MACHINE LEARNING (WORK SET 5)

QUESTION 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?

ANSWER 1

R-squared is better than residual sum of squares (RSS) because r squared is scaled variant statistics that provide proporation of variation in target variables. RSS is residual sum of square of difference in actual value and predicted value, but rss is dependent on scale of target variable. If we can the scale of target variable than rss value will change.

QUESTION 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.

ANSWER

TSS is the sum of square of difference of each data point from the mean value of all the values of target variable (y).

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

ESS is a statistical quantity used in modeling of a process, ESS gives an estimate of how well a model explains the observed data for the process.

ESS = total sum of squares – residual sum of squares

ESS= TSS-RSS

RSS (Residual Sum of Squares)

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.

The smaller the residual sum of squares, the better your model fits your data; the greater the residual sum of squares, the poorer your model fits your data

$$residual\ sum\ of\ squares = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

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

Answer 3

When model is trained than model can get overfitted or underfitted, that is while in training phase it perform well but when it goes through live data than it’s performance gets poor. To avoid this situation regularization technique is used. Regularization techniques help reduce the possibility of overfitting and help us obtain an optimal model.

There are mainly two type of regularization technique.

Lasso (L1)

Ridge (L2)

Question 4. What is Gini–impurity index ?

Answer

Gini impurity index is measure of degree of impurity present. 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.

Gini basically works for categorical data,it does not work for continous data.

Gini Impurity of a dataset is a number between 0-0.5.

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

Answer:-

Yes unregularized decision trees prone to overfitting, when any parameter is not provided to decision tree than in training phase it work to last nodes or we can say to maximum depth this leds to overfitting in decision tree. To avoid overfitting problem we have to put constrain on maximum depth so that tree may not go to last node.

Question 6 What is an ensemble technique in machine learning?

Answer

Ensemble technique in machine learning is a technique that combines several basic model to produce one optimal predictive model .it is used to facalitate accurate and improve decisions.

There are 3 simple ensemble technique

Max voting

Averaging

Weighted averaging

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

Answer

Bagging and Boosting both are ensemble techniques

In Bagging it learns from every model in parallel and reach to final conclusion.

Bagging is a method of merging the same type of predictions and it decrease variance but not bias and solve overfitting model.

Boosting it learns from every model sequentially and reach to final conclusion.

It merges different type of predictions and it decrease bias and not variance.

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

Answer

This out-of-bag score is used in random forest,

Out-of-bag score is measures of samples that are not used in training of the model. these samples are left as samples were choosen randomly.

Out-of-bag error are the estimates of performance of random forest classifier and regressor, The OOB error is computed using the samples that were not included in the training of the individual trees.

QUESTION 9 What is K-fold cross-validation?

ANSWER

K-fold is cross validation technique that divides the data set in k number of times,k

can be any value but generally we take 5 or 10.

Than k-1 folds are trained and remaining one is tested.each time we trained different no. of folds than previous one.

This training and testing is done k no. of times and for each validation result is save. Finally average is taken for each results.

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

ANSWER

Hyperparameter tuning is obtaining the best values of the set of parameters to improve the performance of the model.
Hyperparameter tuning cannot be done directly while model is learning from regular training process.so these parameters are fixed before actually learning process starts

so that model can perform to it best optimization.

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

Answer

When learning rate is too large than it is possible that training error increases and it effect the performance of the model. when learning rate is too large than gradient descent do not reaches to its minima where there is very less error or nearly to zero.

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

ANSWER

Logistic regression for Non-linear data cannot be used because it is used as linear classifier and it has linear decision surface. It divides

the observations from one class to another that is either observation belongs to one class or it may not belong to that class. Decision boundary is linear in logistic regression.

QUESTION 13. Differentiate between Adaboost and Gradient Boosting

ANSWER

Adaboost is ensemble technique that combine many weak learner to one strong learner where each weak learner develop decision stamps that are use to classify the observations. all classfier have different weight.

Gradient boost is also ensemble technique that combine many weak learner to one strong learner but in gradient boost residual of current classifier becomes the input of next classifier. Here all classifier have same weight.

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

ANSWER

Bias is the difference between the predicted value and actual value. Higher the bias value gives larger error in training data and testing data. Algorithm must have low bias value than it perform well

Variance is spreadability of data points. In other word we can say it measure how data points are spread in the data set.

So if variance is high than model perform well in training data but its performance decreases in testing data.

If the algorithm is too simple than bias will be high and variance will be low but if algorithm is too complex than bias will be low and variance will be high.

There is something in between both of these conditions that is bias variance trade off. An

algorithm can’t be more complex and less complex at the same time.

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

Answer.

Linear Kernal :- it is most basic type of kernal.it is one dimentional in nature used when there lots of features.it is mostly preferred for text based classification problems.

Polynominal kernal :- it is popular in image processing . 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.

RBF:-it is used for transformation of data when there is no prior knowledge about the data.it adds radial basic method for transformations of data.