**PRAVEER SAXENA- BATCH 1814**

**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:

A) between 0 and 1 B) greater than -1 C) between -1 and 1 D) between 0 and -1

**Answer: 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

**Answer : D) Ridge Regularisation**

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

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

**Answer : 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

**Answer : 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)

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

**Answer: B) same as old coefficient of ‘X’**

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

**Answer: 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

**Answer: C) Random Forests are easy to interpret**

**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

**Answer: 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?

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.

**Answer: A), C), D)**

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

**Answer: 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.**

**Solution:**

Sometimes a dataset can contain extreme values that are outside the range of what is expected and unlike the other data. These are called outliers.

Any set of data can be described by its [five-number summary](https://www.thoughtco.com/what-is-the-five-number-summary-3126237). These five numbers, which give you the information you need to find patterns and outliers, consist of (in ascending order):

* The minimum or lowest value of the dataset
* The first quartile Q1, 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 Q3, which represents three-quarters of the way through the list of all data
* The maximum or highest value of the data set.

**Inter Quartile Range(IQR) method for outlier detection.**

This is done using these steps:

1. Calculate the interquartile range for the data.( IQR = *Q*3 – *Q*1)
2. Multiply the interquartile range (IQR) by 1.5
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.

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

**Solution:**

The bagging techniques aim to decrease variance i.e it tries to solve over-fitting problem. The boosting techniques aims to reduce bias i.e it tries to solve under-fitting problem.

In bagging, First, we create random samples of the training data set (sub sets of training data set). Then, we build a classifier for each sample. Finally, results of these multiple classifiers are combined using average or majority voting.

Boosting is a sequential process, where each subsequent model attempts to correct the errors of the previous model. The succeeding models are dependent on the previous model.

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

**Solution :**

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 Pseudo R-Squared (adjusted) is a goodness of fit metric for logistic regression**. This approach is similar to **McFadden’s Pseudo R-Squared** but the model is penalized for including too many predictors. This adjustment, however, makes it possible to have negative values for the McFadden’s adjusted Pseudo R-squared.

It is calculated using :

R2adj = 1 – [ln LL(Mˆfull)-K]/[ln LL(Mˆintercept)].

The *LL(Mˆfull)* is the log likelihood of the logit model selected and the *LL(Mˆintercept)* is the log likelihood if the model just had an intercept.Also , *K* is the number of regressors in the model.

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

**Solution**:

Standardisation and Normalization both are feature scaling techniques. Standardization is a scaling technique in which it transforms the data to have mean of zero and standard deviation of 1.In this case, the values are not restricted to a particular range. Example: Standard Scaler. In contrast to standardization, the Normalization is a scaling technique in which it rescales the data to have values between 0 and 1. Example- Min-Max scaling. Also, Standardisation is more robust to outliers.

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

**Solution:**

The purpose of cross validation is to assess how your prediction model performs with an unknown data. Cross validation (CV) is one of the technique used to test the effectiveness of a machine learning models, it is also a re-sampling procedure used to evaluate a model if we have a limited data. To perform cross validation we need to keep aside a sample/portion of the data on which is not used to train the model, later use this sample for testing/validating. Two of the very common approaches of cross validation are Train-Test split and K-Fold cross validation.

One of the advantages of cross validation is in finding the optimal value of hyperparameters to increase the efficiency of the algorithm. The other advantage can be to know, how the model is performing to unseen data so that overfitting can be avoided.

One of the disadvantages of cross validation is increased time to build the final model and other disadvantage is that, some of the data kept aside for testing are not used for training the model.