ASSIGNMENT – 5 MACHINE LEARNING

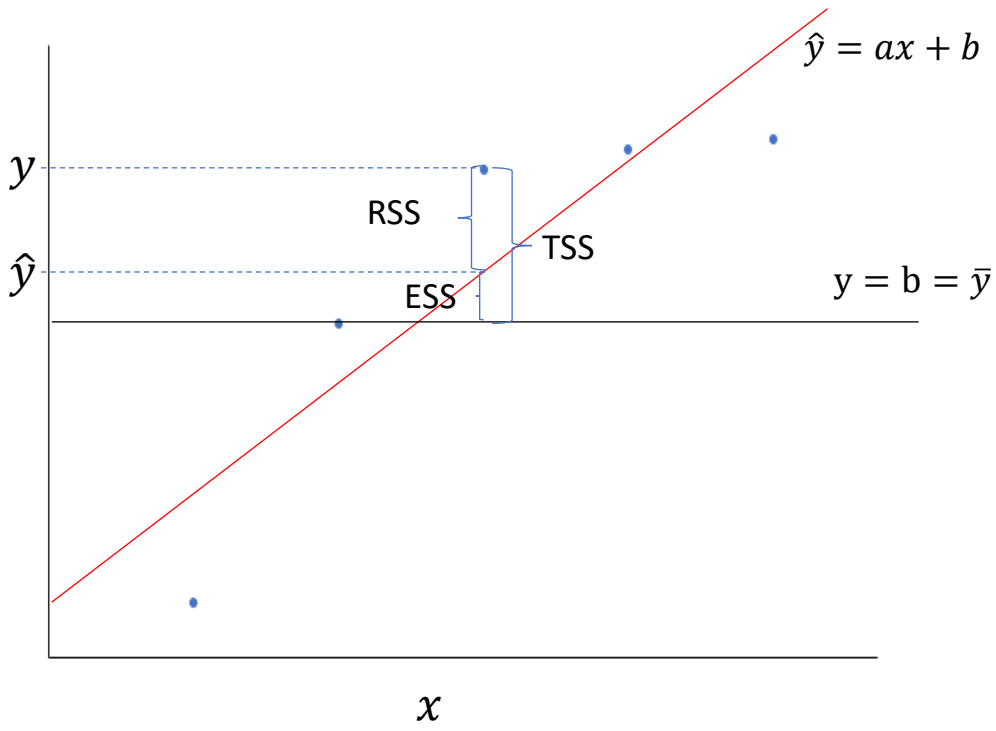
Q1 to Q15 are subjective answer type questions, Answer them briefly.

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?

Ans: Both are good measures of goodness of fit of a regression model. Residual Sum of Squares (RSS) give the goodness of fit as a sum of squares of residual, whereas R-squared measure the goodness of fit in terms of explained sum of square as a percentage of total sum of square that the model can explain. Among different models, the one with lower RSS will be a better fit model, but in case of R-squared, a model with higher value which is as close to 1 as possible is preferable. Both are evaluation metrics for a regression model.

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: Consider a simple linear regression model as shown in the figure below:



- RSS = Residual sum of squares, $\text{RSS} = \sum(y - \hat{y})^2$
- ESS = Explained sum of squares, $\text{ESS} = \sum(\bar{y} - \hat{y})^2$
- TSS = Total sum of squares, $\text{TSS} = \sum(y - \bar{y})^2$

For, simple linear regression, TSS = RSS + ESS

Goodness of fit, $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$

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

Ans: Regularization is an approach to overcome overfitting. In this approach we penalize higher order terms to make the model best fit.

4. What is Gini–impurity index?

Ans: Gini–impurity index tells that how well the classes are classified in a Decision tree. Higher Gini impurity refers a poor classification. It is given by:

Gini impurity index = weighted sum of Gini impurity of individual leaves in a Decision tree

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

Ans: Unregularized decision tree where there is no limit to the depth of the tree from root node to leaf node tend to classify even the noise and outliers. This way the model tends to overfit.

6. What is an ensemble technique in machine learning?

Ans: Ensemble technique trains a number of machine learning models in order to find the best model in terms of accuracy and other performance metrics.

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

Ans: Bagging also known as bootstrap aggregation is an ensemble technique where various machine learning models are trained in series to find the model with best performance.

Boosting is also an ensemble technique where the various machine learning models are trained in parallel to find the best performing model.

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

Ans: Out of bag error in random forest is a method of measuring prediction error using bootstrap aggregating or bagging. The data used for measuring error is called out of bag data. This is the data which was not used during the ensemble training.

9. What is K-fold cross-validation?

Ans: K-fold cross validation is a method to train the model in K different splits of the train and test data. This helps to train the model in a more exhaustive way and hence thereby creating a more accurate model.

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

Ans: In machine learning, certain parameters are specified before the start of the training process. These parameters are called hyperparameters. They control the learning process.

Examples: In KNN algorithm, the value of K has to be specified. Here K is called hyperparameter.

Similarly, learning rate has to be specified to train the neural network. Here, learning rate is called hyperparameter.

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

Ans: Gradient descent is used to minimize the cost function. If there is large learning rate then there is chance that the minimum gets missed.

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

Ans: Non-linear problems can't be solved with logistic regression because it has a linear decision surface.

13. Differentiate between Adaboost and Gradient Boosting.

Ans: Adaboost is an ensemble modelling technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. These weak models are trees with just one node and two leaves. The wrongly classified samples obtained from a weak classifier are given higher weight and passed on to the next classifier. This way this series keeps going on until all the correct classification is obtained.

Gradient boosting is also an ensemble approach to train a model for machine learning. It attempts to build a model using a decision tree with a maximum number of leaves between 8 and 32. It also takes the prediction of each tree and adds it to the weighted prediction of the subsequent tree until there is not much difference obtained by adding any additional tree.

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

Ans: Bias is the error of a model in the training data, whereas variance is the error of the model in test data. When we try to reduce the bias too much then the model tends to overfit, hence its variance increases, that is, it performs poorly on test data. Therefore, it is necessary to trade off properly between bias and variance so that the model performs well both on train and test data.

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

Ans: A kernel is a mathematical function that take data as input and transform it into a higher dimension to help find decision boundary for the classes which are not linearly separable. The different of kernel functions are:

(i) Linear kernel: It is given by:

$$k(x, y) = a^T b$$

where, a(x,y) and b(x,y) are two sample points.

(ii) RBF kernel: The radial basis function or RBF or Gaussian kernel is given by:

$$k(x, y) = e^{-\gamma(a-b)^2}$$

where,

a(x,y), b(x,y) = sample points
γ = determined using cross validation. It acts as a weight to the squared distance between data points x and y.

RBF finds the SVM classifier in infinite dimension. Therefore, this kernel is very useful and widely used.

(iii) Polynomial kernel: It is given by:

$$k(x, y) = (a^T b + r)^d$$

where,
a(x,y), b(x,y) = two different observations in the data set
r = coefficient of the polynomial
d = degree of the polynomial

This kernel transforms a linear dataset into a polynomial, which can be easily separated by a decision boundary.