1. Problem

① Select two datasets that contain examples from multiple classes
② Implement K-means algorithm and project the data onto 2D
③ Implement EM algorithm algorithm and project the data onto 2D
④ Design a method that can determine the number of cluster
⑤ Using a datasets with labels to test error of K-means and EM

2. Proposed Solution

K-Means

① Start with initial guess of cluster center

u_j (cluster j's center u)

② Assign each example to one cluster

$$\hat{y} = \arg\min_{j} | x^{(i)} - u_j |$$

③ Recompute cluster centers by the function:

$$u_j = \frac{1}{m_j} \sum x^{(i)} 1(\hat{y}^{(i)} = j)$$

Recompute ② and ③

Expectation Maximication

M-Step

$$P_l(x^{(i)};\theta_l^g) = \frac{1}{2\pi |\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x^{(i)} - u_l^g)^T |\Sigma|^{-1} (x^{(i)} - u_l^g))$$

$$P(l \,|\, x^{(i)};\theta^g) = \frac{\alpha_l^g p_l(x^{(i)};\theta_l^g)}{\sum_{j=1}^{k} \alpha_j^g P_j(x^{(i)};\theta_l^g)}$$

E-Step

$$\alpha_l^{new} = \frac{1}{m} \sum_{i=1}^{m} P(l \,|\, x^{(i)};\theta^g)$$

$$u_l^{new} = \frac{1}{m\alpha_l^{new}} \sum_{i=1}^{m} P(l \,|\, x^{(i)};\theta^g) x^{(i)}$$

$$\Sigma_l^{new} = \frac{1}{m\alpha_l^{new}} \sum_{i=1}^{m} p(l \,|\, x^{(i)};\theta^g)(x^{(i)} - u_l^{new})(x^{(i)} - u_l^{new})^T$$

Repeat M-Step and E-step until

$$\| tr(\Sigma_l^g - \Sigma_l^{new}) \| > \tau$$

Final get label by assignment:

$$\hat{y}^{(i)} = \arg\max_{l} P(l \mid x^{(i)}; \theta^{g})$$

3. The method for automatically determine number of cluster:

① Initialize Cluster number K

② Do k-means or EM method to get cluster.

③ Keep repeating M-step and E-step ,when the cluster's number changed, save the repeat number r1

④ Keep repeating M-step and E-step. After 2*r1 times if the number of the clusters don't change, then we stop . Currently, we get the number of the clusters. If the number of the clusters changed in 2*r1 times, update repeat number and keep doing step ③.

4. Implementation

① In Python, created a class "clusterInfo" which contains the points information and the centers of clusters

② Read data from ".csv" and ".data" file

③ Depends on the formula in ②, we initial the data for M-step

④ We do the E-step by formula in ②

⑤ Repeat ③,④

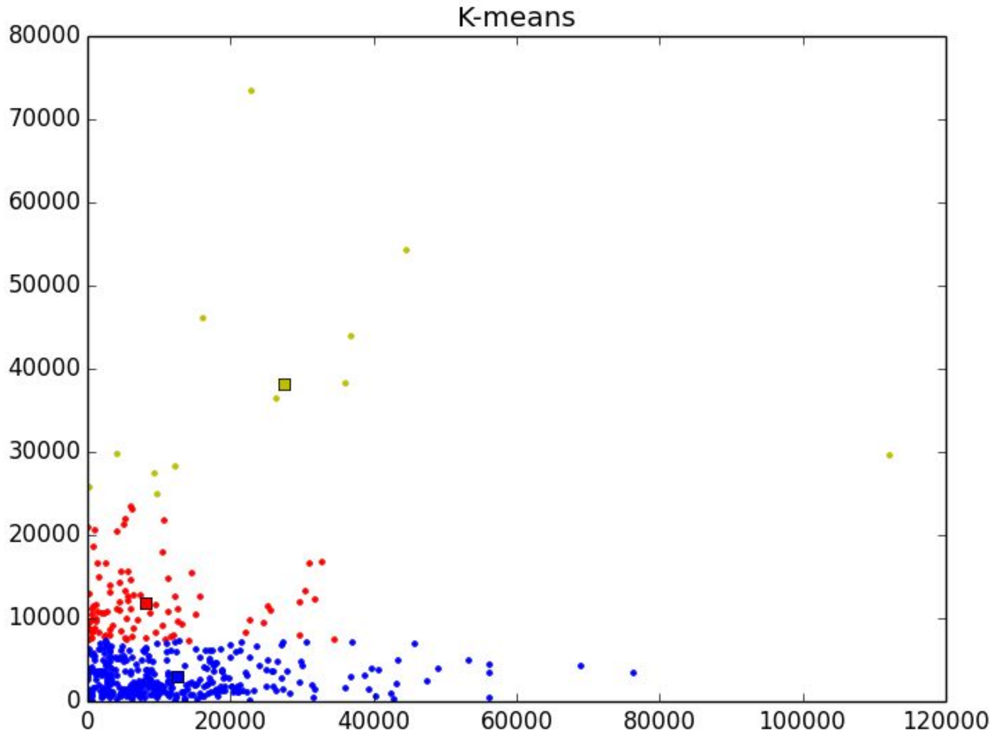
⑥ Read data from "knowdata.csv" which contains labeled data.

