Pattern Recognition

組員 :蔡龍佑 (M10517035)

蕭任均(M10517029)

董屹晨 (M10517028)

Question

Please discuss the similarities and differences between EM. K-means and Bayesian classification?

please use examples or experiments to support your statement.

Types of problems and tasks

- Supervised learning
  - Classification
    - In which the aim is to assign each input vector to one of a finite number of discrete categories.
  - Regression
    - If the desired output consists of one or more continuous variables.
- Unsupervised learning
  - Clustering
    - To discover groups of similar examples within the data.
    - Determine the distribution of data within the input space
    - Known as ***density estimation***
    - Project the data from a high-dimensional space down to two or three dimensions for the purpose of ***Visualization***.
- Reinforcement learning
    - Concerned with the problem of finding suitable actions to take in a given situation in order to maximize a reward.

What is this task?

Thirty people

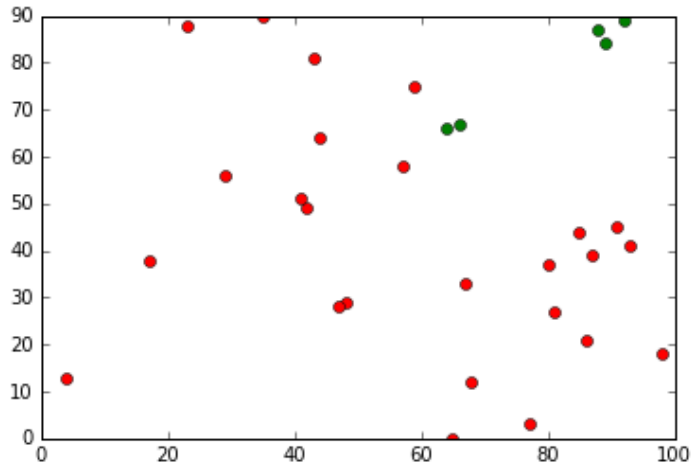
Mathematics		English	
66	67	67	33
86	92	21	89
77	85	3	44
64	42	66	49
59	57	75	58
88	47	87	28
98	89	18	84
80	35	37	90
81	29	27	56
93	87	41	39
23	91	88	45
48	44	29	64
68	43	12	81
17	41	38	51
4	65	13	0

Pass or not

Thirty people

```
plt.plot(getcol(_pass,0),getcol(_pass,1),'ro',c='g')
plt.plot(getcol(_non_pass,0),getcol(_non_pass,1),'ro',c='r')

plt.show()
```

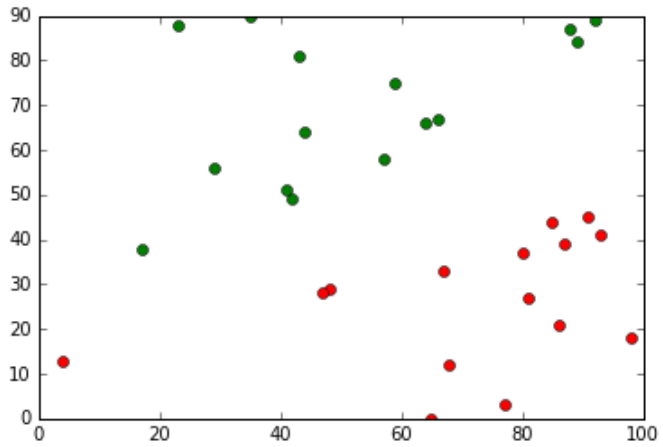


Two clusters

Thirty people

```
plt.plot(getcol(cluster_1,0),getcol(cluster_1,1),'ro',c='g')
plt.plot(getcol(cluster_2,0),getcol(cluster_2,1),'ro',c='r')

plt.show()
```

