Human-Oriented Robotics Winter term 2013/2014
Prof. Dr. Kai Arras, Social Robotics Lab University of Freiburg
Lab instructors: Timm Linder, Luigi Palmieri, Billy Okal Department of Computer Science

Exercise 7: K-Means Clustering

Submission: Send your solution to `palmieri@informatik.uni-freiburg.de` until January 13, 2014 with subject line "[`exercises`] `Sheet 7`". All files (Matlab scripts, exported figures, handwritten notes in pdf/jpg format) should be put into a single zip file named `lastname_sheet7.zip`.

For this exercise, you will need to download a dataset file from the course website.

Exercise 7.1: Clustering with k-means

In this exercise you will implement the k-means algorithm, run it on the three provided datasets `five-clusters.txt`, `R15.txt`, and `aggregation.txt` and display the results.

For the implementation there are two steps:

a) The assignment step where you assign each point $\mathbf{x}_n$ to the closest centroid μ_k and

b) The update step where, based on those assignments, you compute the new centroids.

Initialize the centroids by drawing K data points at random from the dataset (**Hint:** use `randperm`). Iterate until the centroids do not change anymore (**Hint:** use `norm`).

Next, visualize the obtained final clustering.

a) Plot the clusters and centroids in different colors. For the data points, use `plot` or, more elegantly, `scatter` with cluster labels as color indices C into the figure's colormap.

b) Plot the boundaries of the centroids using the built-in command `voronoi`

Exercise 7.2: Plotting the objective function

The k-means algorithm minimizes the objective function J (or distortion measure). Plot J over the iterations and as a function of K.

a) Modify your code to return J in each iteration i and plot the values against i.

b) Run k-means with different K's and plot the final values of J against K. Explain the graph.

c) If you wanted to choose the best K, how would you proceed?

Exercise 7.3: K-Means in Practice

Run k-means on the three datasets with different initializations and describe the algorithm's behavior. What are the problems and how would you solve them?