�
�

FLIP ROBO _ **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?

Ans – The residual standard error (RSE) is another statistical term used to describe the difference in standard deviation of observed values versus predicted values as shown by points in agregression analysis. It is a goodness of fit measure that can be used to analyze how well a set of data points fit with the actual model.

RSE is computed by dividing the RSS by the number of observations in the sample less 2, and then taking the square root: RSE = [RSS/(n-2)]1/2

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 – The sum of squares total (SST) or the total sum of squares (TSS) is the sum of squared differences between the observed dependent variables and the overall mean. Think of it as the dispersion of the observed variables around the mean—similar to the variance in descriptive statistics. But SST measures the total variability of a data set, commonly used in regression analysis and Anova.

Mathematically, the difference between variance and SST is that we adjust for the degree of freedom by dividing by n–1 in the variance formula.

SST=n∑i=1(yi−¯y)2

The sum of squares due to regression (SSR)or explained sum of squares (ESS) is the sum of the differences between the predicted value and the mean of the dependent variable. In other words, it describes how well our line fits the data.

The SSR formula is the following:

SSR=n∑i=1(^yi−¯y)2

Where:

^yi – the predicted value of the dependent variable

¯y – mean of the dependent variable

If SSR equals SST, our regression model perfectly captures all the observed variability, but that’s rarely the case.

The sum of squares error (SSE) or residual sum of squares (RSS, where residual means remaining or unexplained) is the difference between the observed and predicted values.
The SSE calculation uses the following formula:

SSE=n∑i=1ε2i

Where εi is the difference between the actual value of the dependent variable and the predicted value:
εi=yi−^yi

As mentioned, the sum of squares error (SSE)is also known as the residual sum of squares (RSS), but some individuals denote it as SSR, which is also the abbreviation for the sum of squares due to regression.

Although there’s no universal standard for abbreviations of these terms, you can readily discern the distinctions by carefully observing and comprehending them.

Mathematically, SST=SSR+SSE

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

Ans – Regularization is one of the most important concepts of machine learning. It is a technique to prevent the model from over fitting by adding extra information to it.
Sometimes the machine learning model performs well with the training data but does not perform well with the test data. It means the model is not able to predict the output when deals with unseen data by introducing noise in the output, and hence the model is called over fitted. This problem can be deal with the help of a regularization technique.

This technique can be used in such a way that it will allow to maintain all variables or features in the model by reducing the magnitude of the variables. Hence, it maintains accuracy as well as a generalization of the model.

4. What is Gini–impurity index?

Ans – The Gini Index or Gini Impurity is calculated by subtracting the sum of the squared probabilities of each class from one. It favors mostly the larger partitions and are very simple to implement.
In simple terms, it calculates the probability of a certain randomly selected feature that was classified incorrectly.

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

Ans – Decision trees are prone to over fitting when they capture noise in the data. Pruning and setting appropriate stopping criteria are used to address this assumption.

6. What is an ensemble technique in machine learning?

Ans – In this ensemble technique, machine learning professionals use a number of models for making predictions about each data point.

The predictions made by different models are taken as separate votes. Subsequently, the prediction made by most models is treated as the ultimate prediction.

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

Ans – 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?

Ans – A Random Forest is a supervised learning algorithm that works on the concept of bagging. In bagging, a group of models is trained on different subsets of the data set, and the final output is generated by collating the outputs of all the different models. In the case of random forest, the base model is a decision tree.

9. What is K-fold cross-validation?

Ans – K-fold cross-validation is a technique for evaluating predictive models. The data set is divided into k subsets or folds. The model is trained and evaluated k times, using a different fold as the validation set each time. Performance metrics from each fold are averaged to estimate the model's generalization performance.

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

Ans – Hyper parameter tuning consists of finding a set of optimal hyper parameter values for a learning algorithm while applying this optimized algorithm to any data set. That combination of hyper parameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors. Hyper parameters are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning.

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

Ans – If the learning rate is too high, the algorithm may overshoot the minimum, and if it is too low, the algorithm may take too long to converge. Over fitting gradient descent can over fit the training data if the model is too complex or the learning rate is too high.

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

Ans – The reason is that the target label has no linear correlation with the features. In such cases, logistic regression (or linear regression for regression problems) can't predict targets with good accuracy (even on the training data).

13. Differentiate between Adaboost and Gradient Boosting.

Ans – In the case of Ada Boost, the shifting is done by up-weighting observations that were classifieds before, while Gradient Boosting identifies the difficult observations by large residuals computed in the previous iterations.

The adaptable and most used algorithm in AdaBoost is decision trees with a single level. The gradient boosting depends on the intuition which is the next suitable possible model, when get combined with prior models that minimize the cumulative predicted errors.

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

Ans – In statistics and machine learning, the bias–variance tradeoff describes the relationship between a model's complexity, the accuracy of its predictions, and how well it can make predictions on previously unseen data that were not used to train the model.

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

Ans – A kernel function is a crucial component of Support Vector Machines (SVMs), a popular machine learning algorithm used for classification and regression tasks. The choice of the kernel function influences the SVM's ability to effectively separate data points with different class labels.

1. Linear Kernel Function:
A linear kernel function, such as the Linear SVC in SVM, assumes that the data can be separated by a straight line or hyperplane in the input feature space. It computes the dot product between two data points in the original feature space, effectively measuring their similarity. The decision boundary created by a linear kernel is a straight line or hyperplane.

2. Non-linear Kernel Function:
In cases where the data is not linearly separable in the original feature space, a non-linear kernel function comes into play. Non-linear kernels transform the input features into a higher-dimensional space, where linear separation becomes possible. Examples of non-linear kernel functions include the Polynomial, Gaussian Radial Basis Function (RBF), and Sigmoid kernels.