MACHINE LEARNING

Q1 to Q15 are subjective answer type questions, Answer them briefly.

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?
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.
3. What is the need of regularization in machine learning?
4. What is Gini–impurity index?
5. Are unregularized decision-trees prone to overfitting? If yes, why?
6. What is an ensemble technique in machine learning?
7. What is the difference between Bagging and Boosting techniques?
8. What is out-of-bag error in random forests?
9. What is K-fold cross-validation?
10. What is hyper parameter tuning in machine learning and why it is done?
11. What issues can occur if we have a large learning rate in Gradient Descent?
12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?
13. Differentiate between Adaboost and Gradient Boosting.
14. What is bias-variance trade off in machine learning?
15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.

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?

Ans: The answer is R-squared.

Explanation: R-squared test is the standard goodness-of-fit measures for linear models. It

Incorporates the Residual Sum of Squares (RSS), but is preferred over RSS because it is

To interpret and avoids some of RSS's limitation. A smaller RSS indicates, a better fit, while

A value of zero means a perfect fit.

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.

Ans: TSS (Total Sum of Squares)- tells you how much variation there is in the dependent variable. The formula is

$$\textbf{Total SS} = \Sigma(Yi - \text{mean of } Y)^2 \quad \textbf{or} \quad \textbf{TSS} = \Sigma(Y_i - \bar{Y})^2,$$

Where,

Σ = summation
Yi= the actual value of the dependent variable for observation i
$\bar{Y}$ = the predicted value of the dependent variable for observation i

ESS(Explained Sum of Squares)- tells us how much of the variation in the dependent variables our our model is

The formula is

$$\textbf{ESS} = \Sigma(Yi - \text{mean of } Y)^2$$

Residual Sum of Squares- helps determine the suitability of a model for our data by measuring the overall difference between our observed data and the values predicted by the estimation model. More specifically, the RSS provides insight into the amount of unexplained variation in the dependent Variable.

The formula is

$$RSS = \Sigma e^2$$

where,

E= the difference between the actual and predicted Y value.
Σ= the sum of all values

Relation b/w TSS, ESS and RSS

$$TSS = ESS + RSS$$

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

Ans: Regularization is a set of methods used to reduce overfitting in machine learning models. The overall idea of regularization is to help models determine the key features of the data set without fixating on noise or irrelevant detail.

Regularization methods typically focus more on generalizability outside of training data sets than the accuracy of the model. The result of regularization is a more balanced model. It may not perform on the training data, as it's not overly complex, but it will likely do better on new, unseen data, which is the ultimate goal of a practical machine learning model.

4. What is Gini–impurity index?

Ans: Gini Impurity is a method that measures the impurity of a dataset. The more impure the dataset, the higher is the Gini index.

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

Ans. Yes, unregularized decision trees are prone to overfitting. Decision trees tend to overfit when the available training data is limited, as they attempt to extract patterns even from noise. Decision trees can grow to a considerable depth, resulting compatible decision boundaries. As the tree become deeper, it becomes more susceptible to overfitting.

6. What is an ensemble technique in machine learning?

Ans. 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?

Ans. Bagging and Boosting are ensembles learning techniques. Bagging attempts to tackle the over-fitting. Boosting tries to reduce bias. If the classifier is unstable (high variance), then we need to apply bagging. If the classifier is steady and straightforward (high bias), then we need to apply boosting.

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

Ans. Out-of-Bag (OOB) is a method of measuring the prediction error of random forests, bagged decision trees and other machine learning models utilizing bootstrap bagging.

9. What is K-fold cross-validation?

Ans. In K-fold cross-validation, the data set is divided into a number of K-folds and used to assess the model's ability as new data become available. K represents the number of groups into which the data sample is divided. For example, if you find the k value to be 5, you can call it 5-fold cross-validation.

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

Ans. Hyperparameters directly control model structure, function, and performance. 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.

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

Ans. Too small learning rate will lead to very slow learning or even inability to learn at all, while too large learning rate can lead to exploding or oscillating performance over the training epochs and to a lower final performance.

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

Ans. No,
The reason is Logistic regression is simple and easy to implement, but it also as some drawbacks. One of them is that it assumes a linear relationship between the input feature and the output. This means, it cannot capture the complexity and non-linearity of data.

]

13. Differentiate between AdaBoost and Gradient Boosting.

Ans. AdaBoost minimizes loss function related to any classification error and is best used with weak learners. The method was mainly designed for binary classification problems and can be utilized to boost the performance of decision trees.

Gradient Boosting is more robust to outliers and noise since it equally considers all training instances when optimizing the loss function. It is used to solve the differentiable loss function problem.

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

Ans. The bias- variance trade off is about finding the right balance between simplicity and complexity in a machine Learning model.
High bias means the model is too simple and consistently misses the target, while high variance means the model is too complex and shoots all over the place.

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

Ans. Linear Kernel- The linear kernel is the simplest and used when the data is linearly separable. It calculates the dot product between the feature vectors.

Polynomial Kernel- The Polynomial kernel is effective for non linear data. It computes the similarity between Two vectors in terms of the polynomial of the original variables.

RBF Kernel -The RBF kernel is a common type of kernel is SVM for handling non-linear decision boundaries. It maps the data into an infinite- dimensional space.