RWTH Aachen
Lehrstuhl für Informatik 9
Prof. Dr. T. Seidl

November 23rd, 2014

Exercise 5 for the lecture
Data Mining Algorithms
WS 2015/2016

Hand in your solutions on November 30th before the lecture. The tutorial for this exercise will be held on December 4st. Solutions of groups with less than 3 or more than 4 students will not be graded.

Note: All commands for the R-exercises are required to be provided with comments, indicating which task the commands belong to. All R script files should contain a comment-line with the names and matriculation numbers of all group-members. Send all R-files to yifeng.lu@informatik.rwth-aachen.de. The subject of the mail must start with "[DMA1]".

Exercise 5.1) K-Medoid (PAM) — 5=3+2 points

Consider the following 2-dimensional data set:

x_1= (1,4), x_2= (1,6), x_3 = (2,6), x_4= (3,8) , x_5= (4,3), and x_6= (5,2).

a) Perform the first loop of the PAM algorithm (k=2) using the Euclidian distance. Select x_1 and x_3 as initial medoids and compute the resulting medoids and clusters.

b) How can the clustering result $C_1 = \{x_1, x_5, x_6\}, C_2 = \{x_2, x_3, x_4\}$ be obtained with the PAM algorithm (k=2) using the weighted Manhattan distance

$$d(x, y) = w_1 \cdot |x_1 - y_1| + w_2 \cdot |x_2 - y_2|?$$

Assume that x_1 and x_3 are the initial medoids and give values for the weights w_1 and w_2 for the first and second dimension respectively.

Exercise 5.2) Silhouette-coefficient and k-means — 3 points

Construct a 2-dimensional data set D together with a clustering $\{C_1, C_2\}$ computed by k-means with the following property:

- There exists a point $o \in D$ with a negative silhouette coefficient $s(o) < 0$.

Provide the means of the clusters and compute the silhouette coefficient for the corresponding point o.

Hint: It is possible to find such an example with 5 data points.

Exercise 5.3) EM-algorithm — 7=1+1+1+2+1+1 points

Consider the following data points: x_1=2, x_2=3, x_3 = 4, x_4=12, x_5=13, x_6=14, x_7=15, x_8=16, and an initial clustering C_1= {x_1, x_2, x_3, x_4, x_5} and C_2 = {x_6, x_7, x_8}.

a) Compute the initialization $\boldsymbol{\Theta}$ for EM, i.e. means, variance, and mixing coefficients, based on the initial cluster assignment. (use the sample mean, the sample covariance, and the relative frequencies as initializations of the model parameters)
b) Compute the probability that these points have been generated by the model with parameters $\boldsymbol{\Theta}$, i.e. $p(\mathbf{X}|\boldsymbol{\Theta})$.
c) Compute the first *Expectation* step, i.e. compute the new probabilities $\gamma_j^{new}(x_n)$.
d) Compute the first *Maximization* step, i.e. compute a new model $\boldsymbol{\Theta}^{new}$**.**
e) Compute the second *Expectation* step.
f) Compute a partitioning based on the probabilities $\gamma_j^{new}(x_n)$ after the second *Expectation* step. Did it change with respect to the initial clustering?

Note: You can use R, OpenOffice Calc, MS Excel, etc. to do the explicit computations. You only need to submit the results of each step.

Exercise 5.4) Implementation of k-means (R code) 5 points

Implement the two update steps in the *k-means* algorithm. We provided the basic structure in the file *myKmeans.R*.

Download the *myClustering.zip* file from the L2P. You only need to finish the *assignCluster* and *updateMean* functions in the *myKmeans.R* file. You can test your version of *k-means* with the provided data set. More detailed information can be found in the comments of the R script file.

For submission, you only need to provide the *myKmeans.R* file.

Exercise 5.5) Implementation of EM (R code) 10* points (bonus)

Implement the *EM* algorithm. We provided the basic structure in the file *myEM.R* (see L2P).

For submission, you only need to provide the *myEM.R* file.

Note: You can use the package mvtnorm for multivariate normal distributions.