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:

ANS C) between -1 and 1

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

ANS D) Ridge Regularisation

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

ANS C) hyperplane

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

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)

ANS C) old coefficient of ‘X’ ÷ 2.205

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

ANS C) decreases

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

ANS C ) Random Forests 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. B, C

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

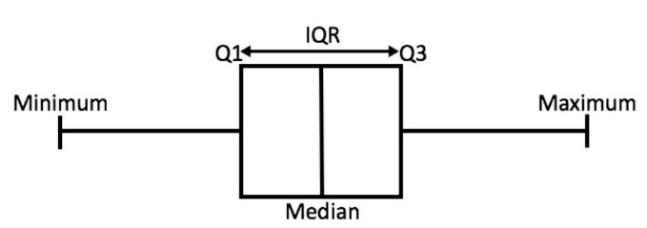
C) n\_estimators D) min\_samples\_leaf

ANS – A,B,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.

ANS - - An **outlier** is an observation that lies an abnormal distance from other values in a random sample from a population. Examination of the **data** for unusual observations that are far removed from the mass of **data.**



**IQR Method** – (Inter-Quartile Range): To detect the outliers using IQR method

* The median is the median (or centre point), also called second quartile, of the data (resulting from the fact that the data is ordered).
* Q1 is the first quartile of the data, i.e., to say 25% of the data lies between minimum and Q1.
* Q3 is the third quartile of the data, i.e., to say 75% of the data lies between minimum and Q3.

We define range here, let’s call it decision range, and any data point lying outside this range is considered as outlier and is accordingly dealt with -

IQR = Q3 - Q1

Lower Bound: (Q1 - 1.5 \* IQR)

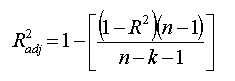
Upper Bound: (Q3 + 1.5 \* IQR)

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

|  |  |
| --- | --- |
| **Bagging** | **Boosting** |
| 1. Simplest way of combining predictions that belong to the same type. | A way of combining predictions that belong to the different types. |
| 1. Aim to decrease variance, not bias. | Aim to decrease bias, not variance. |
| 1. Each model receives equal weight. | Models are weighted according to their performance |
| 1. Each model is built independently. | New models are influenced by performance of previously built models. |
| 1. Different training data subsets are randomly drawn with replacement from the entire training dataset. | Every new subsets contains the elements that were misclassified by previous models |
| 1. Bagging tries to solve over-fitting problem. | Boosting tries to reduce bias. |
| 1. If the classifier is unstable (high variance), then apply bagging. | If the classifier is stable and simple (high bias) the apply boosting. |

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

Logistic regression models are fitted using the method of maximum likelihood - i.e. the parameter estimates are those values which maximize the likelihood of the data which have been observed. McFadden's R squared measure is defined as.



here:

* N is the number of points in your data sample.
* K is the number of independent regressors, i.e. the number of variables in your model, excluding the constant.

**Adjusted R2**:

* Both R2 and the adjusted R2 give you an idea of how many data points fall within the line of the [regression equation](https://www.statisticshowto.com/what-is-a-regression-equation/).
* There is one main difference between R2 and the adjusted R2: R2 assumes that every single variable explains the variation in the [dependent variable](https://www.statisticshowto.com/dependent-variable-definition/).
* The adjusted R2 tells you the percentage of variation explained by only the [independent variables](https://www.statisticshowto.com/independent-variable-definition/) that actually affect the dependent variable.

1. What is the difference between standardisation and normalisation?

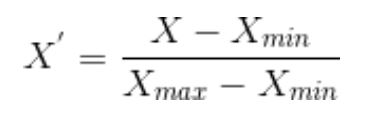
## Normalization

Normalization usually rescales features to [0,1][0,1].[1](https://maristie.com/blog/differences-between-normalization-standardization-and-regularization/#fn:se) That is,

x′=x−min(x)max(x)−min(x)x′=x−min(x)max(x)−min(x)

It will be useful when we are sure enough that there are no anomalies (i.e. outliers) with extremely large or small values. For example, in a recommender system, the ratings made by users are limited to a small finite set like {1,2,3,4,5}{1,2,3,4,5}.

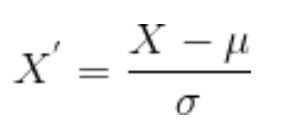
In some situations, we may prefer to map data to a range like [−1,1][−1,1] with zero-mean.[2](https://maristie.com/blog/differences-between-normalization-standardization-and-regularization/#fn:quora_nsr) Then we should choose mean normalization.

****

In this way, it will be more convenient for us to use other techniques like matrix factorization.

## **Standardization**

Standardization is widely used as a preprocessing step in many learning algorithms to rescale the features to zero-mean and unit-variance.[3](https://maristie.com/blog/differences-between-normalization-standardization-and-regularization/#fn:wiki_fs)



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

**Cross Validation**: CV 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:**

* **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.  
    
  **Note:** Chances of overfitting are less if the dataset is large. So, Cross Validation may not be required at all in the situation where we have sufficient data available.
* **Hyperparameter Tuning:** Cross Validation helps in finding the optimal value of hyperparameters to increase the efficiency of the algorithm.

**Disadvantages:**

**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. 