MACHINE LEARNING SUBJECTIVE SET-5

ANS.1 –

Both r-squared and Residual Sum of squared are measures of Goodness of fit in regression analysis, but they capture different aspects of the model's performance.

R-squared measures the proportion of variation in the dependent variable that is explained by the independent variables in the model.

In other word, it indicates how well the model fits the data with values ranging from 0 to 1.

Higher R-squared values indicates a better fit, as they mean that a larger proportion of the variation in the dependent variable is explained by the independent variables in the model.

On the other hand, RSS measure the total sum of squared difference between the actual values of the dependent variable and the predicted value by the model. It represents the amount of unexplained variation in the data, and lower RSS values indicates a better fit as they mean that the model is able to explain more of the variation in the data.

Therefore both measures are useful in evaluating the goodness of fit of a model but they sever different purpose. R-squared is a useful measures to access the overall fit of the model and to compare different model, while RSS is useful to identify the degree of the error in the model's predictions.

In general, a good model should have both a high R-squared value and a low RSS value, indicating that it explain a larger proportion of the variation in the dependent variables and has a low degree of error in its prediction, however in some cases one measure may be more important that the other, depending on the research and the nature of the data being analysed.

ANS.NO.12

Before reading it, though make sure you are not mistaking regression for correlation. If you have got this checked we can get straight into the action.

There are three term must define.

1) TSS (Total Sum Of Squares)
2) ESS (Explained Sum Of Squares)
3) RSS (Residual Sum Of Squares)

1) TSS=: The Total sum of square denoted TSS. Is the squared difference between the observed dependent variables and its mean you can think of this as the dispersion of the observed variables around the mean. Like the variance in descriptive statistics.

It is measure of the total variability of the dataset.

$$\sum$$

2) Explained Sum of Squares=: The term is the sum of the squares of the deviations of the predicted value from the mean value of a response variable in a standard regression model.
ESS is the I predicted value of the response variable.

ANS.NO-3

The concept of regularization is widely used even outside the machine learning domain. In general regularization involves augmenting the input information to enforce generalization. Of course the fancy definition and complicated terminologies are of little worth to a beginner. Instead of beating ourselves over it.

To understand the importance of regularization particular in the machine learning domain, let us consider two extreme cases. An Under fit model and an over fit model. In an under fit model, the machine learning model will be simple but unable to perform well on the training and test data set as it is not

complex enough to represent the patterns in the given data. It can be pictorially represented.

There are two main types regularization.

1) Ridge regularization
2) Lasso (L1) regularization

Ridge regularization is also known as it modifies the over-fitted models by adding the penalty equivalent to the sum of the squares of the magnitude of coefficients.
This means that mathematically function representing our machine learning model is minimized and coefficients are calculated. The magnitude of coefficient is squared and added. Ridge regression perform regularization by shrinking the coefficient present. The function depicted below shows the cost function of ridge regression.
In the cost function the penalty term is represented by lambda.by changing the values of the penalty function we are controlling the penalty term. The higher the penalty it reduces the magnitude of coefficients. It shrinks the parameters therefore it is used to prevent multicollinearity and it reduces the model complexity by coefficient shrinkage.

Cost function = lasso+ lambda multiply sigma w^2

For linear Regression line let's consider two points that are on the line.
Lasso = 0
Lambda= 1
W=1.4
0+1*1.42
1.96

For Ridge Regression, let's assume
Lasso = 0.32+0.22= 0.13
Lambda = 1
W=0.7
Cost function =0.13 +1*0.72
0.62

2) Lasso Regression:-
It modifies the over-fitted or under-fitted models by adding the penalty equivalent to the sum of the absolute values of coefficients.
Lasso regression also performs coefficients minimization, but instead of squaring the magnitudes of the coefficients it takes the true values of coefficients. this means that the

coefficient sum can also be 0, because of the presence of negative coefficients.

Cost function = lasso + lambda *sigma!!w!!
For linear regression line
Lasso = 0
Lambda =1
W = 1.4
The cost function = 0+1*1.4
Ans.1.4

For ridge Regression

Loss = 0.32+0.12 =0.1
Lambda =1
W= 0.7
The cost function =0.1+1*0.7
Ans.0.8

ANS.NO.4

Gini- impurity index: - Gini impurity index measures the diversity in a set. For example a full of basket of several colours. A baskets where all the balls have the same colour is low Gini impurity

index. A baskets where all the balls have different colour has a very high Gini impurity index.

