MACHINE LEARNING

**WORKSHEET – 1**

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

1. The value of correlation coefficient will always be:
   1. between 0 and 1 B) greater than -1

C) between -1 and 1 D) between 0 and -1

Ans: C) between -1 and 1

-----------------------------------------------------------------------------------------------------------------------------------------------

1. Which of the following cannot be used for dimensionality reduction?
   1. Lasso Regularisation B) PCA

C) Recursive feature elimination D) Ridge Regularisation

Ans: C) Recursive feature elimination

-----------------------------------------------------------------------------------------------------------------------------------------------

1. Which of the following is not a kernel in Support Vector Machines?
   1. linear B) Radial Basis Function

C) hyperplane D) polynomial

Ans: C) hyperplane

-----------------------------------------------------------------------------------------------------------------------------------------------

1. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?
   1. Logistic Regression B) Naïve Bayes Classifier

C) Decision Tree Classifier D) Support Vector Classifier

Ans: A) Logistic Regression

-----------------------------------------------------------------------------------------------------------------------------------------------

1. 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)

* 1. 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: A) 2.205 × old coefficient of ‘X’

-----------------------------------------------------------------------------------------------------------------------------------------------

1. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?
   1. remains same B) increases

C) decreases D) none of the above

Ans: C) decreases

-----------------------------------------------------------------------------------------------------------------------------------------------

1. Which of the following is not an advantage of using random forest instead of decision trees?
   1. Random Forests reduce overfitting
   2. Random Forests explains more variance in data then decision trees
   3. Random Forests are easy to interpret
   4. Random Forests provide a reliable feature importance estimate

Ans: C) Random Forest are easy to interpret

-----------------------------------------------------------------------------------------------------------------------------------------------

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

1. Which of the following are correct about Principal Components?
   1. Principal Components are calculated using supervised learning techniques
   2. Principal Components are calculated using unsupervised learning techniques
   3. Principal Components are linear combinations of Linear Variables.
   4. All of the above

Ans: D) All of the above

-----------------------------------------------------------------------------------------------------------------------------------------------

1. Which of the following are applications of clustering?
   1. Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index
   2. Identifying loan defaulters in a bank on the basis of previous years’ data of loan accounts.
   3. Identifying spam or ham emails
   4. Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

Ans: A), B), C), D)

-----------------------------------------------------------------------------------------------------------------------------------------------

1. Which of the following is(are) hyper parameters of a decision tree?
   1. max\_depth B) max\_features

C) n\_estimators D) min\_samples\_leaf

Ans: A) max\_depth , D) min\_samples\_leaf

-----------------------------------------------------------------------------------------------------------------------------------------------

# Q10 to Q15 are subjective answer type questions, Answer them briefly.

-----------------------------------------------------------------------------------------------------------------------------------------------

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

Ans: An outlier is a value in a data set that is very different from the other values. That is, outliers are values unusually far from the middle.

IQR is used to **measure variability** by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.

* Q1 represents the 25th percentile of the data.
* Q2 represents the 50th percentile of the data.
* Q3 represents the 75th percentile of the data.

If a dataset has 2n / 2n+1 data points, then  
 Q1 = median of the dataset.  
 Q2 = median of n smallest data points.  
 Q3 = median of n highest data points.

IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 – Q1. The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.

-----------------------------------------------------------------------------------------------------------------------------------------------

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

Ans: In **Bagging** the result is obtained by averaging the responses **of** the N learners (or majority vote). However, **Boosting** assigns a second set **of** weights, this time **for** the N classifiers, in order to take a weighted average **of** their estimates

-----------------------------------------------------------------------------------------------------------------------------------------------

1. What is adjusted R2 in logistic regression. How is it calculated?

Ans:  The adjusted R-squared adjusts for the number of terms in the model. Importantly, its value increases only when the new term improves the model fit more than expected by chance alone. The adjusted R-squared value actually decreases when the term doesn’t improve the model fit by a sufficient amount.

Adjusted r-square is a modified form of r-square whose value increases if new predictors tend to improve model’s performance and decreases if new predictors does not improve performance as expected.

{R}^{2}\_{adj} = {R}^{2} - (1-{R}^{2})\*{p}/{n-p-1}

Here, p is the no. of regressors and n is the sample size.  
if the newly added variable is good enough to improve model’s performance, then it will overwhelm the decrease due to p. Otherwise, increase in p will decrease adjusted r-square value.

-----------------------------------------------------------------------------------------------------------------------------------------------

1. What is the difference between standardisation and normalisation?
2. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.