MACHINE LEARNING

WORKSHEET SET 5

Ques- 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 - **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.**

Ques- 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 – 2 **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 modelled values.**

Ques- 3- What is the need of regularization in machine learning?

Answer – 3 - **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.**

Ques- 4- What is Gini–impurity index?

Answer – 4 – **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.**

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

Answer – 5– **Decision trees are prone to overfitting, especially when a tree is particularly deep. This is due to the amount of specificity we look at leading to**

smaller sample of events that meet the previous assumptions. This small sample could lead to unsound conclusions.

Ques- 6- What is an ensemble technique in machine learning?

Answer – 6 - Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. This has been the case in a number of machine learning competitions, where the winning solutions used ensemble methods.

Ques- 7- What is the difference between Bagging and Boosting techniques?

Answer – 7 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.

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

Answer – 8 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. This allows the RandomForestClassifier to be fit and validated whilst being trained

Ques- 9- What is K-fold cross-validation?

Answer – 9 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.

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

Answer – 10- Hyperparameters in Machine learning are those parameters that are explicitly defined by the user to control the learning process. These

hyperparameters are used to improve the learning of the model, and their values are set before starting the learning process of the model.

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

Answer – 11- In order for Gradient Descent to work, we must set the learning rate to an appropriate value. This parameter determines how fast or slow we will move towards the optimal weights. If the learning rate is very large we will skip the optimal solution.

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

Answer – 12- It can only be used to predict discrete functions. Hence, the dependent variable of Logistic Regression is bound to the discrete number set. It is very fast at classifying unknown records. Non-linear problems can't be solved with logistic regression because it has a linear decision surface.

Ques- 13 Differentiate between Adaboost and Gradient Boosting?

Answer – 13- 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.

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

Answer – 14 In statistics and machine learning, the bias–variance tradeoff is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters.

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

Answer – 15 The linear, polynomial and RBF or Gaussian kernel are simply different in case of making the hyperplane decision boundary between the classes. The kernel functions are used to map the original dataset

(linear/nonlinear) into a higher dimensional space with view to making it linear dataset.