Computer Exercise 6

Nov. 18, 2021

Task:
Experiment with K-means clustering on synthetic toy data.

Goal:
Find some program to generate synthetic data of multiple Gaussian distributions in 2D space. Write your own program or find some publicly available program for K-means clustering. Adjust the parameters in data generation to simulate different situations of clusters, and observe the performances of K-means clustering in different situations.

Data:
Generated synthetic data.

Experiment 12 (K-means Clustering):

1) Find a program that can generate synthetic data with a Gaussian distribution in the 2D space. Generate several synthetic datasets for this experiment. The datasets should include at least the following situations:
   a) Data of 3 or more clearly separated Gaussian distributions, with means well apart from each other and almost no overlap;
   b) Data of 3 or more Gaussian distributions with heavy overlap.
   c) Data of only one Gaussian distribution with nonequal variances in the 2 dimensions and non-zero covariance. You may generate multiple datasets with different settings of the covariance matrix.

   For each situation, generate two or more versions of the datasets with different sample sizes.
2) Drew the scree-plots for each dataset to see whether the correct number of clusters can be found on the scree-plots.
3) Do K-means clustering on each dataset and observe the result. You may try different initializations to observe their possible influences in the clustering processes and the results.

Experiment Report:

- Write an experiment report to describe and analyze the experiment observations. The report should also include the short essay on parameter choices.
- Provide detailed supplementary materials that should include at least the following:
  - A readme file containing information on all supplementary files, programming environment and parameters used in the experiments (if any)
  - Source codes (should let TAs to be able to run the code and reproduce your experiments)
  - All synthetic data files and experiment result files

Due date: Dec. 1 (Wednesday) 23:00 Beijing time