Pattern Recognition Experiment 3

EECS26 111060024 蔡孟伶

一張含有 文字, 螢幕擷取畫面, 字型, 文件 的圖片

AI 產生的內容可能不正確。

(a)

|  |  |  |
| --- | --- | --- |
| 一張含有 螢幕擷取畫面, 鮮豔, 紫色, 圖形 的圖片  AI 產生的內容可能不正確。 | 一張含有 文字, 螢幕擷取畫面, 鮮豔, 圖形 的圖片  AI 產生的內容可能不正確。 | 一張含有 文字, 螢幕擷取畫面, 鮮豔, 圖表 的圖片  AI 產生的內容可能不正確。 |
|  | 一張含有 文字, 螢幕擷取畫面, 字型, 鮮豔 的圖片  AI 產生的內容可能不正確。 | 一張含有 文字, 螢幕擷取畫面, 字型, 圖表 的圖片  AI 產生的內容可能不正確。 |

(b) Result image

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

AI 產生的內容可能不正確。 一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

AI 產生的內容可能不正確。一張含有 文字, 螢幕擷取畫面, 字型 的圖片

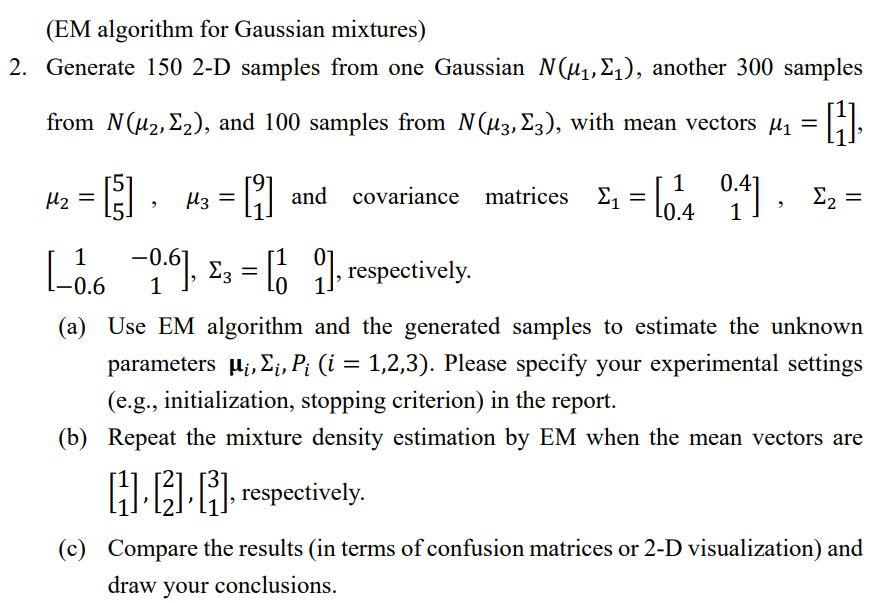
AI 產生的內容可能不正確。 一張含有 文字, 螢幕擷取畫面, 字型 的圖片

AI 產生的內容可能不正確。

(c) Conclusion:

From the two cases tested, we can see that the average construction error of dimension l decreased intensely as l increased. The formula of average construction error is

This reconstruction error is the average of the whole dataset’s data, with each summing up the pixel-wise difference’s square to the reconstructed image. The relationship between dimension l and reconstruction error is reasonable since we start from the eigenvector corresponding to the biggest eigenvector, as more feature involved, the output should become closer the original input.



|  |  |
| --- | --- |
| 一張含有 文字, 螢幕擷取畫面, 圖表 的圖片  AI 產生的內容可能不正確。 | 一張含有 文字, 螢幕擷取畫面, 圖表 的圖片  AI 產生的內容可能不正確。 |
| 一張含有 文字, 螢幕擷取畫面, 圖表 的圖片  AI 產生的內容可能不正確。 | 一張含有 文字, 圖表, 螢幕擷取畫面 的圖片  AI 產生的內容可能不正確。 |
| 一張含有 文字, 螢幕擷取畫面, 圖表 的圖片  AI 產生的內容可能不正確。 | 一張含有 文字, 螢幕擷取畫面, 圖表 的圖片  AI 產生的內容可能不正確。 |

1. Initialization and stopping criterion setting for the EM algorithm:

Three parameters set at initialization of EM are the mean, covariance matrix and the Prior probability. At first, mean is determined by randomly sampling three(since we want three groups) data from the dataset. Covariance matrix is initialized as 2 \* 2 identity matrix for each class, and the prior is initialized as 1 / K, which is equal prior.

