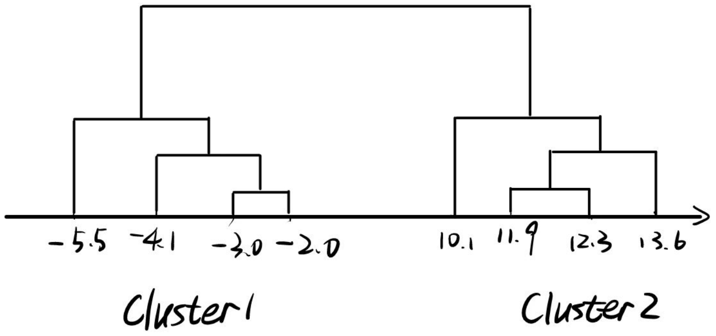
Pattern Recognition and Machine Learning: Homework 2

Qingru Hu   2020012996

2023 年 4 月 12 日

Problem 1

采用 Agglomerative (Bottom-up) Clustering



I want to separate the data points into 2 clusters, because they have small scatter in each cluster and big scatter between them.

Problem 2

I used the `sklearn.mixture.GaussianMixture` model to fit the data. This model uses EM algorithm to estimate the distribution. The code for this problem is in `problem2.py`. The diagram for the prediction error with respect to iteration times t and total number of samples N are shown in Fig.1 and Fig.2.

Problem 3

The code fo this problem is in `problem3.py`.

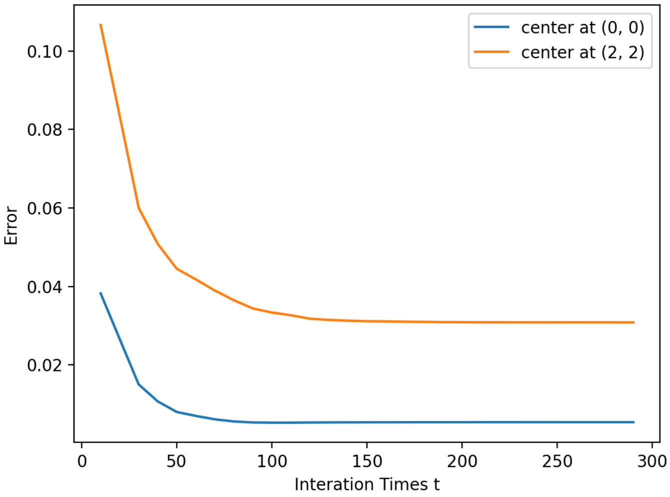


图 1: The prediction error with respect to t

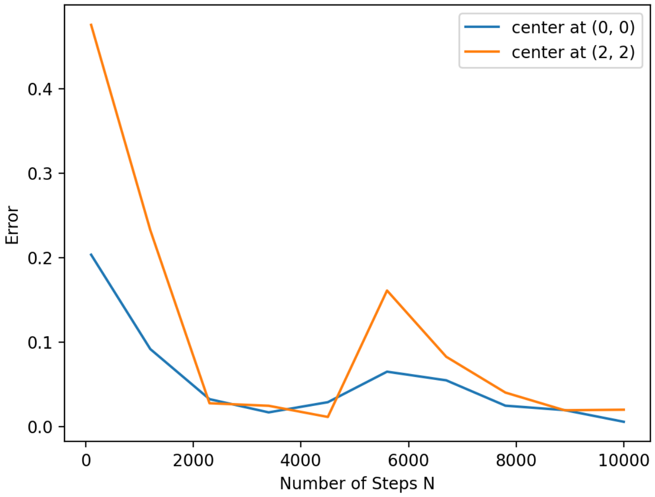


图 2: The prediction error with respect to N

3.1

I used the `sklearn.cluster.KMeans` model to do the clustering on training data points. The diagram of Je with respect to number of clusters is shown in Fig.3. The elbow point is around 15.

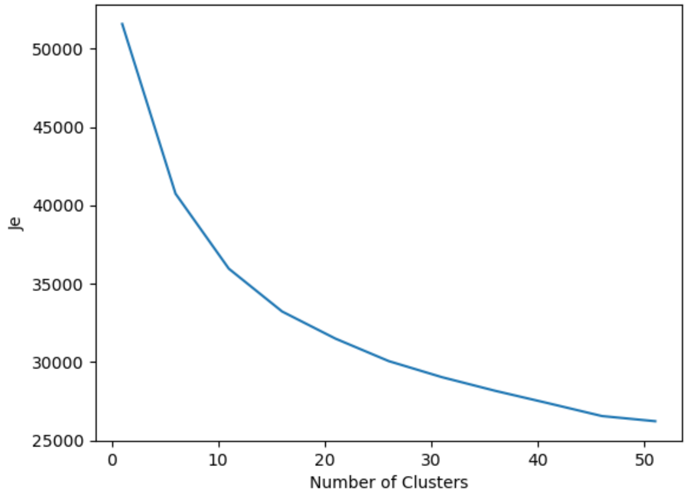


图 3: Je with respect to number of clusters

3.2

The learned means of each cluster are stored in ‘kmeans’ directory. They look like ten hand-written numbers. I used 20 clusters in the prediction and the prediction accuracy on MNIST test dataset of `KMeans` is about 60%.

3.3

I used `sklearn.mixture.GaussianMixture` to use EM on MNIST. The learned means of each cluster are stored in ‘em’ directory. I used 20 components in the prediction and the prediction accuracy on MNIST test dataset of `KMeans` is a little higher than 60%.

Acknowledgement: Thank Jingyuan Zhao (赵泾源) in this class for the discussion about Problem 2 and Problem 3.