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:

R-squared - R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable. So, if the R2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

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.

The residual sum of squares (RSS) is the absolute amount of explained variation, whereas R-squared is the absolute amount of variation as a proportion of total variation.

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 variables explain collectively.

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:

TSS - The Sum of Squared regression is the sum of the differences between the predicted value and the mean of the dependent variable. The Sum of Squared Error is the difference between the observed value and the predicted. In statistical data analysis the total sum of squares (TSS or SST) is a quantity that appears as part of a standard way of presenting results of such analyses.

ESS –It is alternatively known as the model sum of squares or sum of squares due to regression is a quantity used in describing how well a model, often a regression model, represents the data being modelled. In particular, the explained sum of squares measures how much variation there is in the modelled values and this is compared to the total sum of squares (TSS), which measures how much variation there is in the observed data, and to the residual sum of squares, which measures the variation in the error between the observed data and modelled values.

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

RSS – see question 1

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

Ans:

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.

4. What is Gini–impurity index?

Ans:

Gini Index, also known as Gini impurity, calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly. If all the elements are linked with a single class then it can be called pure.

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

Ans:

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?

Ans:

Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model. To better understand this definition let’s take a step back into ultimate goal of machine learning and model building

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

Ans:

Bagging is the simplest way of combining predictions that belong to the same type while Boosting is a way of combining predictions that belong to the different types. Bagging aims to decrease variance, not bias while Boosting aims to decrease bias, not variance.

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

Ans:

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 Random Forest Classifier to be fit and validated whilst being trained

9. What is K-fold cross-validation?

Ans:

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 groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation.

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

Ans:

Hyperparameters in Machine learning are those parameters that are explicitly defined by the user to control the learning process. These hyperparameters are used to improve the learning of the model, and their values are set before starting the learning process of the model.

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?

Ans:

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.

Learning Rate: This is the hyperparameter that determines the steps the gradient descent algorithm takes. Gradient Descent is too sensitive to the learning rate. If it is too big, the algorithm may bypass the local minimum and overshoot.

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

Ans:

It can only be used to predict discrete functions. Hence, the dependent variable of Logistic Regression is bound to the discrete number set. It is very fast at classifying unknown records. Non-linear problems can't be solved with logistic regression because it has a linear decision surface.

13. Differentiate between Adaboost and Gradient Boosting.

Ans:

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?

Ans:

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.

It helps optimize the error in our model and keeps it as low as possible. An optimized model will be sensitive to the patterns in our data, but at the same time will be able to generalize to new data. In this, both the bias and variance should be low so as to prevent overfitting and underfitting.

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 a Large number of Features in a particular Data Set.

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.

In machine learning, the polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.