Assignment: Worksheet 5

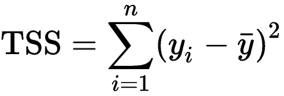
Answer 1. R-squared and Residual Sum of Squares (RSS) are both measures of the goodness of fit of a regression model, but they capture different aspects of the fit.

R-squared, also known as the coefficient of determination, measures the proportion of variance in the dependent variable that is explained by the independent variables in the model. It ranges from 0 to 1, with a higher value indicating a better fit. R-squared is useful for comparing different models or for determining the proportion of the variability in the dependent variable that is explained by the model.

On the other hand, Residual Sum of Squares (RSS) measures the total amount of unexplained variance in the dependent variable that remains after the model has been fit. It is the sum of the squared differences between the actual and predicted values of the dependent variable. A lower value of RSS indicates a better fit. RSS is useful for evaluating the accuracy of the predictions of the model.

In general, both measures are important and should be considered together when evaluating the goodness of fit of a model. However, R-squared is often considered to be a better measure of goodness of fit than RSS because it provides a single number that summarizes the proportion of variance in the dependent variable that is explained by the model, which is more interpretable and easier to compare across models.

Answer 2. TSS finds the squared difference between each variable and the mean.

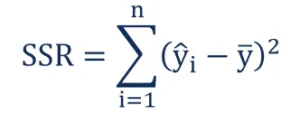


yi = the ith term in the set

ȳ= the mean of all items in the set

TSS as the dispersion of observed variables around the mean, or the variance. So, the goal of TSS is to measure the total variability of the dataset.

Explained Sum of Squares (denoted as ESS). SSR is used to describe the difference between the predicted value and the mean of the dependent variable.

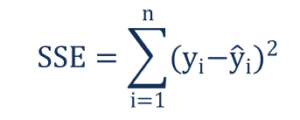


ŷi — the value estimated by the regression line

ȳ — the mean value of a sample

To start, we will again need the mean. The estimated value is the one that lies on the regression line. That means instead of the actual value of each variable, take the value of where that variable would be on the line of regression. This will tell us how well the line fits the data. If the ESS matches the TSS, then that line would be a perfect fit.

RSS finds the difference between the observed, or actual value of the variable, and the estimated value, which is what it should be according to the line of regression.



Where:

yi — the observed value

ŷi — the value estimated by the regression line.

Answer 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.

Answer 4. Gini Impurity of a dataset is a number between 0-0.5, which indicates the likelihood of new, random data being misclassified if it were given a random class label according to the class distribution in the dataset.

Answer 5. Over-fitting in decision tree is called **pruning**which is done after the initial training is complete. In pruning, you trim off the branches of the tree, i.e., remove the decision nodes starting from the leaf node such that the overall accuracy is not disturbed. This is done by segregating the actual training set into two sets: training data set, D and validation data set, V. Prepare the decision tree using the segregated training data set, D. Then continue trimming the tree accordingly to optimize the accuracy of the validation data set, V.

Answer 6. Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. This has been the case in a number of machine learning competitions, where the winning solutions used ensemble methods.

Answer 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.

Answer 8. The out-of-bag (OOB) error is the average error for each (zi)calculated using predictions from the trees that do not contain (zi) in their respective bootstrap sample. This allows the Random Forest Classifier to be fit and validated whilst being trained.

Answer 9. Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation.

Answer 10. Hyperparameter tuning consists of finding a set of optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any data set. That combination of hyperparameters maximizes the model’s performance, minimizing a predefined loss function to produce better results with fewer errors.

Answer 11. If the learning rate is very large we will skip the optimal solution.

Answer 12. No, 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. That can be remedied however if we happen to have a better idea as to the shape of the decision boundary.

Answer 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.

Answer 14. In statistics and machine learning, the bias–variance tradeoff 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.

Answer 15.