Assignment
Unsupervised Learning

Q01-Q12 => 1 mark each

1. Which of the following is a common use of unsupervised clustering?
   a) Detect outliers
   b) Determine a best set of projection for supervised learning
   c) Evaluate the likely performance of a supervised learner model
   d) Determine if meaningful relationships can be found in a dataset
   e) All of the above

2. Which statement is true about the K-Means algorithm?
   a) All attribute values must be categorical
   b) The output attribute must be categorical
   c) Attribute values may be either categorical or numeric
   d) All attributes must be numeric

3. Amongst below data transformation technique which works well when minimum and maximum values for a real-valued attribute are known.
   a) min-max normalization
   b) decimal scaling
   c) z-score normalization
   d) logarithmic normalization

4. This technique uses mean and standard deviation scores to transform real-valued attributes.
   a) decimal scaling
   b) min-max normalization
   c) z-score normalization
   d) logarithmic normalization

5. This unsupervised clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iteration.
   a) agglomerative clustering
   b) conceptual clustering
   c) K-Means clustering
   d) expectation maximization

6. What is the minimum no. of variables/features required to perform clustering?
   a) 0
   b) 1
   c) 2
   d) 3

7. Which of the following algorithm is most sensitive to outliers?
   a) K-means clustering algorithm
   b) K-medians clustering algorithm
   c) K-modes clustering algorithm
   d) K-medoids clustering algorithm

8. The most popularly used dimensionality reduction algorithm is Principal Component Analysis (PCA).

   1. PCA is an unsupervised method
   2. It searches for the directions that data have the largest variance
   3. Maximum number of principal components <= number of features
   4. All principal components are orthogonal to each other

   Which is above is true.

   A. 1 and 2

   B. 1 and 3

   C. 2 and 3

   D. 1, 2 and 3

   E. 1,2 and 4

   F. All of the above

Answer the following using TRUE /FALSE (Q9-12)

9. Given historical weather records, can we predict if tomorrow's weather will be sunny or rainy using K-means.
10. Given a set of news articles from many different websites, using k-means can you find out what topics are the main topics covered.
11. Dimensionality reduction algorithms are one of the possible ways to reduce the computation time required to build a model.
12. PCA can be used for projecting and visualizing data in lower dimensions.

Q13 => 6 mark

13. Point out pros and cons (at least one) for the following unsupervised algorithms

    a) K-Means Clustering
    b) Scatter Plots
    c) Principal Components Analysis

14. MARKET BASKET ANALYSIS:

The dataset called "Online Retail" from UCI Machine Learning repository contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered online retailer.

Perform market basket analysis in python/R with your preference of tool to obtain following results.

- What time do people often purchase online? [1 Mark]
- How many items each customer buy? [1 Mark]
- Top 10 best sellers [1 Mark]
- Share your insights which can help retailer to increase his profits and few association rules [4 Marks]