ASSIGNMENT-5

MACHINE LEARNING

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?

Ans: R2 or R-squared represents the proportion of the variance in our data which is explained by our model, the closer to one, the better the fit. So it always the takes values between 0 and 1. In other words, it represents how much our data is being explained by our model.

The Residual Sum of the Squares(RSS) is the sum of squares in other words it is the sum of the squared distances between actual and the predicted values.

The actual number we get depends largely on the scale of our response variable. Taken alone, the RSS is not so informative.

Therefore, R2 or R-squared is a better measure.

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.

Ans: Total Sum of Squares (TSS) is the dispersion of observed variables around the mean, or the variance. So it measures how much variation is there in the observed data.

Equation: $TSS=\sum_{i=1}^{n}[(y_i-\bar{y})^2]$

Where y_i is the give data point(actual point) and $\bar{y}$ is the mean.

Explained Sum of Squares(ESS) is the sum of the squares of the deviation of the predicted values from the mean value of a response variable, in a standard regression model.

The Residual Sum of the Squares(RSS) is the sum of squares in other words it is the sum of the squared distances between actual and the predicted values.

Equation:RSS=∑i=1 to n [(yi−y'i)^2]

Where yi is the give data point(actual point) and y'i is the fitted value of yi(predicted).

The equation relating these three metrics with each other is expressed as:

TSS=ESS+RSS

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

Ans: When we use regression models to train some data, there is a good chance that the model will over fit the given training data set. So Regularization helps in sorting out this over-fitting problem by restricting the degrees of freedom of a given equation that is simply reducing the number of degrees of a polynomial function by reducing their corresponding weights.

To regularize the model, shrinkage penalty is added to the cost function.

4. What is Gini–impurity index?

Ans: It is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. It is calculated by multiplying the

probability that a given observation is classified into the correct class and sum of all the probabilities when the particular observation is classified into wrong class. Gini impurity or Gini index value lies between 0 and 1, 0 being no impurity and 1 denoting random distribution. The node for which the Gini impurity or Gini index is least selected as the root node to split.0

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

Ans: Yes, unregularized decision trees 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. This small sample could lead to unsound conclusions.

But unlike other algorithms decision tree does not use regularization to fight against over-fitting. Instead it uses pruning. There are mainly two types of pruning performed:-

Pre-pruning(Forward Pruning) that stop growing tree earlier, before it perfectly classifies the training set.

Post-pruning(Backward Pruning) that allows the tree to perfectly classify training set and then post prune the tree.

6. What is an ensemble technique in machine learning?

Ans: We regularly come across many game shows on t.v and you must have noticed an option of 'Audience poll'. Most of the times a contestant goes with the option which has the highest vote from the audience and most of the times they win.

Ensemble technique has a similar underlying idea where we aggregate predictions from a group of predictors which may be Classifiers or Regressors and most of the time prediction is better than one obtained using a single predictor. Such algorithms are called Ensemble methods and such predictors are called Ensembles.

In simple words combining multiple algorithms and taking one decision to improve the overall performance. Bagging and Boosting are most used techniques in machine learning.

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

Ans: Bagging is a type of ensemble technique in which a single training algorithm is used in different subsets of the training data where the subset sampling is done with replacement(Bootstrap). Once the algorithm is trained on all the subsets, then bagging makes the prediction by aggregating all the predictions made by the algorithm on different subsets. In case of regression, bagging prediction is simply the mean of all the predictions and in the case of classifier, bagging prediction is the most frequent prediction(majority vote) among all the predictions.Bagging is also known as parallel model since we run all models parallely and combine their results at the end.

Boosting is an ensemble approach (it involves several trees) that starts from a weaker decision and keeps on building the models such that the final prediction is the weighted sum of all the weaker decision-makers. The weights are assigned based on the performance of an individual tree.

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

Ans: Out-of-bag error, also called Out-of-bag estimate, is a method of measuring the prediction error of random forests, boosted decision trees and other machine learning models utilizing bootstrap aggregation.

9.What is K-fold cross-validation?

Ans: K-fold cross-validation method is used to tackle the high variance of hold-out method. The idea is simple, divide the the whole dataset into 'k' sets preferable of equal sizes. Then first set is selected as the test set ans the rest 'k-1' sets are used to train the data. Error is calculated for this particular dataset. Then the steps are repeated, that is the second set is selected as the test data and the remaining 'k-1'sets are used as the training data. Again the error is calculated. Similarly the process continues for 'k' times. In the end, CV error is given as the mean of the total errors calculated individually. The variance in error decreases with increase in 'k'. The disadvantage of k-fold CV is that it is computationally expensive as the algorithm runs from the scratch for k-times.

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

Ans: While defining parameters, often values are not the ones that give the best result. In machine learning, hyper parameter tuning optimization or tuning is the problem of choosing a set of optimal hyper parameters of a learning algorithm. A hyper parameter is parameter whose value is used to control the learning process. By contrast, the values of other parameters(typically node weights) are learned. It is used to improve the accuracy.

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

Ans: Learning rate that is too large can cause the model to converge too quickly to a sub-optimal solution.

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

Ans: No, logistic regression only forms linear decision surface. Logistic Regression has traditionally been used as a linear classifier, that is when the classes can be separated in the feature space by linear boundaries.

13. Differentiate between Adaboost and Gradient Boosting.

Ans: Gradient Boosting is a generic algorithm to find approximate solutions to the additive modelling problem, while Ada Boost be a special case with a particular loss function. Hence, Gradient Boosting is much more flexible.

On the other hand, Ada Boost can be interpreted from a much more intuitive perspective and can be implemented without the reference to gradients by re-weighting the training samples based on classifications from previous learners.

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

Ans: If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then its going to have high variance and low bias. So we need to find the right/good balance without over-fitting and

under fitting the data. This trade off in complexity is why there is a trade off between bias and variance . An algorithm cant be more complex and less complex at the same time.

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

Ans: 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 many features in a particular Data set.

Gaussian RBF(Radial Basis Function) is another popular kernel method used in SVM models for more. RBF kernel is a function whose value depends on the distance from the origin or from some point.

In the **Polynomial Kernel,** we simply calculate the dot product by increasing the power of the kernel.