ANS.NO- 5

Over-fitting can be one problem that describe if your model no longer generalizes well.
Over-fitting happens when any learning processing overly optimizes training set error at the cost test error. While it's possible for training and testing to perform equality well in cross validation, it could be as the result of the data being very close in characteristics which may not be a huge problem. In the case of decision tree's they can learn a training set to a point of high granularity that makes them easily over-fit.

1) Use a test set that is not exactly like the training set, or difference enough that error rates are going to be easy.
2) Ensure you have enough data.
3) Reduce the complexity of the decision tree model.
4) Use decision tress in an ensemble.
5) Reduce the dimensionality of your data.

ANS.NO.6

Ensemble Technique: - Ensemble technique helps improve machine learning results by combining several models. This approaches allows the production of a better predictive performance compared to a single models. Basic idea is to learn a set of classifier.

1) Statistical problem
2) Computation problem
3) Representation problem

Statistical problem: - the statistical problem arises when the hypothesis space is too larger for the amount of available data. Hence there are many hypothesis with the same accuracy on the data and the learning algorithm.

Computation Problem: - The computation problem arises when the learning algorithm cannot guarantees finding the best hypothesis.

Representation Problem: - The representation problem arises when the hypothesis space does not contain any good approximation of the target class.

Types of Ensemble classifier.

1) Bagging
2) Random forest.

Bagging: -

1) Bagging is used to reduce the variance of a decision tree.
2) A training set D of d tuples is sampled with replacement from D.
3) Then a classifier model M is learned for each training set D<i.
4) Model M returns its class prediction.
5) The bagged classifier M*counts the votes and assigns the class with the most votes to X.

Random Forest: -

1) Random forest is an extension over bagging.
2) Each classifier in the ensemble is a decision tree classifier and is generated using a random selection of attributes.
3) Multiple subset create from the original dataset.

4) Each tree votes and the most popular class is returned.
5) The tree is grown to the largest.
   The goal of machine learning problem is to find a single model that will best predict our wanted outcome. Rather than making one model and hoping this model is the best accurate predictor. We can make ensemble methods take a myriad of models into account and average.

ANS.NO-7

Difference between bagging and boosting technique.

BAGGING: -

1) Bagging is a method of merging the same type of predictions.
2) Bagging decreases variance not bias.
3) Bagging each models receives an equal weight.
4) Models in bagging are built independently.
5) In bagging training data subset are drawn randomly with a replacement for the training set.
6) Bagging is usually applied where the classifier is unstable and has a high variance.

BOOSTING: -

1) Boosting is a method of merging different types of predictions
2) Boosting decreases bias not variance.
3) Boosting models are weighed based on their performance.
4) Boosting models are affected by a previously built models performance.
5) Boosting new subset comprises the elements that were misclassified by previous model.
6) Boosting is usually applied where the classifier is stable and simple and has high bias.

ANS.NO-8

Out-of-bag error also called out-of-bag estimate, is a method of measuring the prediction error of random forest, boosted decision tree and other machine learning models utilizing bootstrap aggregating.

Bagging uses subsampling with replacement to create training samples for the model to learn from. OOB error is the mean prediction error in each training sample.

OUT-OF-BAG-DATASET: -

When bootstrap aggregating is performed, two independent sets are create.one set, the bootstrap sample is the data chosen to be “in-the-bag”

Sampling with replacement. The out of bag set is all data not chosen in the sampling process. When the process is repeated.

Calculating Out-Of-Bag Error: -

Out-Of-Bag set is not used to train the model. It is a good test for the performance of the model. The OOB error is depends on the implementation of the model.

1) Find all models in the case of a random forest that are not trained by the OOB instance.
2) The majority vote of these models result for the OOB instance compared to the true value of the OOB instance.
3) Compile the OOB error for all instances in the OOB dataset.

OUT-Of-Bag Error is used frequently for error estimate with in random forests. Small sample sizes, a large number of predictor variables, small correlation between predictors, week effects.

ANS.NO-9

Cross validation also being referred to as rotation estimation sample testing.

