Business Analytics
Exercise Sheet 7

Josif Grabocka (josif@ismll.de)
Information Systems and Machine Learning Lab (ISMLL)
Universität Hildesheim

18 June 2013
Submission Until 23 June 2013 23:59

Question 1: K-Means Clustering (5 points)

(a) Explain the optimization function of the K-Means clustering. Briefly describe the two steps of the K-Means algorithm. When does the algorithm stop? (**1 point**)

(b) Implement K-Means in R for an arbitrary number of clusters. Assume every instance in the dataset is a vector of arbitrary dimensionality. Provide your code in a brief format. (**2 points**)

(c) Using the *sal-adv-inc.csv* dataset, experiment clustering the dataset instances with varying number of clusters (from 2 to 10). Does the optimization function value decrease by increasing the number of clusters? Comment on advantages and disatvantages of this behavior. (**2 points**)

Question 2: K-Medoid Clustering (5 points)

(a) What is the primary difference between K-Means and K-Medoids? (**1 point**)

(b) Implement a K-Medoids algorithm in R using the Euclidean distance. Provide your code in a brief format. (**2 points**)

(c) Repeat the experiments of Question 1.c) and comment on the differences of the two methods with respect to the number of clusters. (**2 points**)

Submission

- Electronically to josif@ismll.de
- Email title must be *BA2013-NAME-Tutorial-NO*, e.g.: *BA2013- JosifGrabocka-Tutorial-1*
- Report file must be a PDF with file name like the email title *BA2013-NAME-Tutorial-NO.pdf*
- Source codes and other materials must be a ZIP with file name like the email title.