MACHINE LEARNING- 5 Answers

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 (R^2 or the coefficient of determination) 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. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).

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. Instead, it estimates the variance in the residuals, or error term.

Linear regression is a measurement that helps determine the strength of the relationship between a dependent variable and one or more other factors, known as independent or explanatory variables.

Typically, however, a smaller or lower value for the RSS is ideal in any model since it means there's 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.

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 = This gives you the distance from the linear line drawn to each particular variable. You could also describe TSS as the dispersion of observed variables around the mean, or the variance. So,

the goal of TSS is to measure the total variability of the dataset

ESS= Explained sum of square (ESS) or **Regression sum of squares** or Model sum of squares 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.

It tells how much of the variation between observed data and predicted data is being explained by the model proposed. Mathematically, it is the sum of the squares of the difference between the predicted data and mean data.

RSS = The residual sum of squares (RSS) measures the level of variance in the error term, or residuals, of a regression model. 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.

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.

Regularization is a set of techniques that can prevent overfitting in neural networks and thus improve the accuracy of a Deep Learning model when facing completely new data from the problem domain.

4) What is Gini–impurity index?

Answer= Gini Index, also known as Gini impurity, **calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly**. If all the elements are linked with a single class then it can be called pure

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

Answer= There are several approaches to avoiding overfitting in building decision trees. Pre-pruning that stop growing the tree earlier, before it perfectly classifies the training set. Post-pruning that allows the tree to perfectly classify the training set, and then post prune the tree.

6) What is an ensemble technique in machine learning?

Answer= Ensemble methods are **techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model**. The combined models increase the accuracy of the results significantly. This has boosted the popularity of ensemble methods in machine learning.

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

Answer= Bagging is a technique for reducing prediction variance by producing additional data for training from a dataset by combining repetitions with combinations to create multi-sets of the original data. Boosting is an iterative strategy for adjusting an observation's weight based on the previous classification.

Bagging and random forests are “bagging” algorithms that aim to reduce the complexity of models that overfit the training data. In contrast, boosting is an approach to increase the complexity of models that suffer from high bias, that is, models that underfit the training data.

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

Answer = The out-of-bag (OOB) error is **the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample**. every data point is passed for prediction to trees where it would be behaving as OOB and an aggregated prediction is recorded for each row. The OOB_score is computed as the number of correctly predicted rows from the out-of-bag sample. **OOB Error is the number of wrongly classifying the OOB Sample**.

9) What is K-fold cross-validation?

Answer= Cross validation is an evaluation method used in machine learning to find out how well your machine learning model can predict the outcome of unseen data. It is a method that is easy to comprehend, works well for a limited data sample and also offers an evaluation that is less biased, making it a popular choice.

The data sample is split into ‘k’ number of smaller samples, hence the name: K-fold Cross Validation. You may also hear terms like four fold cross validation, or ten fold cross validation, which essentially means that the sample data is being split into four or ten smaller samples respectively.

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

Answer = Hyperparameter tuning consists of **finding a set of optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any data set**. That combination of hyperparameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors.

Hyperparameter tuning takes advantage of the processing infrastructure of Google Cloud **to test different hyperparameter configurations when training your model**. It can give you optimized values for hyperparameters, which maximizes your model's predictive accuracy.

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

Answer = If the execution is not done properly while using gradient descent, it may lead to problems like vanishing gradient or exploding gradient problems. These problems occur when the gradient is too small or too large. And because of this problem the algorithms do not converge.

A learning rate that is too large can cause the model to converge too quickly to a suboptimal solution, whereas a learning rate that is too small can cause the process to get stuck.

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

Answer = 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 observations that belong to a class from all the other observations that do *not* belong to that class. The decision boundary is thus *linear*. Robust and efficient implementations are readily available (e.g. scikit-learn) to use logistic regression as a linear classifier.

While logistic regression makes core assumptions about the observations such as IID (each observation is independent of the others and they all have an identical probability distribution), the use of a linear decision boundary is *not* one of them. The linear decision boundary is used for reasons of simplicity following the Zen mantra – when in doubt simplify. In those cases where we suspect the decision boundary to be nonlinear, it may make sense to formulate logistic regression with a nonlinear model and evaluate how much better we can do.

13. Differentiate between Adaboost and Gradient Boosting.

Answer = AdaBoost can be used **to boost the performance of any machine learning algorithm**. It is best used with weak learners. These are models that achieve accuracy just above random chance on a classification problem. The most suited and therefore most common algorithm used with AdaBoost are decision trees with one level.

Gradient boosting is **a machine learning technique used in regression and classification tasks, among others**. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.

Gradient boosting re-defines boosting as a numerical optimisation problem where the objective is to minimise the loss function of the model by adding weak learners using gradient descent. Gradient descent is a first-order iterative optimisation algorithm for finding a local minimum of a differentiable function.

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

Answer = It is important to understand prediction errors (bias and variance) when it comes to accuracy in any machine learning algorithm.

There is a tradeoff 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.
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. Its recommended that an algorithm should always be low biased to avoid the problem of underfitting. By high bias, the data predicted is in a straight line format, thus not fitting accurately in the data in the data set. Such fitting is known as Underfitting of Data.

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

Answer=

Linear = Linear Kernel is **used when the data is Linearly separable, that is, it can be separated using a single Line**. It is one of the most common kernels to be used. It is mostly used when there are a Large number of Features in a particular Data Set.
The linear kernel is typically used on data sets with large amounts of features as **increasing the dimensionality on these data set does not necessarily improve separability**. Text classification is a typical example of this kind of data set.

RBF = RBF Kernel is **popular because of its similarity to K-Nearest Neighborhood Algorithm**. It has the advantages of K-NN and overcomes the space complexity problem as RBF Kernel Support Vector Machines just needs to store the support vectors during training and not the entire dataset. the RBF kernel **allows us to obtain exactly the same results as though we had added a landmark at each value of the original feature without the need to actually do it .**

Polynomial = 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**. The advantage of using the kernelized version is that **you can specify the degree to be large, thus increasing the chance that data will become linearly separable in this high-dimensional space, without slowing the model down**.