FE540 – Fall 2016 Programming Assignment
K-means Clustering Algorithm

1. Review: Expectation-Maximization

When complete data log likelihood is not available, we use expected complete data log likelihood (a.k.a auxiliary function)

$$Q(\theta, \theta^{t-1}) \equiv E[l_c(\theta)|D, \theta^{t-1}]$$

when θ is the parameter of the model and $l_c(\theta)$ is the log likelihood with complete data. D is the data.

Expectation Maximization is consisted of two steps.

- E-step: compute $Q(\theta, \theta^{t-1})$
- M-step: find $\theta^t = \arg\max_\theta Q(\theta, \theta^{t-1})$

As we learned from the lecture, the expected complete data log likelihood is represented as follow,

$$Q(\theta, \theta^{t-1}) = \sum_i \sum_k r_{ik} \log \pi_k + \sum_i \sum_k r_{ik} logp(x_i|\theta_k)$$

where $r_ik \equiv p(z_i = k|x_i, \theta^{t-1})$ is the responsibility of the k-th cluster for x_i.

- E-step: compute $r_{ik} = \frac{1}{Z}\pi_k p(x_i|\theta_k^{t-1})$
- M-step: compute followings,

$$max_{\pi_k} \sum_i \sum_k r_{ik} \log \pi_k$$

$$max_{\mu_k, \sum_k} \sum_i \sum_k r_{ik} \log p(x_i, \theta_k)$$

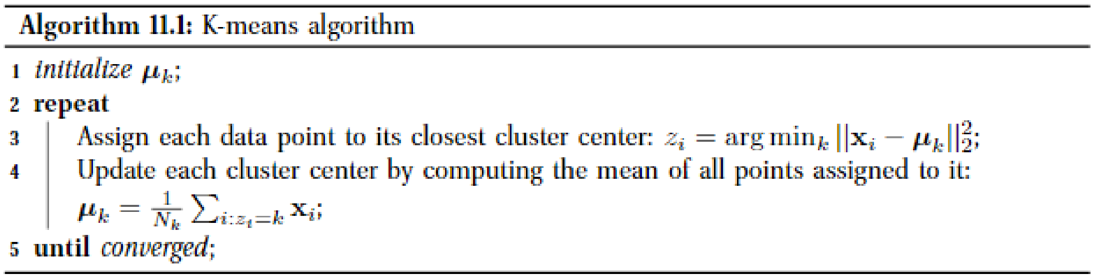
2. K-means clustering

K-means clustering can be thought as a Gaussian Mixture model (GMM) with the following assumptions: $\sum_k = \sigma^2 I_D$ and $\pi_k = 1/K$ are fixed when k is the number of elements in the kth cluster.

The EM update is as follows,

- E-step: compute r_{ik} with following equation.
  $p(z_i = k|x_i, \theta) \approx \mathbb{1}(k = z_i^*)$ where $z_i^* = \arg\max_k p(z_i = k|x_i, \theta)$. In another word, z_i^* can be computed as following
  $$z_I^* = \arg\min_k \|x_i - \mu_k\|_2^2$$
- M-step: compute the new mean of the cluster,
  $$\mu_k = \frac{1}{N_k} \sum_{i:Z_i=k} x_i$$

See the following figure for the pseudo-code.



Algorithm 11.1: K-means algorithm

```
initialize $\mu_k$;
repeat
    Assign each data point to its closest cluster center: $z_i = \arg\min_k ||\mathbf{x}_i - \boldsymbol{\mu}_k||_2^2$;
    Update each cluster center by computing the mean of all points assigned to it:
    $\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i:z_i=k} \mathbf{x}_i$;
until converged;
```

3. Source code template

The part of k-means clustering is written. That is a part of source code is intentionally removed. Please fill the part of codes (E-step and M-step)

Please refer to the source codes provided, 'EM_Kmeans_templet.py' 'EM_Kmeans_templete.ipynb'. Complete the code and submit. Please submit your source code as one zip file named 'hw3_(your student id).zip' which includes your python code file, screen captures and two figures (original sample and final samples after 10 steps of K-means clustering)

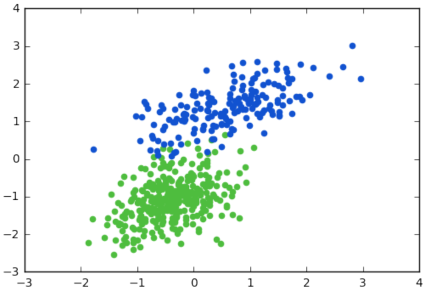


Figure 1: input

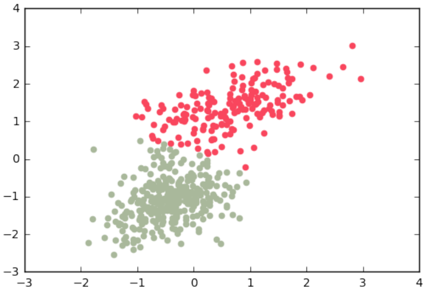


Figure 2: output

4. Bonus question

If you can compute pseud-F statistic with different number of cluster, k. You will receive a small but bonus points. In this case, please provide the chart where x axis is the number of clusters and y axis is the pseudo F statistics.