Worksheet –set5

Machine Learning(Answers)

1. R-squared 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 variable explains collectively. Generally, a higher r-squared indicates more variability. Therefore R-squared is better for goodness of fit measure for linear regression model.
2. **TSS:** Total sum of squares gives the total variation Y. while the variance is the average of the squared sum of difference between actual values and data points, TSS is the total of the squared sum.

   Formula of TSS: $TSS = \sum (Yi - \overline{Y})^2$

   ESS: Explained sum of squares, alternatively known as the model sum of squares or sum of squared due to regression, is a quantity used in describing how well a model will explain the observed data for the process.

   Formula of ESS:

   RSS: Residual sum of squares gives us the total squares of the distance of actual points from the regression line.

   Formula of RSS: $RSS = \sum_{i=1}^{n} (y^i - f(x_i))^2$
3. Regularization refers to techniques that are used to calibrate machine learning models to minimize the adjusted loss function and prevent model from overfitting and underfitting. There are two main types of regularization techniques.

   Ridge regularization: Also known as Ridge regression, it penalizes the model based on the sum of squares of magnitude of the coefficients.

   Lasso Regilarization: Also known as Lasso regression, it penalizes the model based on the sum of magnitude of the coefficients.
4. Gini impurity is a measurement used to build Decision Trees to determine how the feature of a dataset should split nodes to form the tree. More precisely, the Gini impurity of a dataset is a number between 0—0.5, which indicates the likelihood of a new, random data being misclassified if it were given a random class label according to the class distribution in the dataset.
5. Yes, decision trees are prone to overfitting, especially when trees are particularly deep. This happens due to the amount of specificity we look at leading to smaller sample of events that meet the previous assumptions.
6. Ensemble techniques are the technique which create multiple models and then combine them to produce improved results. It usually produces a more accurate solution than a single model would.

7. Difference between Bagging and Boosting are as follows:

   Bagging: i) Training data subsets are drawn randomly with replacement from the entire training dataset.

   ii) Bagging attempts to tackle the over-fitting issue.

   iii) Every model receives an equal weight.

   iv) Objective to decrease variance, not bias.

   v) Every model is built independently.

   Boosting: i) Each new subset contains the components that were misclassified by previous models.

   ii) Boosting tries to reduce bias.

   iii) Models are weighted by their performance.

   iv) Objective to decrease bias, not variance.

   v) New models are affected by the performance of the previously developed model.
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. 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 group that a given data sample is to be split into. As such, the procedure is often called k=fold cross-validation.
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. Hyperparameter tuning takes advantage of the processing infrastructure of google cloud to test different hyperparameter configurations when training a model. It can give us optimized values for hyperparameters, which maximizes the accuracy of models predictive.
11. If the learning rate is large then it can cause undesirable divergent behavior, which means it can cause the model to converge too quickly to a suboptimal solution.
12. Logistic regression is known and used for linear classifier, so we cannot use logistic regression for classification of a non- linear data. It is because it has a linear decision surface, which is rarely found in real-world scenarios.
13. **AdaBoost:** i) In AdaBoost, shift is done by up-weighting observations that were misclassified before.

    ii) In AdaBoost 'shortcoming' are identified by high- weight data points.

    iii) Exponential loss of AdaBoost gives mor weights for those samples fitted worse.

    iv) AdaBoost is considered as a special case of Gradient boost in terms of loss function, in which exponential losses.

    GradientBoost: i)Gradient boost identifies difficult observations by large residual computed in the previous iterations.

    ii) In Gradient boost 'shortcomings' are identified by gradients.

iii) Gradient boost further dissects error components to bring in more explanation.
iv) concepts of gradients are more general in nature.

14. In machine learning 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 SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear classifier.

    RBF: in machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. It is commonly used in supporting vector machine classification.

    Polynomial kernel: In machine learning, the polynomial kernel is a kernel function commonly used with support vector machines (SVMs).