MACHINE LEARNING

**1 In Q1 to Q7, only one option is correct, Choose the correct option:**

1. The value of correlation coefficient will always be:

A) between 0 and 1

B) greater than -1

C) between -1 and 1

D) between 0 and -1

**ANS: C) between -1 and 1**

2. Which of the following cannot be used for dimensionality reduction?

A) Lasso Regularisation

B) PCA

C) Recursive feature elimination

D) Ridge Regularisation

ANS:

3. Which of the following is not a kernel in Support Vector Machines?

A) linear B) Radial Basis Function C) hyperplane D) polynomial

**ANS: C) hyperplane**

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?

A) Logistic Regression B) Naïve Bayes Classifier C) Decision Tree Classifier D) Support Vector Classifier

**ANS: B) Naïve Bayes Classifier**

5. In a Linear Regression problem, ‘X’ is independent variable and ‘Y’ is dependent variable, where ‘X’ represents weight in pounds. If you convert the unit of ‘X’ to kilograms, then new coefficient of ‘X’ will be? (1 kilogram = 2.205 pounds)

A) 2.205 × old coefficient of ‘X’

B) same as old coefficient of ‘X’

C) old coefficient of ‘X’ ÷ 2.205

D) Cannot be determined

**ANS:**

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?

A) remains same B) increases C) decreases D) none of the above

**ANS: B) increases**

7. Which of the following is not an advantage of using random forest instead of decision trees?

A) Random Forests reduce overfitting

B) Random Forests explains more variance in data then decision trees

C) Random Forests are easy to interpret

D) Random Forests provide a reliable feature importance estimate

**ANS: B) Random Forests explains more variance in data then decision trees**

**In Q8 to Q10, more than one options are correct, Choose all the correct options:**

8. Which of the following are correct about Principal Components?

A) Principal Components are calculated using supervised learning techniques

B) Principal Components are calculated using unsupervised learning techniques

C) Principal Components are linear combinations of Linear Variables.

D) All of the above

**ANS: D) All of the above**

9. Which of the following are applications of clustering?

A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index

B) Identifying loan defaulters in a bank on the basis of previous years’ data of loan accounts.

C) Identifying spam or ham emails

D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

**ANS: C) Identifying spam or ham emails**

10. Which of the following is(are) hyper parameters of a decision tree?

A) max\_depth B) max\_features C) n\_estimators D) min\_samples\_leaf

**ANS: A) max\_depth B) max\_features D) min\_samples\_leaf**

**Q11 to Q15 are subjective answer type questions, Answer them briefly.**

**11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.**

**ANS:** An outlier is a value or point that [differs substantially from the rest of the data](https://en.wikipedia.org/wiki/Outlier).

It can hold the useful information about your data.

Some outliers represent true values from natural variation in the population. Other outliers may result from incorrect data entry, equipment malfunctions, or other [measurement errors](https://www.scribbr.com/methodology/random-vs-systematic-error/).

There are four ways to calculate the outliers:

* Sorting Method
* Using Visualization
* Statistical outlier detection
* Using the Inter quartile Range

Inter Quartile Range (IQR):

The**IQR**  of identifying outliers to set up a “fence” outside of Q1 and Q3. Any values that fall outside of this fence are considered outliers. To build this fence we take 1.5 times the IQR and then subtract this value from Q1 and add this value to Q3. This gives us the minimum and maximum fence posts that we compare each observation to. Any observations that are more than 1.5 IQR below Q1 or more than 1.5 IQR above Q3 are considered outliers. This is the method that Minitab uses to identify outliers by default.

**12. What is the primary difference between bagging and boosting algorithms?**

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

Bagging is an acronym for ‘Bootstrap Aggregation’ and is used to decrease the variance in the prediction model. Bagging is a parallel method that fits different, considered learners independently from each other, making it possible to train them simultaneously.

Bagging generates additional data for training from the dataset. This is achieved by random sampling with replacement from the original dataset. Sampling with replacement may repeat some observations in each new training data set. Every element in Bagging is equally probable for appearing in a new dataset.

**Boosting**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.

Boosting is a sequential ensemble method that iteratively adjusts the weight of observation as per the last classification. If an observation is incorrectly classified, it increases the weight of that observation. The term ‘Boosting’ in a layman language, refers to algorithms that convert a weak learner to a stronger one. It decreases the bias error and builds strong predictive models.

Data points mis predicted in each iteration are spotted, and their weights are increased. The Boosting algorithm allocates weights to each resulting model during training. A learner with good training data prediction results will be assigned a higher weight. When evaluating a new learner, Boosting keeps track of learner’s errors.

**13. What is adjusted R2 in linear regression. How is it calculated?**

**ANS:** Adjusted R Squared or Modified R^2 determines the extent of the variance of the dependent variable, which the independent variable can explain. The specialty of the modified R^2 is that it does not consider the impact of all independent variables but only those which impact the variation of the dependent variable. Therefore, the value of the modified R^2 can also be negative, though it is not always negative.

**Adjusted R2 = 1 – [(1-R2)\*(n-1)/(n-k-1)]**

where:

* **R2**: The R2 of the model
* **n**: The number of observations
* **k**: The number of predictor variables

**14. What is the difference between standardisation and normalisation?**

**ANS: Nirmalization:**

Normalization is a feature scaling technique to bring the features in the data to a common range say [0, 1] or [-1, 0] or [-1, 1]. In this section, we’ll go through 3 popular normalization methods are below.

### MinMaxScaler

### MaxAbsScaler

### RobustScaler

**Standardization:**

Standardization is the most commonly used feature scaling technique in machine learning. This is because some of the algorithms assume the normal or near-normal distribution of the data. If the features are normally distributed then the model behaves badly. The StandardScaler and standardization both refer to the same thing.

**15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation**

**ANS:** Cross Validation in Machine Learning is a great technique to deal with overfitting problem in various algorithms. Instead of training our model on one training dataset, we train our model on many datasets. Below are some of the advantages and disadvantages of Cross Validation in Machine Learning:  
  
**Advantages of Cross Validation**  
  
**1. Reduces Overfitting:** In Cross Validation, we split the dataset into multiple folds and train the algorithm on different folds. This prevents our model from overfitting the training dataset. So, in this way, the model attains the generalization capabilities which is a good sign of a robust algorithm.

**2. Hyperparameter Tuning:** Cross Validation helps in finding the optimal value of hyperparameters to increase the efficiency of the algorithm.

**Disadvantages of Cross Validation**  
  
**1. Increases Training Time:** Cross Validation drastically increases the training time. Earlier you had to train your model only on one training set, but with Cross Validation you have to train your model on multiple training sets.   
  
For example, if you go with 5 Fold Cross Validation, you need to do 5 rounds of training each on different 4/5 of available data. And this is for only one choice of hyperparameters. If you have multiple choice of parameters, then the training period will shoot too high.  
  
**2. Needs Expensive Computation:** Cross Validation is computationally very expensive in terms of processing power required.