Programming Assignment 6

Write a program that implements the k-Means clustering with the following optimization criteria:

$$J(D, C) = \sum_{x_i \in D} min_{c_i \in C} \|c_i - x_i\|^2$$

where D is the set of data points and C the set of cluster centers, whereas $|C| = k$. (The optimization criteria can be seen as a kind of Sum of Squared Errors.) The goal is to minimize J, which basically minimizes the sum of intra-cluster distances.
Given are the two data sets[1] named *Example* and *Gauss* as tsv files from the last assignments. You are required to implement a k-Means with $k = 3$ and the initialization set to $c_1 = (0, 5)$, $c_2 = (0, 4)$ and $c_3 = (0, 3)$ for both data sets. Aside from that, use the k-Means algorithm from the lecture slides (slide 4, Clustering). Ignore the first column (class) of the data set.
The output of your program should then be *two* tsv files. One file contains the prototypes for each iteration (including initialization) and the other the optimization criteria given above. Each row then represents one iteration. The prototype values are separated by commas and the prototypes by tabs. The solution for the *Example* data set is given again, which consists of the files `Example-Proto.tsv` (containing the prototypes per iteration) and `Example-Progr.tsv` (containing the optimization criteria per iteration).
If the submitted program fails, the data format is incorrect or I have to change source code, in order to make it work, you will get zero points. Machine learning libraries are not allowed. You can use libraries for handling the CSV/TSV format and the input parameters.
Your program must accept *at least* the following parameters:

1. **data** - The location of the data file (e.g. /media/data/Example.tsv).
2. **output** - The directory, where the output tsv files should be written to. Name your output files `{Example,Gauss}-Progr.tsv` and `{Example,Gauss}-Proto.tsv`.

Please prepare example statements on how to use your program. E.g. for a python program:

```
python3 kMeans.py --data Example.tsv --output /media/data/output/
```

The final program code must be sent via email until Sunday, 20th of January 2019, 23:59 to marcus.thiel@ovgu.de. Please format your e-mail header as follows:

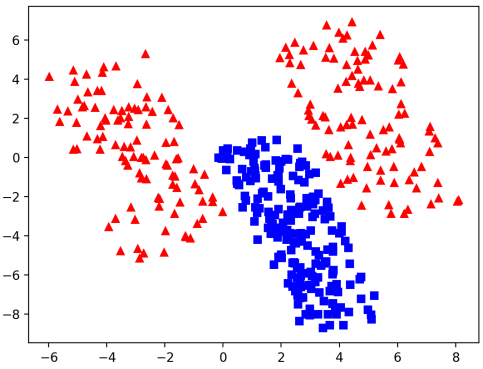
[Exercise Group] ML Programming Assignment 6

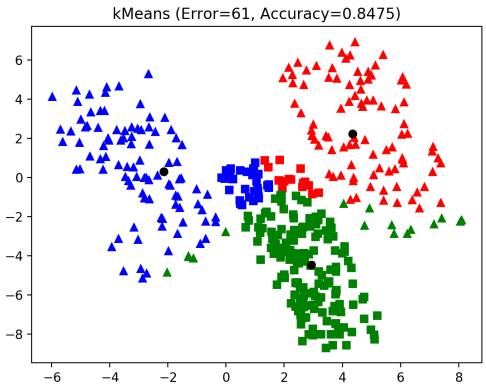
Replace *Exercise Group* with the day and time of your exercise group. E.g for Monday from 13:00 to 15:00 it would be:

[1] http://wwwiti.cs.uni-magdeburg.de/iti_dke/Lehre/Materialien/WS2018_2019/ML/res/kMeans.zip

[Monday 13-15] ML Programming Assignment 6

The figures below show the the *Example* data set and its clustering result for $k = 3$ with the given initialization. The black points are the prototypes. You can only gain one point per data set, if both output tsv files are correct.





2 points