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

Answer:(C)

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

C) Recursive feature elimination D) Ridge Regularisation

Answer:(C)

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

C) hyperplane D) polynomial

Answer:(C)

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

Answer:(D)

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

Answer:(A)

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

Answer:(B)

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

Answer:(A)

# 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

Answer: ( D)

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.

Answer: (B) and (A)

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

Answer: (A) and (D)

# 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.

Answer: An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

IQR:

* The minimum or lowest value of the dataset
* The first quartile *Q*1, which represents a quarter of the way through the list of all data
* The [median](https://www.thoughtco.com/what-is-the-median-3126370) of the data set, which represents the midpoint of the whole list of data
* The third quartile *Q*3, which represents three-quarters of the way through the list of all data
* The maximum or highest value of the data set.

Formula:

IQR =Q3-Q1

Using the Interquartile Rule to Find Outliers

1. Calculate the interquartile range for the data.
2. Multiply the interquartile range (IQR) by 1.5 (a constant used to discern outliers).
3. Add 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier.
4. Subtract 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier.
5. What is the primary difference between bagging and boosting algorithms?

Answer: Bagging is a way to decrease the variance in the prediction by generating additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data. Boosting is an iterative technique which adjusts the weight of an observation based on the last classification.

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

Answer: The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of

predictors in the model. The adjusted R-squared increases only if the new term improves the model more

than would be expected by chance. It decreases when a predictor improves the model by less than

expected by chance.

Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables

(predictors), total sample size. Every time you add a independent variable to a model, the R-squared

increases, even if the independent variable is insignificant.

1. What is the difference between standardisation and normalisation?

Answer: The Difference between normalization and standardization are sometimes used interchangeably, but they usually refer to different things. Normalization usually means to scale a variable to have a values between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1.

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

Answer: Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

Advantage:

1.It ‘validates’ the performance of your model on multiple ‘folds’ of your data.

2.It can balance out the predicted features’ classes if you are dealing with an unbalanced dataset.

DisAdvantage:

K-fold doesn’t really work well with sequential data (a time series of some kind). If you use it you’d be predicting past values with future value training, which is not a good way to go about things.