Math 444 – Data Mining Spring 2017

Assignment 2

1. Load from Blackboard the data file `IrisDataAnnotated.mat`. The data set X is the same as the previously used one, while the annotation vector I indicates the iris species,

$$\begin{aligned} 1 &= \textit{Iris setosa}, \\ 2 &= \textit{Iris versacolor}, \\ 3 &= \textit{Iris virginica}. \end{aligned}$$

   (a) Write your own k-means and k-medoids algorithms. In your k-medoid algorithm, use the ℓ^1-distance,

$$\|x - y\|_1 = \sum_{j=1}^{n} |x_j - y_j|.$$

   (b) Run the k-means and k-medoids algorithms with $k = 3$. Use a random initialization to avoid putting the data by chance in three correct groups.

   If the data were a suitable target for the algorithm, you should have the three species in three different clusters. Investigate the quality of the clustering by using the annotation vector of the data: Decide which iris species in each of your cluster represents by a majority, then count the misclassifications of each iris type. Run the test a couple of times to see that the result is not too sensitive to initial clustering, or, if it turns out to be, report that finding.

2. Download from Blackboard the file `BiopsyDataAnnotated.mat`. The file contains a data matrix X of size 9×699, containing breast tissue needle biopsy data from 699 patients, some of which have breast cancer, some have a benign tumor. The explanation of the columns is as follows:

$$x^{(j)} = \begin{bmatrix} \text{Clump Thickness} \\ \text{Uniformity of Cell Size} \\ \text{Uniformity of Cell Shape} \\ \text{Marginal Adhesion} \\ \text{Single Epithelial Cell Size} \\ \text{Bare Nuclei} \\ \text{Bland Chromatin} \\ \text{Normal Nucleoli} \\ \text{Mitoses} \end{bmatrix} \in \mathbb{R}^9, \quad 1 \le j \le N.$$

   Each attribute takes on a value between 1 and 10. Some data is missing, which is indicated by a '`NaN`' (= not a number) in the file. The vector I is an annotation vector, with the annotation

$$\begin{aligned} 1 &= \text{malignant}, \\ 0 &= \text{beningn}. \end{aligned}$$

   After deleting columns containing missing data, run your k-medoids algorithm. Then check if the clustering corresponds to the annotation.

   Investigate the success of the classifier in terms of misclassification. Calculate the *specificity* (or true negative rate) and *sensitivity* (or recall rate) of the method, defined as follows:

$$\text{sensitivity} = \frac{\#\text{ of malignant cases classified correctly}}{\#\text{ of all malignant}},$$

and

$$\text{specificity} = \frac{\#\text{ of benign cases classified correctly}}{\#\text{ of all benign}}.$$

Again, run your algorithm a couple of times to assess the the robustness of the algorithm to initial partitioning.

3. The Matlab script file `CongressVotes.m` contains the votes of the congress representatives in 1984 on 16 issues. The annotation vector I indicates the party membership (republican $= 0$, democrat $= 1$), and the columns of the Matrix X give a yes/no vote of each congress representative on the following 16 issues ("yes" $= 1$, "no" $= -1$, "missing vote" $= 0$):

    1 handicapped-infants
    2 water-project-cost-sharing
    3 adoption-of-the-budget-resolution
    4 physician-fee-freeze
    5 el-salvador-aid
    6 religious-groups-in-schools
    7 anti-satellite-test-ban
    8 aid-to-nicaraguan-contras
    9 mx-missile
    10 immigration
    11 synfuels-corporation-cutback
    12 education-spending
    13 superfund-right-to-sue
    14 crime
    15 duty-free-exports
    16 export-administration-act-south-africa

    Write the distance matrix between the representatives, using a dissimilarity index between the "yes" and "no" votes. If the vote of a representative is missing, it does not contribute to the dissimilarity. Once you have the distance matrix, run the k-medoids algorithm to cluster the votes in two groups. Then investigate if your clustering corresponds to the party line.