ASSIGNMENT - 5

MACHINE LEARNING

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?

R-squared or Residual Sum of Squares (RSS) both are a better measure of goodness of fit model in regression.

Residual Sum of Squares(RSS):

The residual sum of squares (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.

R-squared:

A statistical measure that determines the proportion of variance in the dependent variable that can be explained by the independent variable.

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: How much variance is there in the dependent variable.
ESS: How much of the variation in the dependent variable your model explained.

RSS: The residual sum of squares tells you how much of the dependent variable's variation your model did not explain. It is the sum of the squared differences between the actual Y and the predicted.

$TSS = \Sigma(Yi - \mu y)^2$

$ESS = \Sigma(\hat{y} - \mu y)^2$

$RSS = (\text{Actual Y} - \text{Predicted Y})^2$

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

Regularization techniques help reduce the chance of overfitting and help us get an optimal model.

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, they are prone to over-fitting. Decision trees are prone to over-fitting, 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.

6. What is an ensemble technique in machine learning?

Ensemble methods are techniques that create multiple models and then combine them to produce improved results

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

Bagging: It is method used to reduce variance in data.

Boosting: It is method used to reduce training errors.

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

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.

Out-of-bag error is a method of measuring the prediction error of random forests.

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.

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

Hyper-parameter tuning is choosing a set of optimal hyperparameters for a

learning algorithm.

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

A learning rate that is too large can cause the model to converge too quickly to a suboptimal solution.

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

No. because it has a linear decision surface and it cannot be predicted with logistic regression.

13. Differentiate between Adaboost and Gradient Boosting.

Adaboost: AdaBoost is the first designed boosting algorithm with a particular loss function.

Gradient Boosting: Gradient Boosting is a generic algorithm that assists in searching the approximate solutions to the additive modelling problem.

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

If the algorithm is too simple (hypothesis with linear eg:) then it may be on high bias and low variance condition and thus is error-prone.

If algorithms fit too complex (hypothesis with high degree eg:.) then it may be on high variance and low bias.

In the later condition, the new entries will not perform well. Well, there is something between both of these conditions, known as Trade-off or Bias Variance Trade-off.

This trade-off in complexity is why there is a trade-off between bias and variance. An algorithm can't be more complex and less complex at the same time.

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

Polynomial Kernel: It represents the similarity of vectors in the training set of data in

a feature space over polynomials of the original variables used in the kernel.

Linear Kernel is used when the data is Linearly separable, that is, it can be

separated using a single Line. It is one of the most common kernels to be used. It is

mostly used when there are a Large number of Features in a particular Data Set.

One of the examples where there are a lot of features, is Text Classification, as

each alphabet is a new feature. So we mostly use Linear Kernel in Text

Classification.

RBF is the default kernel used within the sklearn's SVM classification algorithm and can be

described with the following formula: where gamma can be set manually and has to be >0.