Split the dataset into training data and test data.

The different types of cross-validation technique.

1) Hold-out
2) K-folds
3) Leave-one-out
4) Leave-p-out

K-Fold Cross-validation

1) K-fold cross-validation is when the dataset is a split into a K number of folds and used to evaluate the model's ability when given new data.
2) Choose your k value
3) Split the dataset into the number.
4) Train the model on the training dataset and validation it on the test dataset.
5) Save the validation score.

ANS.NO-10

Hyper parameter tuning: -

1) Hyper parameter tuning is an essential part of controlling the behaviour of a machine learning model.
2) 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.

HYPER PARAMETER TYPES.

- Number of hidden layers.
- Number of Nodes.
- Learning rates
- Momentum

1) Number of hidden layers: - it's a trade-off between keeping our neural network as simple as possible. We can start with value of four to six and check our data's prediction accuracy.

2)Number of Nodes: -More isn’t always better when determine how many neurons to use per layer. Increasing neuron count can help up to a point. But layers that are too wide may memorize the training dataset, causing the network to be less accurate on new data.

3)Learning rate: - Model parameter are adjusted iteratively and the learning rate controls the size of the adjustment at each steps. The lower the learning rate. The lower the changes to parameter estimates.

4)Momentum: - Momentum helps us avoid falling into local minima by resisting rapid changes to parameter values. It encourage parameter to keep changing in the direction they were already changing.

ANS.No-11

The learning rate is an important hyper parameter that greatly affects the performance of gradient descent. it

determines how quickly or slowly our model learns and it play an important role in controlling both convergence and divergence of the algorithm. When the learning rate is too large, gradient descent can suffer from divergence. This means that weight increase exponentially resulting in exploding gradients which can cause problem such as instabilities and overly high loss-value. In order avoid these issues with different learning rates for each parameter. We use adaptive technique. These adaptive measures ensure better results than standard gradient descent while avoiding potential pitfalls in term of either massive gains or slow losses.

ANS.NO-12

Logistic Regression has traditionally been used as a linear classifier when the classes can be separated in the feature space by linear boundaries. That can be remedies however if we happen to have a better idea as to the shape of the decision boundary.

Logistic regression is known and used as a linear classifier.it is used to come up with a hyper plane in feature space to separate observation that belong to a class from all the other observation that do not belong to that class. The decision boundary is thus linear.

ANS.NO-13

Differentiate between Adaboost and Gradient Boosting.

ADABOOST

1) The shift is made by up-weighting the observation that are miscalculated prior.
2) The tree are called decision stumps.
3) Every classifier has different weight assumption to its final prediction that depends on the performance.
4) Maximum weighted data points are used to identify the shortcomings.

5) The exponential loss provides maximum weights for the samples which are fitted in worse conditions.

GRADIENT BOOSTING

1)It identifies complex observation by huge residual calculated in prior iterations.
2)The tree with week learners are constructed using a greedy algorithm based on split points and purity scores.
3)The classifiers are weighted precisely and their prediction capacity is constrained to learning rate and increasing accuracy.
4)The gradient themselves identify the shortcomings.
5)Gradient boosting cut down the error components to provide clear explanations and its concepts are easier to adapt and understand.

ANS.NO-14

It is important to understand prediction errors bias and variance. When it comes to accuracy in any machine learning algorithm. There is a trade-off between a model's ability to minimize bias and variance. Which is referred to as the best solution for selecting a value of Regularization constant.

Bias: -

The bias is known as the difference between the Prediction of the values by the machine learning model and the correct value. Being high in biasing gives a larger error in training as well as testing data. By high bias the data predicted is in a strength line format thus not fitting accurately in the data in the data set. Such fitting is known as under-fitting of data.

VARIANCE: -

The variability of model prediction for a given data point which tells us spread of our data is called the variance of the model. When a model is high variance, it is then said to as Over-fitting of data. Over-fitting is fitting

the training set accurately via complex curve and high order hypothesis. But is not solution as the error with unseen data is high.

Bias variance trade-off: -

If the algorithm is too simple then it may be o high bias and low variance condition.it may be on high variance and low bias. The condition the new entries will not perform well. There is something between both of these conditions known as trade-off or bias variance trade-off.

ANS.NO.15