⑦ Using K-means method to get the cluster and compute accuracy.

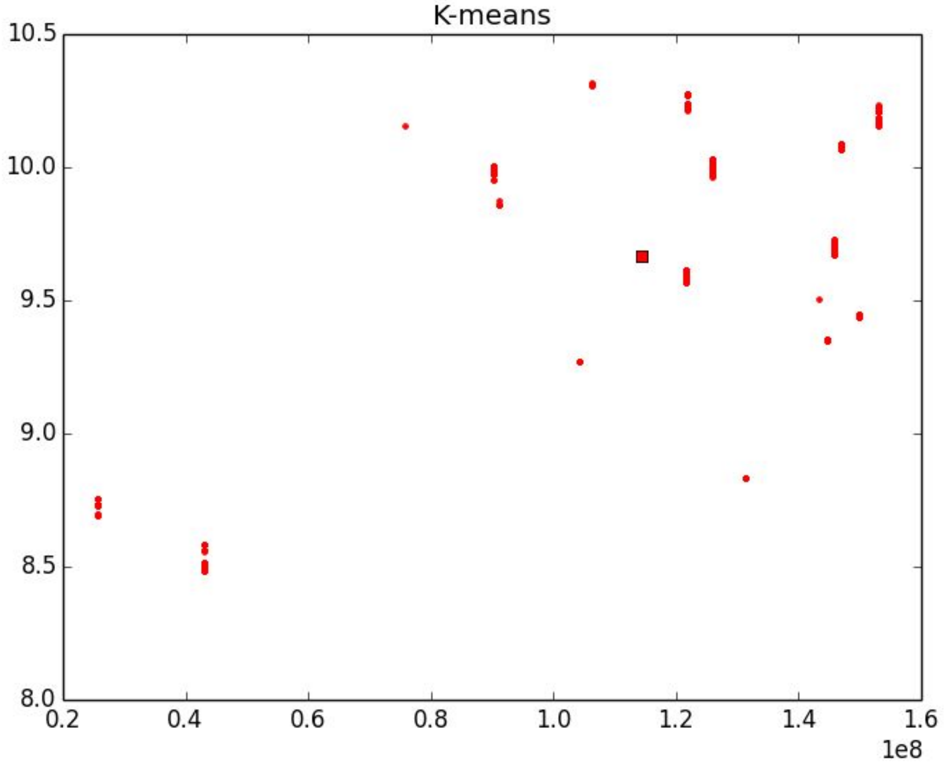
⑧ Using EM method to get the cluster and compute accuracy.

