model. They usually produce more accurate solutions than a single model could do.

7. Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions. Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance. In Bagging, each model receives an equal weight. In Boosting, models are weighed based on their performance. Models are built independently in Bagging. New models are affected by a previously built model’s performance in Boosting. In Bagging, training data subsets are drawn randomly with a replacement for the training dataset. In Boosting, every new subset comprises the elements that were misclassified by previous models. Bagging is usually applied where the classifier is unstable and has a high variance. Boosting is usually applied where the classifier is stable and simple and has high bias.

8. Out of bag (OOB) score is a way of validating the Random forest model. It is the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample.

9. Cross Validation is a very useful technique for assessing the performance of machine learning models. It helps in knowing how the machine learning model would generalize to an independent data set. K-Fold Cross Validation is a common type of cross validation that is widely used in machine learning. In the k-fold cross validation method, all the entries in the original training data set are used for both training as well as validation. Also, each entry is used for validation just once.

10. Hperparameter settings could have a big impact on the prediction accuracy of the trained model. Optimal hyperparameter settings often differ for different datasets. Therefore they should be tuned for each dataset. Since the training process doesn't set the hyperparameters, there needs to be a meta process that tunes the hyperparameters. This is what we mean by hyperparameter tuning. Hyperparameter tuning is a meta-optimization task. Each trial of a particular hyperparameter setting involves training a model—an inner optimization process. The outcome of hyperparameter tuning is the best hyperparameter setting, and the outcome of model training is the best model parameter setting.

11. When the learning rate is too large, gradient descent can inadvertently increase rather than decrease the training error.

12. No, Logistic regression is considered a generalized linear model because the outcome always depends on the sum of the inputs and parameters. It is used when the classes can be separated in the feature space by linear boundaries.

13. Adaboost is an Boosting algorithim which increases the accuracy by giving more weightage to the target which is misclassified by the model. Gradient Boosting Algorithim increases the accuracy by minimizing the loss function(error which is difference of actual and predicted value) and having them as target for next decision tree building.

14. In statistics and machine learning, the bias–variance tradeoff is the property of a model that the variance of the parameter estimates across samples can be reduced by increasing the bias in the estimated parameters. The bias–variance problem is the conflict in trying to simultaneously minimize these two sources of error that prevent supervised learning algorithms from generalizing beyond their training set. There is no escaping the relationship between bias and variance in machine learning. Increasing the bias will decrease the variance and increasing the variance will decrease the bias.

15. **Linear:** It is the most basic type of kernel which proves to be the best function when there are lots of features. It is mostly preferred for text-classification problems as most of these kinds of classification problems can be linearly separated.

Polynomial: 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: It 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.