Unsupervised Learning – Individual Assignment 1

Submission Deadline: 27th February, 11:55 PM

- This is an individual assignment. **Honor code: 3N-a**.
- Technology glitches can happen and marriages too 😊. So, plan in advance.
- Please upload your work on the LMS by the deadline as specified on LMS.
- Use R to work on your assignment.
- In your submission, you must include the R file, the answer sheet in pdf. Add relevant output from R wherever possible in your answer sheet.
- An answer without a justification will not be awarded any credit even if the answer is correct.

---

1. Hierarchical Clustering: 6 x 5=30

The dataset on American College and University Rankings contains information on 1302 American colleges and universities offering an undergraduate program. For each university, there are 17 continuous measurements (such as tuition and graduation rate) and 2 categorical measurements (such as location by state and whether it is a private or public school). Note that many records are missing some measurements. Our first goal is to estimate these missing values from "similar" records. This will be done by clustering the complete records and then finding the closest cluster for each of the partial records. The missing values will be imputed from the information in that cluster.

a. Remove all records with missing measurements from the dataset.

b. For all the continuous measurements, run hierarchical clustering using complete linkage and Euclidean distance. Make sure to normalize the measurements. From the dendrogram, how many clusters seem reasonable for describing these data?

c. Compare the summary statistics for each cluster and describe each cluster in this context (e.g., "Universities with high tuition, low acceptance rate…").

d. Use the categorical measurements that were not used in the analysis (State and Private/Public) to characterize the different clusters. Is there any relationship between the clusters and the categorical information?

e. What other external information can explain the contents of some or all of these clusters?

f. Consider Tufts University, which is missing some information. Compute the Euclidean distance of this record from each of the clusters that you found above (using only the measurements that you have). Which cluster is it closest to? Impute the missing values for Tufts by taking the average of the cluster on those measurements.

2. **Principal Component Analysis**: 5x5=25

The file ToyotaCorolla.csv contains data on used cars (Toyota Corollas) on sale during late summer of 2004 in the Netherlands. It has 1436 records containing details on 38 attributes, including Price, Age, Kilometers, HP, and other specifications. The goal will be to predict the price of a used Toyota Corolla based on its specifications.

a. Identify the categorical variables.

b. Explain the relationship between a categorical variable and the series of binary dummy variables derived from it.

c. How many dummy binary variables are required to capture the information in a categorical variable with N categories?

d. Use R to convert the categorical variables in this dataset into dummy variables, and explain in words, for one record, the values in the derived binary dummies.

e. Use R to produce a correlation matrix and matrix plot. Comment on the relationships among variables.

3. **Test your concepts**: 4 x 5=20

Consider the 3-means algorithm on a set S consisting of the following 6 points in the plane: a= (0,0), b = (8,0), c=(16,0), d=(0,6), e=(8,6), f=(16,6). The algorithm uses the Euclidean distance metric to assign each point to its nearest centroid; ties are broken in favor of the centroid to the left/down. A starting configuration is a subset of 3 starting points from S that form the initial centroids. A 3-partition is a partition of S into 3 subsets; thus {a,b,e}, {c,d}, {f} is a 3-partition; clearly any 3-partition induces a set of three centroids in the natural manner. A 3-partition is stable if repetition of the 3-means iteration with the induced centroids leaves it unchanged.

a. How many starting configurations are there?

b. What are the stable 3-partitions?

c. What is the number of starting configurations leading to each of the stable 3-partitions in (b) above?

d. What is the maximum number of iterations from any starting configuration to its stable 3-partition?

---End--