5. Result and discussion

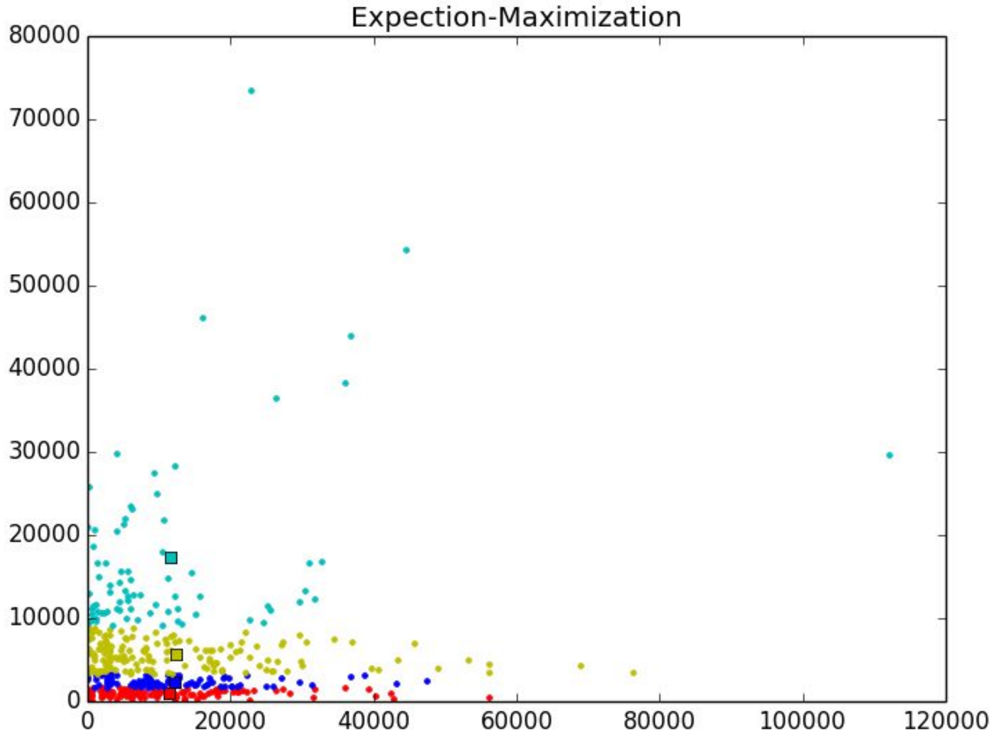
Using K-means for "customersdata.csv"(square is center, point is data)



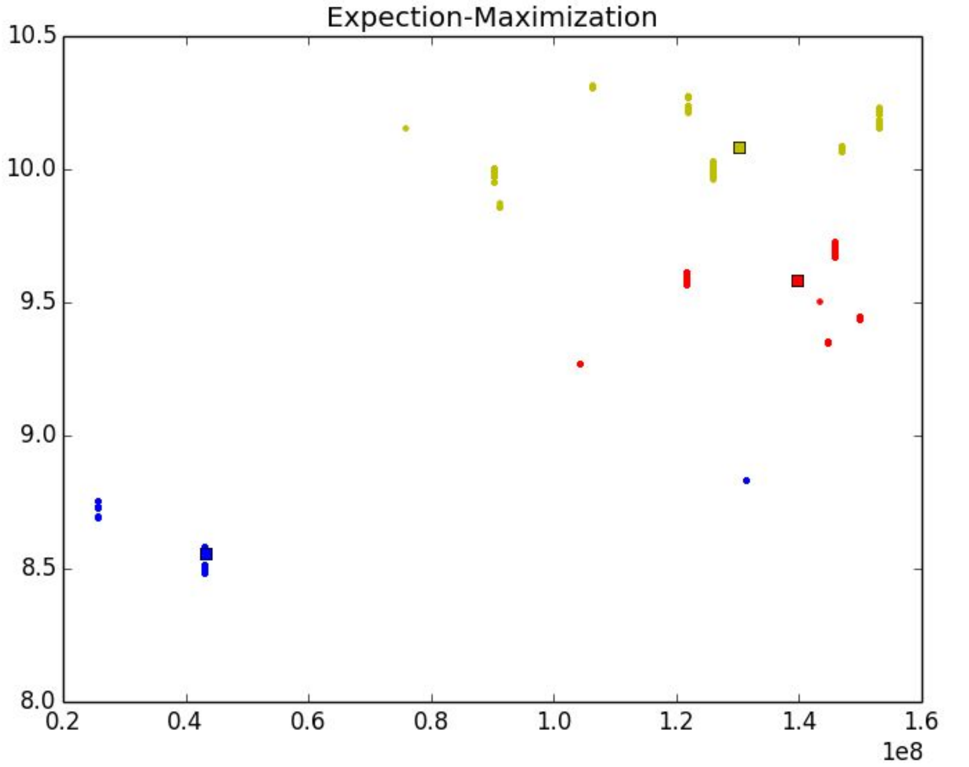
Using k-means for "D_spatial_networ.csv"



