Machine Learning
1st-Term Semester 2017-2018
Assignment 3

SYM

December 3, 2017

General instructions: All course participants are requested to handle their exercise solutions as follows:

- Write your report as PDF using one of the following text processing tools: MS-Word, LibreOffice, or Latex. Always mention your **name** and **student ID** in the PDF file. We use the report as the main basis for grading: All your results should be in the report. We also look at the code, but we won't however go fishing for results in your code.
- The code needs to be submitted as a runnable file or set of files (command to run it given in the report).
- In your report, the results will be mostly either in the form of a figure or program output. In both cases, add some sentences which explain what you are showing and why the results are the answer to the question.
- The deadline is Thursday 7.12.2017 at 21.00 UTC+7.
- Submit your work (only code and report) through email to the lecturer (**sym.milo.at.gmail.com**) before the deadlines. Late submission yields **penalty** by 10 points per hour.
- **All forms of cheating are strictly prohibited**.

Section 3: CLO3 (Totally points)

We have 10 datasets for exercise 1. Each data set has 3 columns where 1st and 2nd columns are attributes while 3rd column is the label. The datasets are (0) Aggregation[1], (1) Compound[2], (2) D31[3], (3) Flame[4], (4) Heart-1[5], (5) Heart-2[6], (6) Jain[7], (7) Pathbased[8], (8) R15[9], and (9)Spiral[10].

Use your student ID to select the datasets as follows: let t is **the last digit of your student ID**, then use the data set (t) for the exercise.

[1] A. Gionis, H. Mannila, and P. Tsaparas, Clustering aggregation. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007. 1(1): p. 1-30.

[2] C.T. Zahn, Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Transactions on Computers, 1971. 100(1): p. 68-86.

[3] C.J. Veenman, M.J.T. Reinders, and E. Backer, A maximum variance cluster algorithm. IEEE Trans. Pattern Analysis and Machine Intelligence 2002. 24(9): p. 1273-1280.

[4] L. Fu and E. Medico, FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. BMC bioinformatics, 2007. 8(1): p. 3.

[5] M.S. Mubarok, Machine Learning Course at Informatics Faculty Telkom University, 2017

[6] M.S. Mubarok, Machine Learning Course at Informatics Faculty Telkom University, 2017

[7] A. Jain and M. Law, Data clustering: A user's dilemma. Lecture Notes in Computer Science, 2005. 3776: p. 1-10.

[8] H. Chang and D.Y. Yeung, Robust path-based spectral clustering. Pattern Recognition, 2008. 41(1): p. 191-203.

[9] C.J. Veenman, M.J.T. Reinders, and E. Backer, A maximum variance cluster algorithm. IEEE Trans. Pattern Analysis and Machine Intelligence, 2002. 24(9): p. 1273-1280.

[10] H. Chang and D.Y. Yeung, Robust path-based spectral clustering. Pattern Recognition, 2008. 41(1): p. 191-203.

1. In this exercise, we implement partitional clustering method: K-means algorithm.

   (a) Load the selected data set. Visualize all data points using scatter plot in one color (**no** need to give different color for different label). Use only attribute 1 as x -axis and attribute 2 as y -axis. This visualization might give you a brief thought of existed clusters. [**5 points**]

   (b) Again, visualize all data points using scatter plot. Use attribute 1 as x -axis, attribute 2 as y -axis. Use different color and/or different symbol for each label. This visualization shows you the real existed clusters. [**5 points**]

   (c) Apply K-means on the selected data set. Your codes have to clearly contain

      i. Function that takes as inputs: the data matrix and initial centroids; as outputs: the final centroids and the cluster assignments specifying which data vectors are assigned to which centroids after convergence of the algorithm. (To speed up the algorithm sufficiently, use matrix operations wherever possible and avoiding explicit loops). [**15 points**]

      ii. Function to calculate the objective function of K-means that is Sum of Squared Errors (SSE). It takes as inputs: all data vectors and final centroids resulted from learning. [**10 points**]

   (d) Run your K-means algorithm, using K equals to the number of different label in dataset. The initial centroids taken from randomly selected K data points. After convergence, visualize the centroid of each cluster as well as all data points assigned to that cluster (it should be easily distinguished between the centroids and the data points, also give different color/symbol to different cluster). One may run the algorithm several times in order to obtain the best result (hints: use the SSE as the measure). [**10 points**]

   (e) Based on visualization resulted from point 1(d), to what extent do the K clusters correspond to the K different labels? (Hints: Use the visualization from point 1(b) to compare and get a view/thought of clustering results shown by point 1(d).) [**5 points**]

   (f) Re-run K-means but selecting randomly one instance of each label as the initial centroid (so that the initial centroids all represent distinct label). After convergence, visualize the centroid of each cluster as well as all data points assigned to that cluster (it should be easily distinguished between the centroids and the data points, also give different color/symbol to different cluster). One may run the algorithm several times in order to obtain the best result (hints: use the SSE as the measure). [**10 points**]

   (g) Based on visualization resulted from point 1(f), to what extent do the K clusters correspond to the K different labels? (Hints: Use the visualization from point 1(b) to compare and get a view/thought of clustering results shown by point 1(f).) [**5 points**]

   (h) By visually comparing figures created from point 1(d) and 1(f), what do you think of the clustering results? Give explanations. [**5 points**]