Batch 6 – Data Mining 1 – Unsupervised Learning - Assignment 1

- **This is an INIDIVIDUAL assignment.**
- **Please upload your work on the LMS by deadline as specified on LMS.**
- **You may use any tool to work on your assignment.**
- **In your submission you must include the output or refer to the relevant output in the excel file while answering any question or justifying your answers.**
- **An answer without justification will not be awarded full credit.**

---

1. **Marketing to Frequent Fliers.** The file **EastWestAirlinesCluster.xls** (available on the textbook website http://dataminingbook.com/) contains information on 4000 passengers who belong to an airline's frequent flier program. For each passenger the data include information on their mileage history and on different ways they accrued or spent miles in the last year. The goal is to try to identify clusters of passengers that have similar characteristics for the purpose of targeting different segments for different types of mileage offers.
   a) Apply hierarchical clustering with Euclidean distance and Ward's method. Make sure to standardize the data first. How many clusters appear?
   b) What would happen if the data were not standardized?
   c) Compare the cluster centroids to characterize the different clusters and try to give each cluster a label.
   d) To check the stability of the clusters, remove a random 5% of the data (by taking a random sample of 95% of the records), and repeat the analysis. Does the same picture emerge?
   e) Use k-means clustering with the number of clusters that you found above in Part (a). Does the same picture emerge? If not, how does it contrast or validate the finding in Part c above?
   f) Which cluster(s) would you target for offers, and what type of offers would you target to customers in that cluster? Include proper reasoning in support of your choice of cluster(s) and the corresponding offer(s).

2. **Step 1:** Download the Wine data from the UCI machine learning repository (http://archive.ics.uci.edu/ml/datasets/Wine)

   Step 2: Do a Principal Components Analysis (PCA) on the data. Please include (copy-paste) the relevant software outputs in your submission while answering the following questions.
   a. Enumerate the insights you gathered during your PCA exercise. (*Please do not clutter your report with too MANY insignificant insights as it will dilute the value of your other significant findings*)

b. What are the social and business values of those insights, and how the value of those insights can be harnessed?

Step 3: Do a cluster analysis using (i) all chemical measurements (ii) using two most significant PC scores. Please include (copy-paste) the relevant software outputs in your submission while answering the following questions.

c. Any more insights you come across during the clustering exercise?

d. Are there clearly separable clusters of wines? How many clusters did you go with? How the clusters obtained in part (i) are different from or similar to clusters obtained in part (ii), qualitatively?

e. Could you suggest a subset of the chemical measurements that can separate wines more distinctly? How did you go about choosing that subset? How do the rest of the measurements that were not included while clustering, vary across those clusters?