MACHINE LEARNING ASSIGNMENT

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 ---

R-squared and Residual Sum of Squares (RSS) are both measures of the goodness of fit of a regression model, but they serve slightly different purposes.

R-squared (also known as the coefficient of determination) measures the proportion of variation in the dependent variable that is explained by the independent variables in the model. In other words, it indicates how well the model fits the data, with values ranging from 0 to 1. Higher R-squared values indicate a better fit of the regression model to the data. Therefore, R-squared is often used to compare different models and select the best one.

On the other hand, RSS measures the total sum of squared differences between the actual values of the dependent variable and the predicted values by the model. It represents the amount of unexplained variation in the data, and lower RSS values indicate a better fit, as they mean that the model is able to explain more of the variation in the data. It represents the sum of the squared differences between the actual and predicted values of the dependent variable. The goal is to minimize the residual sum of squares to obtain a better model fit.

In terms of determining the goodness of fit of a model, R-squared is generally considered a better measure than RSS. This is because R-squared provides an overall measure of the proportion of variance in the dependent variable that is explained by the model, whereas RSS only measures the magnitude of the residuals. Additionally, R-squared is a standardized measure and ranges from 0 to 1, making it easy to compare the fit of different models. In contrast, the magnitude of the RSS value depends on the scale of the dependent variable and can't be easily compared across models.

However, it's worth noting that neither R-squared nor RSS is a perfect measure of model fit. R-squared can be influenced by outliers or data points that don't fit the model well, while RSS doesn't take into account the number of variables or degrees of freedom in the model. Therefore, it's important to consider multiple metrics when evaluating a regression model's goodness of fit.

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.

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---

The total sum of squares (TSS) measures how much variation there is in the observed data, while the residual sum of squares measures the variation in the error between the observed data and modeled values. Regression sum 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 modeled data. A higher regression sum of squares indicates that the model does not fit the data well.

TSS(Total Sum of Squares)= ESS (Explained Sum of Squares) + RSS (Residual Sum of Squares)

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

Answer-----

Regularization refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting. Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.

While training a machine learning model, the model can easily be overfitted or under fitted. To avoid this, we use regularization in machine learning to properly fit a model onto our test set. Regularization techniques help reduce the chance of overfitting and help us get an optimal model.

Regularization is one of the most important concepts of machine learning. It is a technique to prevent the model from overfitting 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.

4. What is Gini–impurity index?

Answer-----

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. More precisely, the Gini Impurity of a dataset is a number between 0-0.5, which indicates the likelihood of new, random data being misclassified if it were given a random class label according to the class distribution in the dataset.

Gini Impurity tells us what is the probability of misclassifying an observation. Note that the lower the Gini the better the split. In other words the lower the likelihood of misclassification.

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

Answer----

So when it comes to decision trees the thing is, it makes very few assumptions about training data (linear model assumes that the data you will be feeding will be linear). If you don't constraint it, the tree will adapt itself to the training data, which will lead to overfitting.

Decision trees are a popular and powerful method for data mining, as they can handle both numerical and categorical data, and can easily interpret the results. However, decision trees can also suffer from overfitting, which means that they learn too much from the training data and fail to generalize well to new data

6. What is an ensemble technique in machine learning?

Answer----

In statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.

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

Answer----

Bagging helps to decrease the model's variance and Boosting helps to decrease the model's bias. Bagging is the simplest way of combining predictions that belong to the same type while Boosting is a way of combining predictions that belong to the different types. Bagging aims to decrease variance, not bias while Boosting aims to decrease bias, not variance.

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

Answer----

Out-of-bag (OOB) error, also called out-of-bag estimate, is a method of measuring the prediction error of random forests, boosted decision trees, and other machine learning models utilizing bootstrap aggregating (bagging). Bagging uses subsampling with replacement to create training samples for the model to learn from.

9. What is K-fold cross-validation?

Answer----

K-fold cross-validation is a technique for evaluating predictive models. The dataset 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?

Answer------

Hyperparameter tuning allows data scientists to tweak model performance for optimal results. This process is an essential part of machine learning, and choosing appropriate hyperparameter values is crucial for success. For example, assume you're using the learning rate of the model as a hyperparameter.

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

Answer-----

When the learning rate is too large, gradient descent can suffer from divergence. This means that weights increase exponentially, resulting in exploding gradients which can cause problems such as instabilities and overly high loss values.

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

Answer-----

Non-linear problems can't be solved with logistic regression because it has a linear decision surface.Logistic regression may not be accurate if the sample size is too small. If the sample size is on the small side, the model produced by logistic regression is based on a smaller number of actual observations. This can result in overfitting.Logistic Regression does not assume linear relationship between dependent and independent variables as it applies a non linear log transformation. But there could be some linear relationship among the independent variables i.e. multicollinearity which anyways should be avoided, in general, in any model.

13. Differentiate between Adaboost and Gradient Boosting.

Answer----

AdaBoost or Adaptive Boosting is the first Boosting ensemble model. The method automatically adjusts its parameters to the data based on the actual performance in the current iteration. Meaning, both the weights for re-weighting the data and the weights for the final aggregation are re-computed iteratively. In practice, this boosting technique is used with simple classification trees or stumps as base-learners, which resulted in improved performance compared to the classification by one tree or other single base-learner.

Gradient Boosting

Gradient Boost is a robust machine learning algorithm made up of Gradient descent and Boosting. The word 'gradient' implies that you can have two or more derivatives of the same function. Gradient Boosting has three main components: additive model, loss function and a weak learner. The technique yields a direct interpretation of boosting methods from the perspective of numerical optimisation in a function space and generalises them by allowing optimisation of an arbitrary loss function.

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

Answer-----

In machine learning, as you try to minimize one component of the error (e.g., bias), the other component (e.g., variance) tends to increase, and vice versa. Finding the right balance of bias and variance is key to creating an effective and accurate model. This is called the bias-variance tradeoff.

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

Answer---

Linear kernels are simpler and computationally efficient, suitable for linearly separable data. Non-linear kernels provide flexibility for capturing complex patterns in the data and are preferred when the relationships are not linear.

In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification.

In machine learning, the polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.