MACHINE LEARNING

ASSIGNMENT - 5

Q1 to Q15 are subjective answer type questions, Answer them briefly.

1. R-squared or Residual Sum of Squares (RSS) which one of these two is a better measure of goodness of fit model in regression and why?

Residual Sum of Squares (RSS) is a better measure as the lower the sum of squared residuals , the better the regression model is at explaining the data.

2. What are TSS (Total Sum of Squares), ESS (Explained Sum of Squares) and RSS (Residual Sum of Squares) in regression. Also mention the equation relating these three metrics with each other.

The total sum of squares (TSS) measures how much variation there is in the observed data, while the residual sum of squares (RSS) measures the variation in the error between the observed data and modeled values and the ESS is where the value estimated by the regression line .

Total sum of squares (TSS) = explained sum of squares (ESS) + residual sum of squares (RSS).

3. What is the need of regularization in machine learning?

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. What is Gini–impurity index?

Gini Impurity is a measurement used to build Decision Trees to determine how the features of a dataset should split nodes to form the tree.

5. Are unregularized decision-trees prone to overfitting? If yes, why?

Yes as ,Decision trees are prone to overfitting, especially when a tree is particularly deep. This is due to the amount of specificity we look at leading to smaller sample of events that meet the previous assumptions. This small sample could lead to unsound conclusions.

6. What is an ensemble technique in machine learning?

Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model.

7. What is the difference between Bagging and Boosting techniques?

Bagging and Boosting: Differences

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.

8. What is out-of-bag error in random forests?

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. What is K-fold cross-validation?

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. What is hyper parameter tuning in machine learning and why it is done?

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.

11. What issues can occur if we have a large learning rate in Gradient Descent?

In order for Gradient Descent to work, we must set the learning rate to an appropriate value. This parameter determines how fast or slow we will move towards the optimal weights. If the learning rate is very large we will skip the optimal solution.

12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?

No, as although the variable you are targeting in logistic regression is a classification, logistic regression does not actually individually classify things for you: it just gives you probabilities (or log odds ratios in the logit form).

13. Differentiate between Adaboost and Gradient Boosting.

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.

14. What is bias-variance trade off in machine learning?

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. Give short description each of Linear, RBF, Polynomial kernels used in SVM.

Linear Kernel:

The application of a support vector machine with a linear kernel is to perform classification or regression. It will perform best when there is a linear decision boundary or a linear fit to the data, thus the linear kernel.

RBF kernel:

Unlike linear or polynomial kernels, RBF is more complex and efficient at the same time that it can combine multiple polynomial kernels multiple times of different degrees to project the non-linearly separable data into higher dimensional space so that it can be separable using a hyperplane

Polynomial kernel:

The polynomial kernel is a kernel function commonly used to represent the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.