CML-CC-516 MACHINE LEARNING LABORATORY

Lab Cycle

1. Consider the dataset in the following table, of houses represented by five training examples. The target attribute is 'Acceptable', which can have values 'Yes' or 'No'. This is to be predicted based on the other attributes of the house.

House	Furniture	No.rooms	New kitchen	Acceptable
1	No	3	Yes	Yes
2	Yes	3	No	No
3	No	4	No	Yes
4	No	3	No	No
5	Yes	4	No	Yes

Compute the entropy of the target attribute

i. Construct a decision tree from the above examples, that would be learned by the ID3 algorithm.

ii. Show the value of information gain for each candidate attribute at each step in the construction of the tree.

2. Maharashtra has the first highest GDP coefficient in India. It is very high segregation in the annual income of rich and poor. In this study, we have to come up with an accurate predictive algorithm to estimate the annual income of each individual in Maharashtra. The income is given,
   1. Below rupees 4,00,000
   2. Rupees between 4,00,000 - 15,00,000
   3. More than rupees15,00,000

   Other relevant information for an individual are,
   1. Age, 2. Gender, 3. Highest educational qualification, 4. Forking in Industry, 5. Residence in Metro/Non-metro.
   You are requested to design an algorithm which gives an accurate predictive for an individual who has following traits:
   1. Age: 35 years, 2, Gender: Male, 3. Highest educational qualification: Diploma holder, 4. Industry: Manufacturing, 5. Residence: Metro
   You can select a random forest technique to make this prediction in this study. Repeat the process several times and make a prediction on each observation.

3. Given the following data, which specify classifications for nine combinations of VAR1 and VAR2 predict a classification for a case where VAR1=0.906 and VAR2=0.606, using the result of k- means clustering with 3 means (i.e., 3 centroids) calculate the precision, recall and F1 score of the above classification method

VAR1	VAR2	CLASS
1.713	1.586	0
0.180	1.786	1
0.353	1.240	1
0.940	1.566	0

1.486	0.759	1
1.266	1.106	0
1.540	0.419	1
0.459	1.799	1
0.773	0.186	1

4. Calculate the classification accuracy of K-means algorithm with IRIS dataset.

5. The word is a sequence of characters. The sequence is important when a word is meaningful. For example, 'qc' may have less chance to occur than 'is'. We can say the probability of occurrence of the second word is greater than the first. How can we identify such a sequence of occurrences in words? Design an SVM that classify t two-letter words in English or not from an input word.

6. Limitation of the K-means algorithm is that it is very sensitive to the initial centers. Changing the initial centers strongly influences the results. Use genetic algorithm for solving the K-means clustering problem given above (Question no:7) and find out the optimal centre for 3 clusters.

7. Apply the concept of ANN for diagonizing whether a patient has diabetics or not. Evaluate the model based on its performance matrices also. (Hint : Create your own dataset)

8. Apply the concept of perceptron for performing Sentiment analysis using perceptron.

9. Demonstrate the diagonosis of heart disease using a Standard Heart Disease database from Kaggle/UCI repository and find the best machine learning algorithm among Decsion tree and Support Vector Machines in the particular scenario. Evaluate the performance matrices also.

10. Apply fuzzy logic for detecting Spam e-mails. Use this knowledge to classify a sample and evaluate the performance metrices.

11. Implement different machine learning algorithms to perform online fraud detection and find the best machine learning algorithm by evaluating the performance matrices of each.