Intelligent Data Analysis — Summer 2016
C. Braune, F. Witt

Exercise Sheet 8

Additional Exercise Application of Clustering and Classification Algorithms

This exercise will be discussed on June 14th/20th. The deadline for submissions is June 13th, 23:59:00 CEST.

- As presented during the lecture on May 25th: sign up at

  `http://learning-challenge.de/`

  with your URZ login name as Nick- or Teamname and your real name as real name.
- Choose the group `IDA2016`.
- You can now select the *Classification* and the *Clustering* rounds. Parameterize the algorithms available in each round to obtain the best possible result for each data set.
- *Classification* tasks will be ranked by the achieved accuracy; *Clustering* tasks by the minimum of Homogenity and Completeness.
- The five best students per round (across all data sets) (right hand side, scored with *stars*), will receive bonus credit for the exercise.

Exercise 29 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 30 Expectation Maximization

Consider again the one-dimensional data set used in exercise 29, which we want to process in this exercise with the expectation maximization algorithm to estimate the parameters of a mixture of two normal/Gaussian distributions. Let the prior probabilities of the two clusters be fixed to $\theta_i = \frac{1}{2}$ and the variances to $\sigma_i^2 = 1$, $i = 1, 2$. (That is, only the expected values of the normal distributions — the cluster centers — are to be adapted.) Use the same values for the initial cluster centers as in exercise 29, that is, 1 and 5. Computeone expectation step and one maximization step, i.e.:

a) Compute the posterior probabilities of the data points for the initial cluster centers!

b) Estimate new cluster centers from the data point weights computed in this way!

Exercise 31 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

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

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 $\forall i \in \{1, \ldots, c\} : \forall j \in \{1, \ldots, n\} : u_{ij} \in \{0, 1\}$.
(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.)

Exercise 32 Agglomerative Clustering

Let the following one-dimensional data set be given:

$$2, 5, 11, 12, 17, 21, 32.$$

Process this data set with hierarchical agglomerative clustering using

a) the centroid method,

b) the single linkage methode,

c) the complete linkage methode!

Draw a dendrogram for each case!

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

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

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.)