Student Name: - Kiran Digambar Yadav.

Internship: - 35

Assignment No:- Worksheet_Set_5

Assignment No: - 5

Machine Learning

Q1 Answer: -

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 has no relation to express the effect of a bad or least significant independent variable on the regression. Thus, even if the model consists of a less significant variable say, for example, the person's Name for predicting the Salary, the value of R squared will increase suggesting that the model is better.

Whereas the RSS measures the amount of error remaining between the regression function and the data set after the model has been run. A smaller RSS figure represents a regression function that is well-fit to the data. The RSS, also known as the sum of squared residuals, essentially determines how well a regression model explains or represents the data in the model. any model might have variances between the predicted values and actual results. Although the variances might be explained by the regression analysis, the RSS represents the variances or errors that are not explained.

Since a sufficiently complex regression function can be made to closely fit virtually any data set, further study is necessary to determine whether the regression function is, in fact, useful in explaining the variance of the dataset. Typically, however, a smaller or lower value for the RSS is ideal in any model since it means there is less variation in the data set. In other words, the lower the sum of squared residuals, the better the regression model is at explaining the data. Thus, we can conclude RSS is the better measure of goodness of fit model in regression analysis.

Q2 Answer: -

The total sum of squares is a variation of the values of a dependent variable from the sample mean of the dependent variable. Essentially, the total sum of squares quantifies the total variation in a sample. It can be determined using the following formula,

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Where:

Yi – the value in a sample

$\bar{y}$ – the mean value of a sample

Regression sums of squares (also known as the sum of squares due to regression or explained sum of squares): - The regression sum of squares describes how well a regression model represents the modelled data. A higher regression sum of squares indicates that the model does not fit the data well.

The formula for calculating the regression sum of squares is:

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Where:

- $\hat{y}i$ – the value estimated by the regression line
- $\bar{y}$ – the mean value of a sample

Residual sum of squares (also known as the sum of squared errors of prediction): - The residual sum of squares essentially measures the variation of modelling errors. In other words, it depicts how the variation in the dependent variable in a regression model cannot be explained by the model. Generally, a lower residual sum of squares indicates that the regression model can better explain the data, while a higher residual sum of squares indicates that the model poorly explains the data.

The residual sum of squares can be found using the formula below:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Where:

- yi – the observed value
- $\hat{y}i$ – the value estimated by the regression line

The relationship between the three types of sums of squares can be summarized by the following equation:

$$TSS = SSR + SSE$$

Q3 Answer: -

One can easily overfit or underfit a machine learning model while it is being trained. To prevent this, we use regularization in machine learning to accurately fit a model onto our test set. Regularization describes methods for calibrating machine learning models to reduce the adjusted loss function and avoid overfitting or underfitting.

Often the linear regression model comprising of a large number of features suffers from some of the following:

- Overfitting: Overfitting which results in model failing to generalize on the unseen dataset
- Multicollinearity: Model suffering from multicollinearity effect
- Computationally Intensive: A model becomes computationally intensive.
- Model lack of generalization: Model found with higher accuracy fails to generalize on unseen or new data.
- Model instability: Different regression models can be created with different accuracies. It becomes difficult to select one of them.

The above problem makes it difficult to come up with a model which has higher accuracy on unseen data and which is stable enough. Variance, i.e., variability, is a characteristic of a standard least squares model. this model will not generalize well for a data set different than its training data. Regularization, significantly reduces the variance of the model, without a substantial increase in its bias. Therefore, the regularization techniques described above use the tuning parameter λ to control the effect of bias and variance. As the value of lambda increases, the value of the coefficients decreases, lowering the variance. Up to a point, this increase in λ is advantageous because it only reduces variance (avoiding overfitting) while maintaining all of the data's significant properties. But once the value reaches a certain point, the model begins to lose crucial characteristics, leading to bias and underfitting. As a result, care should be taken when choosing the value of & lambda.

Q4 Answer: -

Gini Index or Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen. Let us understand what is actually meant by 'impurity'?

If all the elements belong to a single class, then it can be called pure. The degree of Gini Index varies between 0 and 1, where, '0' denotes that all elements belong to a certain class or there exists only one class (pure), and '1' denotes that the elements are randomly distributed across various classes (impure). A Gini Index of '0.5 'denotes equally distributed elements into some classes.

Q5 Answer: -

Yes, unregularized decision-trees prone to overfitting. Unlike other regression models, decision tree does not use regularization to fight against overfitting. Instead, it employs tree pruning. Selecting the right hyperparameters (tree depth and leaf size) also requires experimentation, e.g., doing cross-validation with a hyperparameter matrix.

Q6 Answer: -

Ensemble learning is a general meta-approach to machine learning that seeks better predictive performance by combining the predictions from multiple models. Although there are a seemingly unlimited number of ensembles that you can develop for your predictive modelling problem, there are three methods that dominate the field of ensemble learning. The three main classes of ensemble learning methods are bagging, stacking, and boosting, and it is important to both have a detailed understanding of each method and to consider them on your predictive modelling project.

Let us understand in general types of ensemble techniques avail to us.

- Bagging involves fitting many decision trees on different samples of the same dataset and averaging the predictions.

- Stacking involves fitting many different model's types on the same data and using another model to learn how to best combine the predictions.

- Boosting involves adding ensemble members sequentially that correct the predictions made by prior models and outputs a weighted average of the predictions.

Q7 Answer: -

Sr. No	Bagging	Boosting
1	Bagging is a learning approach that aids in enhancing the performance, execution, and precision of machine learning algorithms.	Boosting is an approach that iteratively modifies the weight of observation based on the last classification.

