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?

   ANS: The residual sum of squares (RSS) is the absolute amount of explained variation, whereas R-squared is the absolute amount of variation as a proportion of total variation. RSS is better than r-squared as 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.

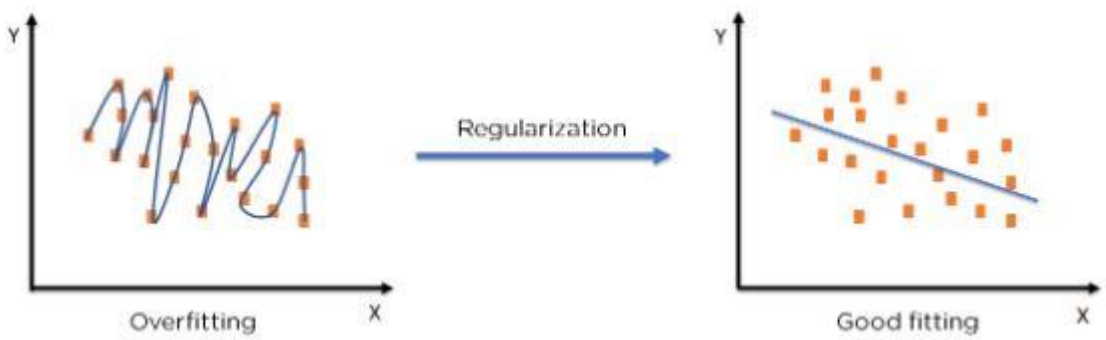
2. What are TSS (Total Sum of Squares), ESS (Explained Sum of Squares) and RSS (Residual Sum of Squares) in regression. Also mention the equation relating these three metrics with each other.

   ANS: explained sum of squares (ESS), alternatively known as the model sum of squares or sum of squares due to regression (SSR**) .** residual sum of squares (RSS) or sum of squares of errors), is a quantity used in describing how well a model, often a regression model, represents the data being modelled. In particular, the explained sum of squares measures how much variation there is in the modelled values and this is compared to the total sum of squares (TSS), which measures how much variation there is in the observed data, and to the residual sum of squares, which measures the variation in the error between the observed data and modelled values.

   TSS=ESS+RSS

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

   ANS: 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.



4. What is Gini–impurity index?

   ANS: A measurement used to build Decision Trees to determine how the features of a dataset should split nodes to form the tree. More precisely, the Gini Impurity of a dataset is a number between 0-0.5, which indicates the likelihood of new, random data being misclassified if it were given a random class label according to the class distribution in the dataset.

   Consider a dataset D that contains samples from k classes. The probability of samples belonging to class i at a given node can be denoted as pi. Then the Gini Impurity of D is defined as:

   $$Gini(D)=1-\sum_{i=1}^{k} p_i^2$$

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

   ANS: Regularization basically adds the penalty as model complexity increases. Regularization parameter (lambda) penalizes all the parameters except intercept so that model generalizes the data and won't overfit. Model overfitting is a serious problem and can cause the model to produce misleading information. One of the techniques to overcome overfitting is Regularization. Regularization, in general, penalizes the coefficients that cause the overfitting of the model.

6. What is an ensemble technique in machine learning?

   ANS: Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model. Types of ensemble techniques are:

   1- BAGGing, or Bootstrap AGGregating. BAGGing gets its name because it combines Bootstrapping and Aggregation to form one ensemble model.

   2- Random Forest Models. Random Forest Models can be thought of as BAGGing, with a slight tweak. When deciding where to split and how to make decisions, BAGGed Decision Trees have the full disposal of features to choose from.

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

   ANS: Bagging and Boosting are two types of Ensemble Learning. These two decrease the variance of a single estimate as they combine several estimates from different models. So the result may be a model with higher stability.

   Bagging: It is a homogeneous weak learners' model that learns from each other independently in parallel and combines them for determining the model average.

   Boosting: It is also a homogeneous weak learners' model but works differently from Bagging. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm.

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

   ANS: 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?

