MACHINE LEARNING ASSIGNMENT - 5

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?**

The R-squared is the better measure of goodness of fit model in regression .R-squared is a statistical measure that represents the goodness of fit of a regression model. The ideal value for r-square is 1. The closer the value of r-square to 1, the better is the model fitted. R-square is a comparison of the residual sum of squares (SS_{res}) with the total sum of squares(SS_{tot}). The total sum of squares is calculated by summation of squares of perpendicular distance between data points and the average line.

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.**

Total Sum of Squares: TSS explains the variation between observations or dependent variable's values and its mean. Regression Sum of Squares: It explains how well a regression model represents the data. A higher value indicates that the model does not fit the data well and vice versa.
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.
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.

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

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.

4. **What is Gini–impurity index?**

Gini-impurity or gini-index in machine learning is a metric to measure the randomness in a feature. It determines whether a particular feature adds value towards the predictability of the model or not. Higher the value of gini-impurity, lower is the predictive power of the variable or higher is the randomness.

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

In the case of decision tree's they can learn a training set to a point of high granularity that makes them easily overfit. Allowing a decision tree to split to a granular degree, is

the behavior of this model that makes it prone to learning every point extremely well — to the point of perfect classification — ie: overfitting.

6. What is an ensemble technique in machine learning?

Ensemble learning is a machine learning technique that creates a model made up of multiple learning algorithms. The model is created by combining the outputs of these algorithms and averaging them to get the final prediction. With ensemble learning, a single machine can be trained on many models, each with different strengths and weaknesses.

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

Bagging gives equal weight to each model, whereas in Boosting technique, the new models are weighted based on their results. In boosting, new subsets of data used for training contain observations that the previous model misclassified. Bagging uses randomly generated training data subsets.

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

Out-of-Bag Error in Random Forest The out-of-bag error is the average error for each predicted outcome calculated using predictions from the trees that do not contain that data point in their respective bootstrap sample. This way, the Random Forest model is constantly being validated while being trained.

9. What is K-fold cross-validation?

K-Fold Cross Validation. K-Fold Cross Validation is a common type of cross validation that is widely used in machine learning. K-fold cross validation is performed as per the following steps: Partition the original training data set into k equal subsets. Each subset is called a fold. Let the folds be named as f 1, f 2, …, f k .

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

Hyperparameter tuning in machine learning is almost identical to tuning a guitar. It is a procedure where we change the values of some of the important parameters in the machine learning model so that it would have an impact on the overall performance of the algorithm respectively.

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

The learning rate can seen as step size, η. As such, gradient descent is taking successive steps in the direction of the minimum. If the step size η is too large, it can (plausibly)

"jump over" the minima we are trying to reach, ie. we overshoot. This can lead to osculations around the minimum or in some cases to outright divergence.

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

Yes, it might work, but logistic regression is more suitable for classification task and we want to prove that logistic regression yields better results than linear regression. Let's see how logistic regression classifies our dataset. Now we have 2 models trained on the same dataset, one by linear regression, and another by logistic regression.

13. Differentiate between Adaboost and Gradient Boosting.

AdaBoost is the first designed boosting algorithm with a particular loss function. On the other hand, 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.

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

Bias Variance Tradeoff is a design consideration when training the machine learning model. Certain algorithms inherently have a high bias and low variance and vice-versa. In this one, the concept of bias-variance tradeoff is clearly explained so you make an informed decision when training your ML models.

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

Linear Kernel

It is the most basic type of kernel, usually one dimensional in nature. It proves to be the best function when there are lots of features. The linear kernel is mostly preferred for text-classification problems as most of these kinds of classification problems can be linearly separated.Linear kernel functions are faster than other functions.

Polynomial Kernel

It is a more generalized representation of the linear kernel. It is not as preferred as other kernel functions as it is less efficient and accurate.

Gaussian Radial Basis Function (RBF)

It is one of the most preferred and used kernel functions in svm. It is usually chosen for non-linear data. It helps to make proper separation when there is no prior knowledge of data.