Statistics worksheet 5

Q1 (d)

Q2 (c)

Q3 (c)

Q4 (b)

Q5 (c)

Q6 (b)

Q7 (a)

Q8 (a)

Q9 (b)

Q10 (a)

Machine learning

Ans1) One of generic way to evaluate the fit of a linear model is by compting the R-squared value.It explains the proportion of variance in the observed data.R squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variables in a regression model.The value of r-squared lies between 0 and 1.A value closer to 1 is better as it means that more variance is explained by the model.

RSS-Residual sum of squares

Also known as sum of squared estimate of errors ,is the sum of squares of residuals .It is measure of the discrepancy between the data and an estimation model.

R squared is a useful metric for multiple regression,a higher r-squared indicates more variability is explained by the model .R-squared is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable.

R-squared is a goodness of fit measure for linear regression models .This statistic indicates the percentage of variance in the dependent variable that the independent variables explain collectively.Also it is a popular metric used for evaluating the performance of linear regression models.

Ans2)TSS(Total sum of squares) :it is used to determine the variation and is calculated by summing the squared difference between the observations and the mean.

Formulae=>TSS=square of(summation of(Xi-mean of X))

Where summation is adding up ,Xi is observations

ESS:-ESS is explained sum of squares ie total sum of squares minus sum of squared residuals.

Formulae: TSS-RSS/TSS

Where RSS is residual sum of squares and TSS is Total sum of squares.

RSS: residual sum of squares calculates the degree of variance in a regression model .It estimates the level of error in the model. The smaller the RSS the better is the model

Formulae –RSS=(square of e1)+(square of e2)+(square of e3)

Where e is error or residue

Tss(total sum of squares)=ESS+RSS

Ans 3)When we use regression models to train some data ,there is good chance that the model will overfit the given training dataset .Regularization helps sort this overfitting problem by restricting the degrees of freedom of a given equation ie simply simply reducing the number of degrees of a polynomial function.

Different types of regularization are:

Lasso

Ridge

Elasticnet (less popular)

Ans 4)gini impurity is a measure of how often a randomly choosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset .It is calculated by multiplying the probability of each class from 1.Gini impurity value lies between 0 and 1.0 indicated no impurity and 1 denotes random distribution .It favours larger partitions and very simple to implement.

It calculates the probability of a certain random selected features that was classified incorrectly.

Ans5)Decision tree are prone to overfitting ,especially when the tree is particularly deep.Decision trees are one of such models which have low bias bt high variance that is why decision trees tend to overfit the data .

Ans 6)Ensemble techniques aims at improving the accuracy of results in models by combining the multiple models instead of using a single model .The combined model increase the accuracy of the results .The most popular ensemble methods are boosting and bagging.It is ideal for both regression and classification,where they reduce bias and variance to increase the accuracy of the models.

We aggregate predictions from a group of predictors ,which may be classifier or regressor and most of the times the prediction is better than the one obtained using a single predictor .such a algorithm are called ensemble methods and predictors are class ensembles.

Ans 7)BAGGING:is a type of ensemble technique in which a single training algorithmis used on different subsets of training data where the subset sampling is done with replacement(bootstep).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 the case of regression,bagging prediction is simply the mean of all the predictions and in the case of classifier ,bagging predictions is the most frequent prediction (majority vote)among all the predictions .

Bagging is also known as parallel model since we run all the models parallel and combine their results at the end.

Where as Boosting:is an ensemble technique that starts from the weaker decision and keeps on building the models such that the final prediction is the weighted sum of all the weaker decision makers.Performance is based on a individual tree.

Ans8)In bagging ,when different samples are collated,no samples contains all the data but a fraction of original dataset.There might be some data which are never sampled at all.The remaining data which are not sampled at all are called out of bag instances .Since the model never trains over these data,they can be used for evaluating the accuracy of the model.

Ans9)To tackle the high variance k-fold cross validation method is used.The basic idea behind this is very simple ie divide the dataset into ksets preferably of equal sizes .Then the first set is selected as the test set and the rest k-1 sets are used to train the data .Error is calculated for this particular dataset.Then the steps are repeated ie the second set is selected as the test dataset and the remining k-1 sets are used as the training data .At the end the cv error (mean of the total errors calculated individually . The variance in error decreases with the increase in k.

Ans10)Hyperparameter tuning improves the performance of the model and minimizes the chances of loss .It is a parameter whose value is set before the machine learning process begins.

They are important as it impacts the performance of the model.GridSearchis the basic hyperparameter tuning method .

Ans11) Learning rate affect how quickly our model can converge to a local minima if the learning rate is large in Gradient Descent ,it takes too longer to train.A large learning rate allows to learn faster and the accuracy is affected.

Ans12)No,Logistic Regression doesn’t use for non linear data because it seperates obsevations that belong to a particular class from all other observations that do not belong to that class.

Ans 13)Adaboost is the first boosting algorithm with a loss function whereas Gradient Boosting searches the solutions to the problems.

Ans 14)Bias variance tradeoff is a situation that involves losing one quality for gains in other forms . Simply we can say that one thing increases and other decreases.

Ans 15)Linear SVM:It is used when data can be linearly seperable ie it can be separated using a single line . It is used when there are somany features.

RBF SVM:It changes with distance from location.It is somewhat similar to kfold .

Polynomial Kernels SVM:It represents the similarity of vectors in the training set of data.

#####thank you##################3