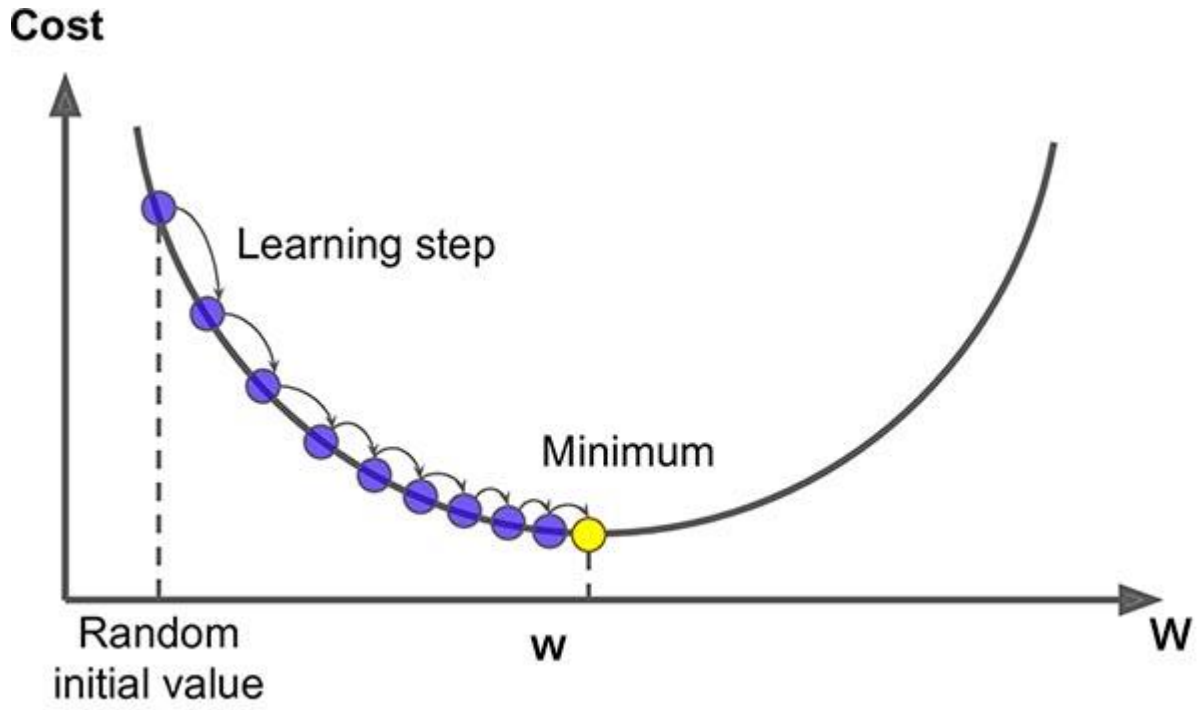
   ANS: Cross-validation is a statistical method used to estimate the skill of machine learning models**.** It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods. 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. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

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

    ANS: Hyper parameter tuning is choosing a set of optimal hyper parameters for a learning algorithm. A hyper parameter is a model argument whose value is set before the learning process begins. The key to machine learning algorithms is hyper parameter tuning. It is done as Model parameters are learned from data and hyper-parameters are tuned to get the best fit. Searching for the best hyper-parameter can be tedious, hence search algorithms like grid search and random search are used.

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

    ANS: Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function.



    If learning rate is too small, gradient descent can be slow. If learning rate is too large, gradient descent can overshoot the minimum. It may fail to converge and even diverge.

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

    ANS: Logistic Regression has traditionally been used as a linear classifier, i.e. when the classes can be separated in the feature space by linear boundaries. That can be remedied however if we happen to have a better idea as to the shape of the decision boundary.

13. Differentiate between Adaboost and Gradient Boosting.

    ANS: 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.
    AdaBoost minimises 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 utilised to boost the performance of decision trees. Gradient Boosting is used to solve the differentiable loss function problem. The technique can be used for both classification and regression problems**.**

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

    ANS: 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. This happens when the hypothesis is too simple or linear in nature.

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

    ANS: Linear: Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

RBF : Radial Basis Function (RBF), RBF kernels are the most generalized form of kernelization and is one of the most widely used kernels due to its similarity to the Gaussian distribution.

Polynomial kernels : represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.