There are two stopping criterions for the EM algorithm. First, a maximum of iteration = 100 is set, and second, when the new parameters are updated at M-step is close to the old parameters. The threshold is set as 10^-4, when the total of np.linalg.norm(X) of mean, covariance and prior is small, the loop stops early. The norm operation is shown below.

(c) Conclusion:

From the three test run shown above, we can see that although the initialization of mean may not be very nice or well-separated, EM could still find the cluster of the data. However, it only clusters data into group but didn’t guarantee the correctness of classification class. Though the data that belongs to class 1 at ground truth is grouped together, it may be classified as group 2 depending on the initialization sample data. For dataset (b), we can see the following result. The clustering result is not as clear as dataset (a), since there are many overlapping data that are hard to be separate by using usual linear methods. EM intends to cluster data points that are close to one group, making it harder to deal with overlapping data.

|  |  |
| --- | --- |
| 一張含有 螢幕擷取畫面, 文字, 圖表, 鮮豔 的圖片  AI 產生的內容可能不正確。 | 一張含有 文字, 螢幕擷取畫面, 圖表 的圖片  AI 產生的內容可能不正確。 |
| 一張含有 文字, 螢幕擷取畫面, 圖表, 鮮豔 的圖片  AI 產生的內容可能不正確。 | 一張含有 文字, 螢幕擷取畫面, 圖表, 鮮豔 的圖片  AI 產生的內容可能不正確。 |
| 一張含有 文字, 螢幕擷取畫面, 圖表, 鮮豔 的圖片  AI 產生的內容可能不正確。 | 一張含有 文字, 螢幕擷取畫面, 圖表 的圖片  AI 產生的內容可能不正確。 |

一張含有 文字, 字型, 螢幕擷取畫面, 行 的圖片

AI 產生的內容可能不正確。

Setting the random seed in shared\_dataset.py, we will be able to generate same dataset even executing the file separately. Below will show the result of problem 2: EM algorithm along with problem 3: K-means. Two dataset are on two columns, and the result of each row is in order of ground\_truth, EM\_algorithm, K-means(K = 2), K-means(K = 3), and K-means(K = 4).

Conclusion:

From the result, I think the result of EM algorithm and K-means(K=3) under dataset 1 are both acceptable, while for dataset 2, K-means(K=3) seems to cluster more data from the same group in ground truth than EM does in this case. However, we should still seek for better representation data or else such clustering doesn’t really fit the dataset. It is also a problem worth trying of what K to decide, since in real world, we don’t have the ground truth label. Another observation on the K-means result is that it didn’t take into account of the prior of each class, from the K-means result(K = 2, K= 3, K=4), I see that each class has similar portion. However, inside EM algorithm, we take into account prior variable and update it each time, therefore for some dataset with larger prior differences among classes, I think EM algorithm might perform better on each cluster’s portion compared with K-means. Last, since the principal of clustering inside K-means is about distance, it is not likely to have results like what we have in EM algorithm for dataset(b). Some red data are separated from other red ones by blue data. When facing some distribution to have one class caught the others in the middle, K-means might not be appropriate for that.

|  |  |
| --- | --- |
| 一張含有 文字, 螢幕擷取畫面, 圖表 的圖片  AI 產生的內容可能不正確。 | 一張含有 螢幕擷取畫面, 文字, 圖表, 鮮豔 的圖片  AI 產生的內容可能不正確。 |
| 一張含有 文字, 螢幕擷取畫面, 圖表 的圖片  AI 產生的內容可能不正確。 | 一張含有 文字, 螢幕擷取畫面, 圖表, 鮮豔 的圖片  AI 產生的內容可能不正確。 |
| 一張含有 文字, 螢幕擷取畫面, 圖表 的圖片  AI 產生的內容可能不正確。 | 一張含有 文字, 螢幕擷取畫面, 圖表 的圖片  AI 產生的內容可能不正確。 |
| 一張含有 文字, 螢幕擷取畫面, 圖表 的圖片  AI 產生的內容可能不正確。 | 一張含有 文字, 螢幕擷取畫面, 圖表, 鮮豔 的圖片  AI 產生的內容可能不正確。 |
| 一張含有 文字, 螢幕擷取畫面, 圖表 的圖片  AI 產生的內容可能不正確。 | 一張含有 文字, 螢幕擷取畫面, 圖表, 鮮豔 的圖片  AI 產生的內容可能不正確。 |