MACHINE LEARNING WORKSHEET –

1. The value of correlation coefficient will always be:

**Ans. C) between -1 and 1**

2. Which of the following cannot be used for dimensionality reduction? **ANS.D) Ridge Regularisation**

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

**Ans .C) hyperplane**

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

**Ans. A) Logistic Regression**

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)

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

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

**Ans. B) increases**

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

**Ans. C) Random Forests are easy to interpret D) Random Forests provide a reliable feature importance estimate**

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

Ans. **B) Principal Components are calculated using unsupervised learning techniques**

**C) Principal Components are linear combinations of Linear Variables.**

9. Which of the following are applications of clustering?

**Ans. 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.**

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

**Ans. A) max\_depth**

**B) max\_features**

**D) min\_samples\_leaf**

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

**Ans.** **OUTLIER** is a commonly used terminology by analysts and data scientists as it needs close attention else it can result in wildly wrong estimations. Simply speaking, Outlier is an observation that appears far away and diverges from an overall pattern in a sample.

Let’s take an example, we do customer profiling and find out that the average annual income of customers is $0.8 million. But, there are two customers having annual income of $4 and $4.2 million. These two customers annual income is much higher than rest of the population. These two observations will be seen as Outliers.

There are many ways to detect outliers such as –Z score,extreme value analysis(parametric),statistical modeling,IQR.

**Inter Quartile Range (IQR):**

IQR = Q3-Q1

Where, Q1 = 25th %ile of the data

Q2 = 50th %ile (a.k.a. median)

Q3 = 75th %ile of the data.

Upper bound = Q3 + 1.5\*Q3

Lower Bound = Q1 – 1.5\*Q1

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

Ans. **Bagging** and **boosting** are both ensemble learning methods in machine learning.

**Bagging** and **boosting** are similar in that they are both ensemble techniques, where a set of weak learners are combined to create a strong learner that obtains better performance than a single one.

Ensemble learning helps to improve machine learning model performance by combining several models. This approach allows the production of better predictive performance compared to a single model.

The basic idea behind ensemble learning is to learn a set of classifiers (experts) and to allow them to vote. This diversification in Machine Learning is achieved by a technique called **ensemble learning**. The idea here is to train multiple models, each with the objective to predict or classify a set of results.

**Bagging** and **boosting** are two types of ensemble learning techniques. These two decrease the variance of single estimate as they combine several estimates from different models. So the result may be a model with higher stability.

The main causes of error in learning are due to **noise, bias and variance**. Ensemble helps to minimize these factors. By using ensemble methods, we’re able to increase the stability of the final model and reduce the errors mentioned previously.

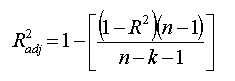
* **Bagging helps to decrease the model’s variance.**
* **Boosting helps to decrease the model’s bias.**

These methods are designed to improve the stability and the accuracy of Machine Learning algorithms. Combinations of multiple classifiers decrease variance, especially in the case of unstable classifiers, and may produce a more reliable classification than a single classifier.

Q.13. What is adjusted R2 in logistic regression. How is it calculated?

**Answer.** R2 shows how well terms (data points) fit a curve or line. Adjusted R2 also indicates how well terms fit a curve or line, but adjusts for the number of terms in a model. If you add more and more **useless** [variables](https://www.statisticshowto.com/probability-and-statistics/types-of-variables/)to a model, adjusted r-squared will decrease. If you add more **useful** variables, adjusted r-squared will increase.  
Adjusted R2 will always be less than or equal to R2.

You only need R2 when working with[samples](https://www.statisticshowto.com/sample/). In other words, R2 isn’t necessary when you have data from an entire [population](https://www.statisticshowto.com/what-is-a-population/).

The formula is:  
[](https://www.statisticshowto.com/wp-content/uploads/2013/09/r-squared-adjusted.jpg)

Where, n = number of data points in the dataset

K = Number of features in the dataset excluding the constant term

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

**Answer.** In **Normalization** a dataset is scaled in such a way that all the data points lie between 0 and 1. Normalization is often called min-max scaling. Formula for Normalization is as follows:



Whereas, In **Standardization** a dataset is scaled in such a way that the mean of data points becomes 0 and standard deviation is 1. The transformed data may be positive as well as negative in standardization. The formula for standardization is as follows:



Where, = ith data point

= sample mean

= sample standard deviation

15. 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.

The three steps involved in cross-validation are as follows:

1. Reserve some portion of sample data-set.
2. Using the rest data-set train the model.
3. Test the model using the reserve portion of the data-set.

**Advantage** -There is no need of separate validation data, cross validation reduces chances of overfitting and gives a more generic model.

**Disadvantage** -It takes more time to fit the model over a large dataset and the model built is more complex than the basic model.