MACHINE LEARNING WORKSHEET – 1

1. The value of correlation coefficient will always be:

Answer: C) between -1 and 1

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

Answer: D) Ridge Regularisation

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

Answer: C) hyperplane

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

Answer: C) Decision Tree Classifier

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?

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

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

Answer: B) increases

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

Answer: C) Random Forests are easy to interpret

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

Answer: D) All of the above

1. Which of the following are applications of clustering?

Answer:

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.

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

Answer: A) max\_depth B) max\_features and D) min\_samples\_leaf

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

Answer: Inter Quartile Range is the difference between the third quartile and the first quartiles. Difference between Q3 and Q1 is the IQR. It is a measure of dispersion.

IQR = Q3 – Q1.

It plays a role when dealing with varieties of problems, it also helps to find the outliers in the data.

Ex. Let’s take a number set and will find IQR

Numbers: 4, 4, 10, 11, 15, 7, 14, 12, 6

Sorting data into ascending order 4, 4, 6, 7, 10, 11, 12, 14, 15

Let’s find the median: Median is 10, as 10 is in the middle. Now, will find min and max. Here min is 4 and max is 15. Now, will find Q1 to do so we will calculate 4+6/2 = 5. Q3 = 12+14/2 = 13.

IQR = 13-5 = 8.

Outlier: An outlier is the data which lies outside the overall distribution. Outlier is the data point which lies outside a rule which says that a data point is an outlier if it is more that 1.5 \* (IQR) above the third quartile (Q3) and below the first quartile. To define outlier low outliers is Q1-1.5\*(IQR), and above Q3 + 1.5(IQR).

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

|  |  |
| --- | --- |
| Bagging | Boosting |
| In Bagging every model is independent constructed. | In Boosting new models are affected by the performance of previous model. |
| It decreases the variance. | It decreases the bias. |
| It applies to high variance dataset. | It applied to high bias dataset. |
| It tries to reduce the over-fitting issue | It tries to reduce bias. |
| It give equal weight to each model | It gives weight according to model performance. |
| Random Forest | AdaBoost, GMB, XGB |

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

Answer: Before defining adjusted R2, let’s define R2: It is the proportion of the variation in our dependent variable explained by our independent variable.

Adjusted R2: It measures the proportion of variation explained by the independent variable which helps us to explain our independent variable.

Adjusted R2 can be calculated based on value of R-squared, number of independent variables to total sample size.

R2 = (1-R2)(N-1)/N-p-1

Where R2 = sample R sq.

P = Number of predictors

N = Total number of sample.

14. What is the difference between standardisation and normalisation?

|  |  |
| --- | --- |
| Standardisation | Normalisation |
| It does not bound with the range. | Scales values between [0, 1] or [-1, 1] |
| It transform data to have a mean of zero and standard deviation of 1. | It scales value between 0 and 1 |
| It is also called as scaling | It is also called as z-score |
| It is less affected by the outliers | It affects by outliers |
| It is used when we want 0 mean and standard deviation | It is used when data is at different scale. |
| StandardScaler | MinMaxScaler |

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

Answer: Cross Validation is used to assess the predictive performance of the models and to judge how they perform outside the sample to a new data set also known as test data. It also called as rotation estimation. It separates our data set into two subsets. One subset we use for training and other for testing, it performs the exercise again by swapping the data sets. Then at last it return the average test result. This is called as k fold cross validation. Similarly if we divide our entire data set in to five sub sets and perform the exercise five times and report the average test result  then that would be 5 fold cross validation.