WS '16/'17 / 03-ME-712.07 / Prof. Dr. Frank Kirchner
Machine Learning for Autonomous Robots

Exercise sheet 4
due 04.12.2016

General remarks

The data sets used for this exercise sheets are found in the file "cluster_dataset2d.txt" and the file "cluster_dataset4d.txt" which can be downloaded from Stud.IP.

Problem 4.1 (k-Means) **(20 p.)**

a) Implement the k-means **algorithm** and write **tests** that apply the *dataset2d* data set on it.

b) Create a plot of the obtained clustering for $k = 35$.

Note: It may happen that in one iteration one of the clusters gets assigned no data points. In this case, reinitialize k-Means and start anew.

Problem 4.2 (C-Index) **(25 p.)**

The C-Index is a measure of the quality of a clustering. The following explains how to calculate the C-Index:

- Let S_{cl} be the sum of the distances between all pairs of points that belong to the **same clusters** (all clusters considered).
- Let N be defined as the number of distances used to calculate S_{cl} (number of **intra-cluster** point-pairs).
- Let D be the set of the distances between **all** point-pairs.
- Let S_{min} be the sum of the N smallest distances in D.
- Let S_{max} be the sum of the N largest distances in D.

Finally, the C-Index is defined as:

$$C = \frac{S_{\text{cl}} - S_{\text{min}}}{S_{\text{max}} - S_{\text{min}}}$$

The larger the intersection between the **set of intra-cluster pair-distances** and the **set** containing the N **smallest distances** among **all** pair-distances the smaller C will be (best case: $C = 0$).

a) For each $k = \{2, \dots, 9\}$: Run your k-means implementation 50 times with random initialization on the *dataset2d* data set.
Compute the C-Index for all runs and compute the **minimal and average** for each k. (5 p.)

b) Plot the minimal and average C-Index versus k. (5 p.)

c) How can the results be interpreted? What is a good value of k based on the values of the C index? (*Note:* The C values might become very small and indistinguishable in the plots. Please consider also the numerical values.) (10 p.)

d) Repeat a) and b) for the *dataset4d* data set. (5 p.)

On the hand-in date, **04.12.2016**, you must hand-in the following: [1]

a) a text file stating how much time you (all together) used to complete this exercise sheet
b) your solutions / answers / code

for problem 4.1 and 4.2.

[1] upload via StudIP (if there are problems with the upload contact me **beforehand**: krell@uni-bremen.de)