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 – Solutions –

Exercise 7.1: Clustering with k-means

The assignment and update step of k-means are implemented in `kMeansStep.m`.

You can run the k-means algorithm by executing the `RunKMeans` script. This script also performs the visualization of the centroids and their boundaries.

Exercise 7.2: Plotting the objective function

a) Plotting of the objective function is implemented in the `RunKMeans` script. The computation of the objective function can be found in `computeJ`.

b) Running k-means with different K values is implemented in `RunKMeansJev`. We can see from the graph that J decreases monotonically with growing K, which shows that a more complex model can fit the input data more closely.

c) For choosing the optimal K, just looking at the final cost value J is not sufficient. The reason for that is that J does not properly capture the complexity of the model. In general, a more complex model (i.e. larger number of clusters K) may reduce the overall error of the model, but may lead to overfitting and provide a bad abstraction from the underlying data. Instead, to properly choose K, it is better to use a measure such as BIC (Bayesian Information Criterion), which was shortly mentioned in the lecture in the context of X-Means. BIC also takes the model complexity into account, thus striving to find the optimal balance between low model complexity and low data fitting error.

Exercise 7.3: K-Means in practice

Even if the correct number of centroids K is known in advance, K-Means does not necessarily create an "optimal clustering" in the way a human would expect it (although the clustering should be optimal in terms of J). Reasons for this might be:

- Suboptimal initial centroid positions,
- The Euclidean distance measure (that is used in the calculation of J) being inadequate for the given dataset.