WORKSHEET 4

MACHINE LEARNING

INTERNSHIP 32

FLIP ROBO TECHNOLOGIES



Submitted by: Shahla M

SME : Khushbooo Garg

**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 :

C) Recursive feature elimination

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 :

D) Support Vector Classifier

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 :

C) old coefficient of ‘X’ ÷ 2.205

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 :

B) Random Forests explains more variance in data then decision trees

**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) 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.

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

D) min\_samples\_leaf

**Q10 to Q15 are subjective answer type questions, Answer them briefly.**

11. 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.. The outliers may suggest experimental errors, variability in a measurement, or an anomaly. The age of a person may wrongly be recorded as 200 rather than 20 Years. Such an outlier should definitely be discarded from the dataset.  
However, not all outliers are bad. Some outliers signify that data is significantly different from others. For example, it may indicate an anomaly like bank fraud or a rare disease..

IQR is used to **measure variability** by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.

* Q1 represents the 25th percentile of the data.
* Q2 represents the 50th percentile of the data.
* Q3 represents the 75th percentile of the data.

If a dataset has **2n / 2n+1** data points, then  
Q1 = median of the dataset.  
Q2 = median of n smallest data points.  
Q3 = median of n highest data points.

IQR is the range between the first and the third quartiles namely

**Q1 and Q3: IQR = Q3 – Q1.**

The data points which fall below

**Q1 – 1.5 IQR or above Q3 + 1.5 IQR**

are outliers.

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

ANSWER :

**Bagging** is a learning technique that helps in improving the performance, implementation, and accuracy of machine learning algorithms. Or in other words, we can say that it is basically a machine learning ensemble meta-algorithm crafted to enhance the stability and accurateness of algorithms utilised in statistical classification and regression. It is also known as Bootstrap aggregating.

**Boosting** is an ensemble technique that iteratively alters the weight of observation based on the last classification. It helps in enhancing the power of a machine learning program by counting more complicated or qualified algorithms. This process can also help in reducing bias and variance in machine learning.

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

ANSWER :

Adjusted R2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs.

R2 tends to optimistically estimate the fit of the linear regression. It always increases as the number of effects are included in the model. Adjusted R2 attempts to correct for this overestimation. Adjusted R2 might decrease if a specific effect does not improve the model.

Adjusted R squared is calculated by dividing the residual mean square error by the total mean square error (which is the sample variance of the target field). The result is then subtracted from 1.

14. What is the difference between standardisation and normalisation?

ANSWER :

In **Normalisation**, the change in values is that they are at a standard scale without distorting the differences in the values. Whereas, **Standardisation** assumes that the dataset is in Gaussian distribution and measures the variable at different scales, making all the variables equally contribute to the analysis.

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

ANSWER :

**Cross-Validation** is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model.

Advantage :

**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

Disadvantage:

**Changing Needs:** One of the most significant disadvantages of data validation is that data must be re-validated once specific changes to the data are made. As new data types and inputs are added, schema models and mapping documentation will need to be updated.