Math 444 – Data Mining Spring 2020

Assignment 2

1. Write **your own** k-means and k-medoids algorithms. In your report **include the whole code** for both algorithms in the appendix. Since you may be asked to use can use them later, write your algorithms as a function. If you are coding in MATLAB, the structure of the k-medoids routine could be

   ```
   function [I_assign,I_bar] = my_k_medoids(k,D,tau),
   ```

   where the input consists of

`k`	=	number of clusters,
`D`	=	distance matrix of size $p \times p$,
`tau`	=	tolerance, a small number used for the stopping criterion (see the lecture notes),

   and the output variables are

`I_assign`	=	assignment vector of length p, indicating the cluster of each data vector,
`I_bar`	=	index vector of length k, indicating the medoids.

   To initialize the algorithm, include the following step:

   (i) Choose the initial clusters by picking randomly k data vectors from your data to be the initial medoids, computing the corresponding tightness.

   (ii) Repeat the initialization 20 times, and pick the initial clustering that corresponds to the lowest tightness.

2. Test your k-means and k-medoids algorithm with $k = 3$ on the data file `WineData.mat` containing a matrix $\mathsf{X} \in \mathbb{R}^{13\times 178}$ of chemical analysis data of wines derived from the same area of Italy but originating from three different cultivars. The attributes in each column of X are concentrations/levels of the following substances:

   1 Alcohol
   2 Malic acid
   3 Ash
   4 Alcalinity of ash
   5 Magnesium
   6 Total phenols
   7 Flavonoids
   8 Nonflavonoid phenols
   9 Proanthocyanins
   10 Color intensity
   11 Hue
   12 OD280/OD315 of diluted wines
   13 Proline

   Comments on how well, or badly, your algorithms were able to identify the three different clusters by comparing your results with the true annotation recorded in the vector I. Report whether the three wine types were easy to cluster based on the recorded attributes.

3. Download the data file `CardiacData.mat`. It will contain a matrix X of size 23×187. The dataset reports 22 attributes extracted from cardiac Single Proton Emission Computed Tomography (SPECT) images. Each patient was classified into on eof two categories: normal or abnormal. The database of 187 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images. As a result, 44 continuous feature pattern was created for each patient. The pattern was further processed to obtain 22 binary feature patterns. The last entry contains the annotation of the patient, also in the form of 0 or 1.

   (a) Write the distance matrix between the patients, using a dissimilarity index between the 0 and 1 as a distance measure.

   (b) Once you have the distance matrix, run the k-medoids algorithm to cluster the votes in two groups. Then investigate if and how well your clustering corresponds to the classification given by the cardiologist. Write a matrix C of size 2×2 with the following entries:

$$\begin{aligned} c_{11} &= \text{1's in your cluster 1,} \\ c_{12} &= \text{1's in your cluster 0,} \\ c_{21} &= \text{1's in your cluster 0,} \\ c_{22} &= \text{1's in your cluster 1.} \end{aligned}$$

   In the light of your clustering, how well do the attributes represent the state of the patients?