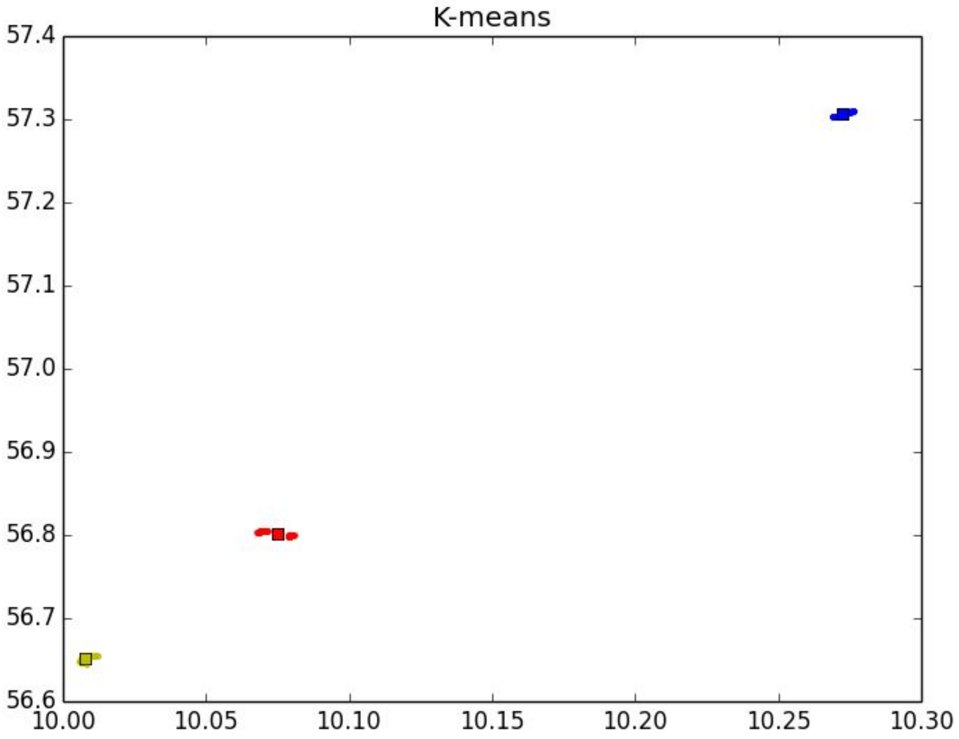
Using EM for "customersdata.csv"



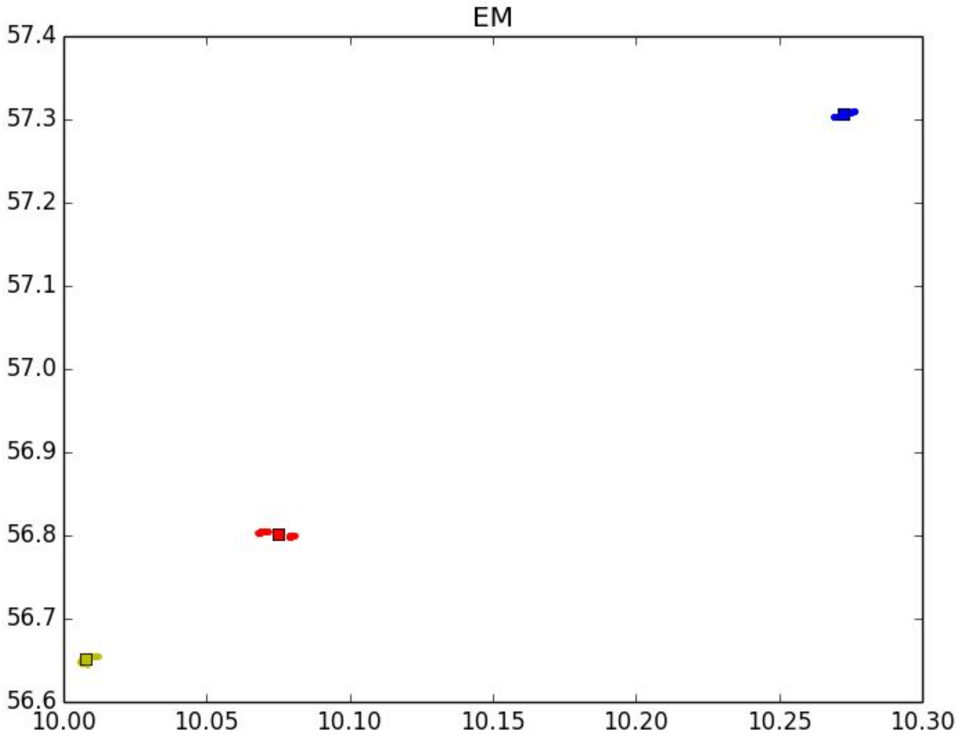
Using EM for “D_spatial_network”



We have labeled data “knowdata.csv”
Using K-means ,the accuracy is :0.807692307692



Using EM accuracy is :0.807692307692


EM