Machine Learning Assignment – 5 Answers

1. The Residual Sum of Squares is the most informative method when compared to the R^2 method. This is because the residuals can be used descriptively, usually by looking at histograms or scatter plots of residuals, and also form the basis of several other methods that we can examine.
2. **TSS(Total Sum of Squares):** In statistical data analysis the total sum of squares is a quantity that appears as part of a standard way of presenting results of such analyses. For a set of observations, yi, i≤n, it is defined as the sum over all squared differences between the observations and their overall mean.

   $$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

   ESS(Explained Sum of Squares): The ESS alternatively known as the model sum of squares or sum of squares due to regression, is the sum of squares of the deviations of the predicted values from the mean value of a response variable, in a standard regression model.

   $$\mathrm{ESS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

   RSS(Residual Sum of Squares): The RSS is a statistical technique used to measure the amount of variance in a data set that is not explained by a regression model itself. Instead, it estimates the variance in the residuals, or error term.

   $$RSS = \sum^{n}_{i=1} (y^{i} - f(x_i))^2$$

   The equation relating these 3 was:

   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, also known as Gini impurity, calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly. If all the elements are linked with a single class then it can be called pure.
5. Decision trees are prone to overfitting, especially when a tree is particularly deep. This is due to the amount of specifically we look at leading to smaller sample of events that meet the previous assumptions. This small sample could lead to unsound conclusions.
6. Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produce 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.
7.

   - While Bagging and boosting make seem similar due to the use of N learners in both techniques, They are inherently quite different. While the Bagging technique is a simple way of combining predictions of the same kind, boosting combines predictions that belong to different types.
   - In Bagging, each model is created independent of the other, But in boosting new models, the results of the previously built models are affected.
   - Bagging gives equal weight to each model, whereas in Boosting technique, the new models are weighted based on their results.

Machine Learning Assignment – 5 Answers

- Bagging gives equal weight to each model, whereas in Boosting technique, the new models are weighted based on their results.
- Bagging tends to decrease variance, not bias. In contrast, Boosting reduces bias, not variance.
- The bagging technique tries to resolve the issue of overfitting training data, whereas Boosting tries to reduce the problem of Bias.

8. The Out-of-Bag (OOB) 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 RandomForestClassifier to be fit and validated whilst being trained.
9. K-fold 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. When a specific value for k is chosen, it may be used in place of k in reference to the model, such as k=10 becoming 10-fold cross-validation.
10. Hyperparameter tuning is an essential part of controlling the behavior of a machine learning model. If we don’t correctly tune our hyperparameters, our estimated model parameters produce suboptimal results, as they don’t minimize the loss function. This means our model makes more errors.
11. A high learning rate can cause the model to converge too quickly to a suboptimal solution, whereas low learning rate can cause the process to get stuck.

    The learning rate can seen as step size n. As such, gradient descent is taking successive steps in the direction of the minimum. If the step size n is too large, it can jump over the minima we are trying to reach.
12. We cannot use Logistic regression for non-linear data because the outcome always depends on the sum of the inputs and parameters. In other words, the output cannot depend on the product of its parameters.
13.

Adaboost	**Gradient boost**
Both ada boost and gradient boost use a base weak learner and they try to boost the performance of a weak learner by iteratively shifting the focus towards problematic observations that were difficult to predict. At the end, a strong learner is formed by addition (or weighted addition) of the weak learners.	
In adaboost, shift is done by up-weighting observations that were misclassified before.	Gradient boost identifies difficult observations by large residuals computed in the previous iterations.
In adaboost “shortcomings” are identified by high-weight data points.	In gradient boost “shortcomings” are identified by gradients.
Exponential loss of adaboost gives more weights for those samples fitted worse.	Gradient boost further dissect error components to bring in more explanation.
Adaboost is considered as a special case of gradient boost in terms of loss function, in which exponential losses.	Concepts of gradients are more general in nature.

Machine Learning Assignment – 5 Answers

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.
15. **Linear kernel:** It is used when the data is linearly separable, that is, itt can be separated using a single line. It is one of the most common kernels to be used. It it mostly used when there are a large number of features in a particular data set.

    RBF(Radial Basis Function): RBF kernel is a function whose value depends on the distance from the origin or from some point.

    Polynomial Kernel: The polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity og vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.