# CS5487 Machine Learning - Programming Assignment 2

Zhanghan Ke (55460880)

zhanghake2-c@my.cityu.edu.hk

**Abstract.** This is the report for my Programming Assignment 2 of CS5487 Machine Learning. The code also available in my github. Link: https://github.com/ZHKKKe/PA2.git.

## 1 Clustering Synthetic Data

### 1.1 Problem (a)

I implemented K-Mean and Mean-Shift algorithms, and the EM-GMM algorithm of my assignment from the *scikit learn* package.

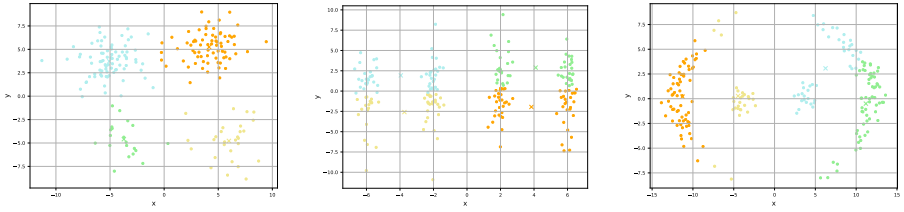
### 1.2 Problem (b)

The Fig. 1 shows my results of this problem. The results shows the Mean-Shift with careful selection of bandwidth $h$ could clustering data best. And the EM-GMM algorithm also work pretty good. The K-Mean gains a bad result in the *dataB_X*.

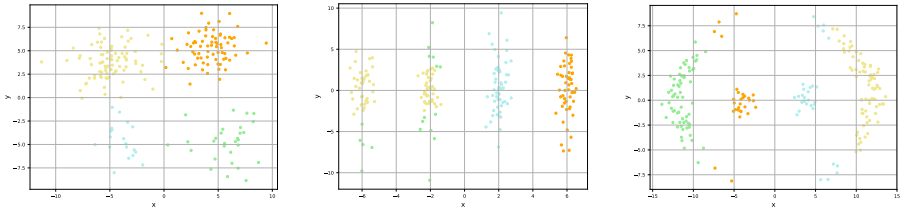
K-Mean assigns each sample to a particular cluster on convergence. It works better in the high dimensional data nad it is easy to interpretation and implementation. However, K-Mean might cause some samples mis-matching to any cluster. EM-GMM assigns each sample to clusters softly by a probability, it works well with non-linear geometric distribution as well. But since the all components must be accessed, it is difficult in the high dimensions. Mean-Shift has a hyperparameter bandwidth $h$, which could control the clustering results, it could be set case by case to gain great results. But this algorithm is very sensitive to $h$, so we need to choose it carefully.

### 1.3 Problem (c)

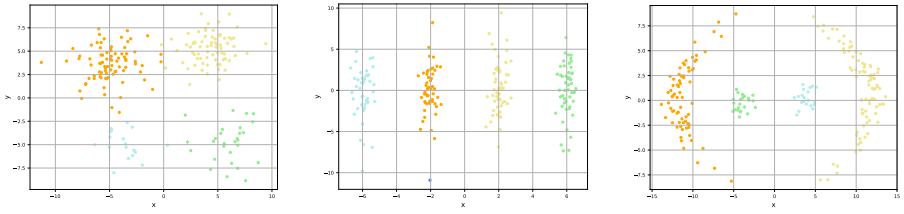
Mean-Shift is sensitive for the bandwidth, changing it will cause bad resutls. The Fig. 2 shows the result of *dataC_X* under different $h$.



(a) K-Mean - dataA_X (b) K-Mean - dataB_X (c) K-Mean - dataC_X

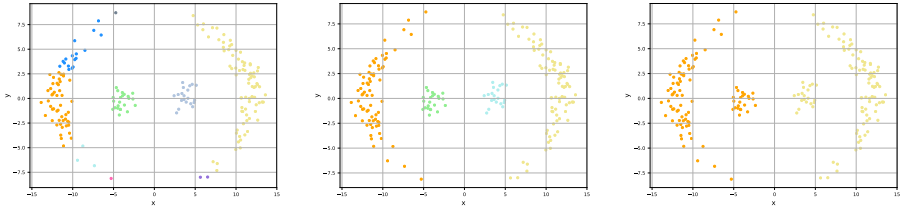


(d) EM-GMM - dataA_X (e) EM-GMM - dataB_X (f) EM-GMM - dataC_X



(g) Mean-Shift - dataA_X (h) Mean-Shift - dataB_X (i) Mean-Shift - dataC_X

**Fig. 1.** Result of clustering synthetic data by three algorithm.



(a) bandwidth=2.0 (b) bandwidth=2.5 (c) bandwidth=3.0

**Fig. 2.** Result of clustering synthetic data of Mean-Shift under different bandwidth. (a) Too small bandwidth will generate too many clusters. (b) Suitable bandwidth will generate matching clusters. (c) Too large bandwidth will generate too less clusters.

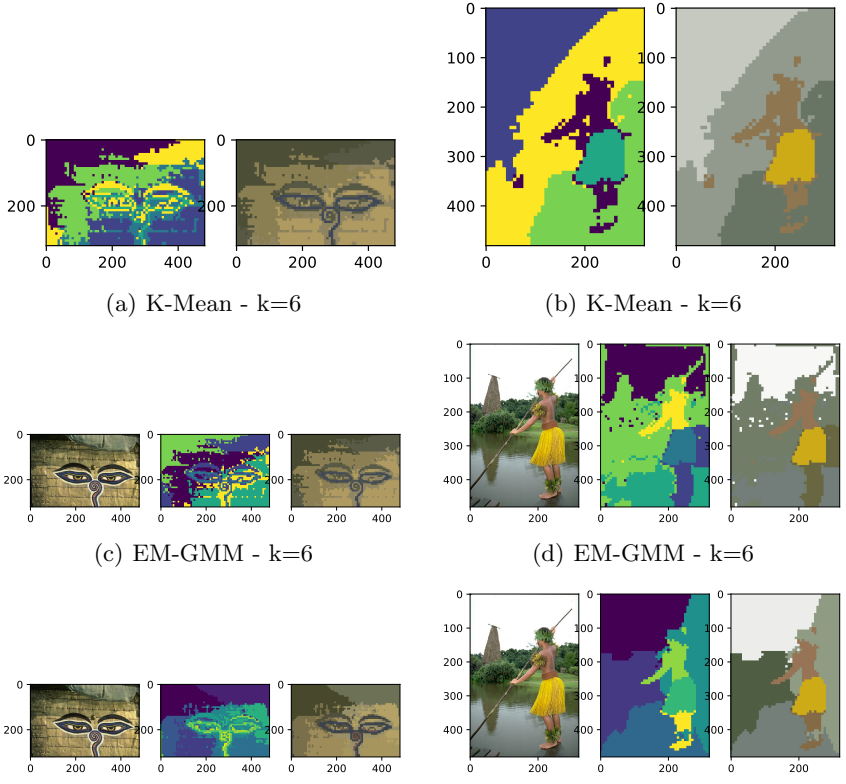
# 2 A Real World Clustering Problem Image Segmentation

## 2.1 Problem (a)

The Fig. 3 shows the results in 2 different images by three algorithm.

## 2.2 Problem (b)

I try to use different setting of the distance and kernel and find these parts need to be setting carefully, otherwise the results will be bad.



(a) K-Mean - k=6 (b) K-Mean - k=6

(c) EM-GMM - k=6 (d) EM-GMM - k=6

(e) Mean-Shift - h=1 (f) Mean-Shift - h=1

**Fig. 3.** Result of image segmentation by three algorithm.