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?

RSS is a better measure of goodness of fit model in regression because it measures the level of variance in the error term and smaller the residual sum of squares improves fitting of data into the model.

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.

TSS: in regression, TSS is the sum of square of the difference of each data from the mean value of all the values of the target variable.

$$TSS = (Y1 – Ymean)^2 + (Y2 – Ymean)^2 + ……. (Yn – Ymean)^2$$

RSS: residual sum of squares (RSS) is a statistical technique used to measure the amount of variance in a data set. It estimates the variance in the residuals, or error term.

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

ESS: The explained sum of squares (ESS) is the sum of the squares of the deviations of the predicted values from the mean value of a response variable, in a standard regression model.

Explained SS = Σ(Y-Hat – mean of Y)^2.

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 over fitting or under fitting

4. What is Gini–impurity index?

Gini impurity index is used for the amount of probability of a specific feature that is classified incorrectly when selected randomly.

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

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?

In machine learning, ensembling technique is the method of creation of multiple models and combining them to produce improved results.

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

Bagging technique is a method of combining predictions that belong to the same type, while boosting technique combines predictions of different types. In bagging every model has equal weight but in boosting technique, models are weighted as per their performance.

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.

9. What is K-fold cross-validation?

K-fold Cross-Validation is a technique in which 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.

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

Hyperparameter tuning (or hyperparameter optimization) is the process of determining the right combination of hyperparameters that maximizes the model performance. In machine learning, hyperparameter tuning is done when the model produce sub optimal results.

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

When the learning rate is too high in Gradient descent, the algorithm may bypass the local minimum and overshoot.

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

Logistic regression cannot be used for classification of non-linear data because it is used as a linear classifier. It is used to come up with a hyperplane in feature space to separate observations that belong to a class from all the other observations that do not belong to that class.

13. Differentiate between Adaboost and Gradient Boosting.

In Ada Boost technique, each classifier has different weights assigned to the final prediction based on its performance. While, in Gradient Boost all classifiers are weighed equally and their predictive capacity is restricted with learning rate to increase accuracy.

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

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Bias-variance trade-off is tension between the error introduced by the bias and the error produced by the variance.