MACHINE LEARNING
ASSIGNMENT NO.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?
Answer : R-squared is the value which will define the best measure of goodness of fit model in regression. R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively.
Residuals means observed value – Predicted value

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 = ESS + RSS, where TSS is Total Sum of Squares, ESS is Explained Sum of Squares and RSS is Residual Sum of Suqares. The aim of Regression Analysis is explain the variation of dependent variable Y. The sum of squares is used to calculate whether a linear relationship exists between two variables, and any unexplained variability is referred to as the Residual Sum of Square.The RSS allows you to determine the amount of error left between a regression function and the data set after the model has been run. You can interpret a smaller RSS figure as a regression function that is well-fit to the data while the opposite is true of a larger RSS figure.

Here is the formula for calculating the residual sum of squares:

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

Answer : Regularization is a technique to prevent the model from overfitting by adding to extra information to it. This technique can be used in such a way that it will allow to maintain all variables or features in the model by reducing the magnitude of the variables. Hence, it maintains accuracy as well as a generalization of the model.

4. What is Gini–impurity index?
Answer : The Gini Index is a proportion of impurity or inequality in statistical and monetary settings. In machine learning, it is utilized as an impurity measure in decision tree algorithms for classification tasks. The Gini Index measures the probability of a haphazardly picked test being misclassified by a decision tree algorithm, and its value goes from 0 (perfectly pure) to 1 (perfectly impure).

5. Are unregularized decision-trees prone to overfitting? If yes, why?
Answer : Yes unregularized decision-trees prone to overfitting.
Overfitting happens when any learning processing overly optimizes training set error at the cost test error. While it's possible for training and testing to perform equality well in cross validation, it could be as the result of the data being very close in characteristics, which may not be a huge problem. 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 behaviour of this model that makes it prone to learning every point extremely well to the point of perfect classification and i.e. overfitting.

6. What is an ensemble technique in machine learning?
Answer : Ensemble technique are techniques that create multiple models and then combine them to produce improved results. Ensemble methods in machine learning usually produce more accurate solutions than a single model.
Bagging
Boosting
Stacking

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

Bagging	Boosting
1.Independent trees/modals are independent to each other.	Independent trees/modals are not independent to each other.
2.Aim to decrease variance not bias	Aim to decrease bias not variance
3.Bagging tries to solve overfittig problem	Boosting tries to reduce bias.
4. Base learner training parallel.	Base learner training sequential.

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. This allows the random forest classifier to be fit and validated whilst being trained.

9. What is K-fold cross-validation?
Answer : K-fold cross-validation is a technique for evaluating predictive models.It is a resampling procedure used to evaluate machine learning models on a limited data sample. If you have a machine learning model and some data, you want to tell if your model can fit. You can split your data into training and test set.

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.

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 : No we can’t use because the generation of this dataset is pretty simple, there were plenty of data points, and no noise at all, the logistic regression model performed poorly on this dataset. The reason is that the target label has no linear correlation with the features.

13. Differentiate between Adaboost and Gradient Boosting.

Adaboost is computed with a specific loss function and becomes more rigid when comes to few iterations. But in gradient boosting, it assists in finding the proper solution to additional iteration modeling problem as it is built with some generic features.

14. What is bias-variance trade off in machine learning?
Answer : It struggles to grasp the intricacies of the data and thus fails to provide an accurate prediction. Striking a balance between accuracy and the ability to make predictions beyond the training data in an ML model is called the bias-variance tradeoff.

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

Answer : In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. RBF Kernel is popular because of its similarity to K-Nearest Neighborhood Algorithm.
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.