Intelligent Data Analysis Summer 2017
C. Braune, T. Sabsch

Exercise Sheet 9

Exercise 32 c-Means Clustering

Consider the following two-dimensional data set:

x	1	6	8	3	2	2	6	6	7	7	8	8
y	5	2	1	5	4	6	1	8	3	6	3	7

Process this data set with c-means clustering with $c = 3$ (i.e., try to find 3 clusters)! Use the first three data tuples als initial positions for the cluster centers and observe the migration of the centers.

Exercise 33 c-Means Clustering

In exercises 21 and 22 on sheet 6 we considered a simple two-dimensional data set. Reconsider this data set, but assume that no class information is available for the data points. That is, consider the following data set:

x	3	3	4	4	5	6	7	7	8	9	1	2	2	3	4	5	5	6	7	7
y	1	2	2	3	3	4	4	6	5	7	3	4	5	6	6	7	8	8	8	9

a) Which problem of c-means clustering becomes obvious when this data set is processed with $c = 2$ (i.e., if one tries to find two clusters)?
   Hint: What is the desired result? What is produced by c-means clustering?
   (You need not compute the exact result of the algorithm, a qualitative description suffices. Compare the result to a naive Bayes classifier.)

b) How could one try to cope with this problem?
   Hint: Recall what distinguishes a full and a naive Bayes classifier.

Exercise 34 Fuzzy Clustering

Consider the one-dimensional data set

$$1, 3, 4, 5, 8, 10, 11, 12.$$

We want to process this data set with fuzzy c-means clustering with $c = 2$ (two clusters) and a fuzzifier of $w = 2$. Assume that the cluster centers are initialized to 1 and 5. Execute one step of alternating optimization as it is used for fuzzy clustering, i.e.:

a) Compute the membership degrees of the data points for the initial cluster centers!

b) Compute new cluster centers from the membership degrees computed in this way!

Exercise 35 Fuzzy Clustering

Consider the objective function of fuzzy clustering with a fuzzifier $w = 1$, that is,

$$J(\mathbf{X}, \mathbf{B}, \mathbf{U}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}\, d^2(\beta_i, \vec{x}_j),$$

which is to be minimized under the constraint

$$\forall j \in \{1, \dots, n\} : \quad \sum_{i=1}^{c} u_{ij} = 1.$$

Show that one obtains a hard/crisp assignment of the data points even if the membership degrees u_{ij} may come from the interval $[0, 1]$. That is, show that for the minimum of the objective function J it is $\forall i \in \{1, \dots, c\} : \forall j \in \{1, \dots, n\} : u_{ij} \in \{0, 1\}$.
(Hint: You may find it easier to consider the special case $c = 2$ (two clusters) and to examine the term for a single data point $\vec{x}_j$. Then generalize the result.)

Additional Exercise Fuzzy Clustering

Consider the objective function

$$J(\mathbf{X}, \mathbf{B}, \mathbf{U}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}\, d^2(\beta_i, \vec{x}_j),$$

which is to be minimized under the constraint

$$\forall j \in \{1, \dots, n\} : \quad \sum_{i=1}^{c} \sqrt{u_{ij}} = 1.$$

Derive the update formulae for the membership degrees and the cluster centers using the Euclidean distance. How does the result differ from standard fuzzy clustering with a fuzzifier $w = 2$? (In particular, consider the cluster centers.)