Computer Exercise 7

Nov. 25, 2021

Task:

1. Dimension reduction and clustering on data from the 10 categories of handwritten digits (0~9).
2. (Optional) Dimension reduction, visualization and clustering on the ICU data.

Goal:

Exercise on the comprehensive use of dimensionality reduction, visualization and clustering methods on data with homogeneous features and with heterogeneous features. Use publicly available packages (e.g., scikit-learn) when possible.

Data:

For task 1:

Folder	Sample size	Feature data file	Class label file
Data2forEx	60,000	train-images-idx3-ubyte.gz	train-labels-idx1-ubyte.gz

For task 2:

You may use either the data with the 108 features in TrainingSet-1 (5000 samples) or in TestSet-1 (1097 samples) or both (6097 samples) for the exercise. Do not use any class label information.

Experiment 13 (Dimensionality reduction and clustering of digits):

1) Select one category of handwritten digits ("0", "1", "2", …, or "9") in the file as experiment data.
2) Perform PCA on the data. Choose a proper number of PCs as the features for clustering.
3) Perform clustering analyses on the data with K-means, hierarchical clustering and GMM. Choose the proper number of clusters for each method.
4) Visually check the handwritten digits within each cluster and between different clusters. Discuss whether the clustering results are reasonable or reflect any interesting information or subtypes in the data.
5) Repeat above experiment on another digit category and check whether you can get similar results.

Optional: You are encouraged to go beyond the above-mentioned experiments. Be creative and explore whether you can make new discoveries on the dataset.

(Optional) Experiment 14 (Dimensionality reduction and clustering of ICU data):

1) Study the nature of the features and design your own way of doing dimensionality reduction on these features based on what we have learned in this course. (Hint: One possibility is to define a distance measure among the samples using certain weighted sum of differences of all features. Then you can compute the distance matrix of all samples, and use MDS to represent the samples in a low dimensional space for further analysis or for visualization. You may also work out other possible solutions.)
2) Apply clustering analysis on the obtained low-dimensional representation of the samples and see whether/what clusters can be obtained. Try to explain the clusters if you obtained more than one.
3) Visualize the samples and the clustering results in a 2D plot. Discussion your observations on the data and on the methods in the 3 steps.

(Note on the optional exercise: It is a good training for yourself so do not work for the credit. But if, just in case, you do not perform well in the final, extra bonus may be added for the optional exercise if completed with high quality.)

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 experiment result files

Due date: Dec. 8 (Wednesday) 23:00 Beijing time