2	In bagging, each model is assembled independently.	In boosting, the new models are impacted by the implementation of earlier built models.
3	It helps in solving the over-fitting issue.	It helps in reducing the bias.
4	In the case of bagging, if the classifier is unstable, then we apply bagging.	In the case of boosting, If the classifier is stable, then we apply boosting.
5	Here, every model has equal weight.	Here, the weight of the models depends on their performance.
6	It is the easiest method of merging predictions that belong to the same type.	It is a method of merging predictions that belong to different types.

Q8 Answer: -

OOB (out-of-bag) score is a performance metric for a machine learning model, specifically for ensemble models such as random forests. It is calculated using the samples that are not used in the training of the model, which is called out-of-bag samples. These samples are used to provide an unbiased estimate of the model's performance, which is known as the OOB score. The OOB error is computed using the samples that were not included in the training of the individual trees. This is different from the error computed using the usual training and validation sets, which are used to tune the hyperparameters of the random forest.

The OOB error can be useful for evaluating the performance of the random forest on unseen data. It is not always a reliable estimate of the generalization error of the model, but it can provide a useful indication of how well the model is performing.

Q9 Answer: -

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. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model. It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

Q10 Answer: -

Whenever a machine learning algorithm is implemented on a specific dataset, the performance is judged based on how well it generalizes, i.e., how it reacts to new, never-before-seen data. In case the performance of the learning algorithm is not satisfactory or there is room for improvement, certain parameters in the algorithm need to be changed/tuned/tweaked. These parameters are known as 'hyperparameters' and the process of varying these hyperparameters to better the learning algorithm's performance is known as 'hyperparameter tuning'. While this is an important step in modelling, it is by no means the only way to improve performance.

Q11 Answer: -

Large learning rates puts the model at risk of overshooting the minima so it will not be able to converge, what is known as exploding gradient. This has the effect of your model being unstable and unable to learn from your training data.

Q12 Answer: -

Non-linear problems cannot be solved with logistic regression because it has a linear decision surface. Linearly separable data is rarely found in real-world scenarios.

Q13 Answer: -

Sr.No.	Adaptive Boosting	Gradient Boosting
1	AdaBoost is the first designed boosting algorithm with a particular loss function	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.

2	it minimises the exponential loss function that can make the algorithm sensitive to the outliers	Gradient Boosting algorithm is more robust to outliers than AdaBoost
3	AdaBoost minimises loss function related to any classification error and is best used with weak learners.	Gradient Boosting is used to solve the differentiable loss function problem
4	The method was mainly designed for binary classification problems and can be utilised to boost the performance of decision trees	The technique can be used for both classification and regression problems.
5	In the case of AdaBoost, the shifting is done by up-weighting observations that were misclassified before	Gradient Boosting identifies the difficult observations by large residuals computed in the previous iterations.

Q14 Answer: -

The bias is known as the difference between the prediction of the values by the ML model and the correct value. Being high in biasing gives a large error in training as well as testing data. It is recommended that an algorithm should always be low biased to avoid the problem of underfitting. 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 on variance, it is then said to as Overfitting of Data. Overfitting is fitting the training set accurately via complex curve and high order hypothesis but is not the solution as the error with unseen data is high.

Bias variance trade-off is a relationship between the expected test error and the variance and the bias. Both contribute to the level of test error and ideally should be small as possible. Expected test error = variance + $bias^{2}$ + Irreducible error. But as the model complexity increases the bias decreases and the variance increases which lead to overfitting. And vice versa, model simplification helps to decrease the variance but it increases the bias which leads to underfitting.

Q15 Answer: -

- Linear kernels:

Linear kernels are the most basic type of kernel function and are used when the data is linearly separable. A linear kernel maps the data points from the original feature space to a higher-dimensional space using a linear function. The decision boundary in this case is a hyperplane, which is a subspace of one dimension less than the original space.

Example:

To illustrate how a linear kernel is used in an SVM, let's consider an example. Suppose we are given a dataset containing two classes of points, represented by red and blue dots in the figure below. We can use an SVM with a linear kernel to find a decision boundary that maximally separates the two classes.

In this example, the decision boundary is a straight line that maximally separates the two classes of points.

- Polynomial kernels:

Polynomial kernels are used when the data is not linearly separable and can be

separated by a polynomial function. A polynomial kernel maps the data points from the original feature space to a higher-dimensional space using a polynomial function of degree d. The decision boundary in this case is a polynomial curve of degree d-1.

Example:

To illustrate how a polynomial kernel is used in an SVM, let's consider an example. Suppose we are given a dataset containing two classes of points, represented by red and blue dots in the figure below. We can use an SVM with a polynomial kernel to find a decision boundary that maximally separates the two classes.

In this example, the decision boundary is a polynomial curve that maximally separates the two classes of points.

- Radial basis function (RBF) kernels:

Radial basis function (RBF) kernels are used when the data is not linearly separable and cannot be separated by a polynomial function. RBF kernels map the data points from the original feature space to a higher-dimensional space using a radial basis function, which is a function that is zero at the origin and increases with distance from the origin. The decision boundary in this case is a non-linear curve.

Example:

To illustrate how an RBF kernel is used in an SVM, let's consider an example. Suppose we are given a dataset containing two classes of points, represented by red and blue dots in the figure below. We can use an SVM with an RBF kernel to find a decision boundary that maximally separates the two classes.

In this example, the decision boundary is a non-linear curve that maximally separates the two classes of points.

ASSIGNMENT_NO_5

STATTISTICS_WORKSHEET_5

Question No	Answers
1	d) Expected
2	c) Frequencies.
3	c) 6
4	b) Chi squared distribution.
5	C) F Distribution.
6	b) Hypothesis.
7	a) Null Hypothesis.
8	a) Two tailed.
9	b) Research hypothesis.
10	a) np