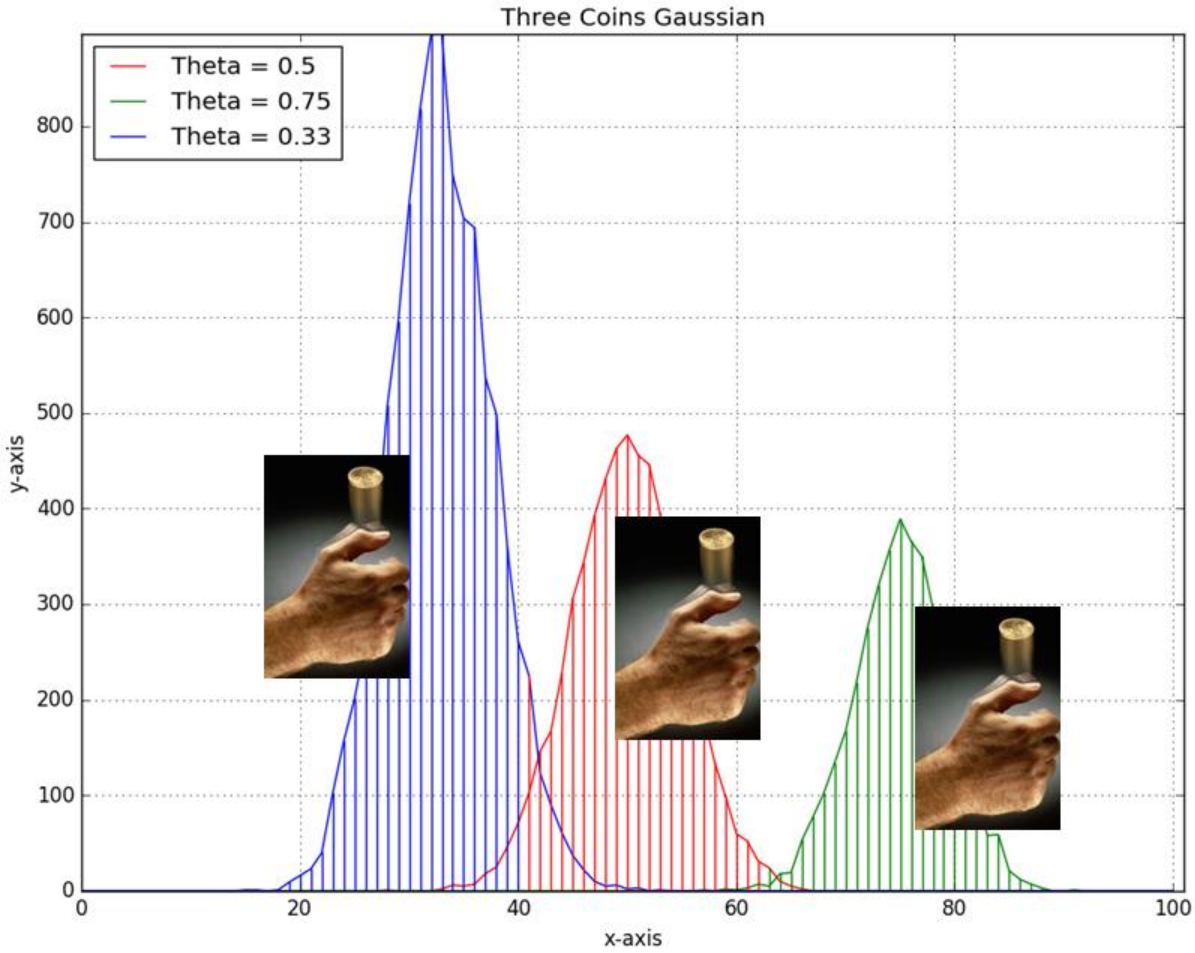


Discuss Question

Please discuss the similarities and differences between EM ,K-means and Bayesian classification?

Classification	Clustering
Bayesian Classification	K-means EM

Classification



Clustering

- Partitioning method
- Hierarchical method
- Density-base method
- Grid-based method
- Model-based method

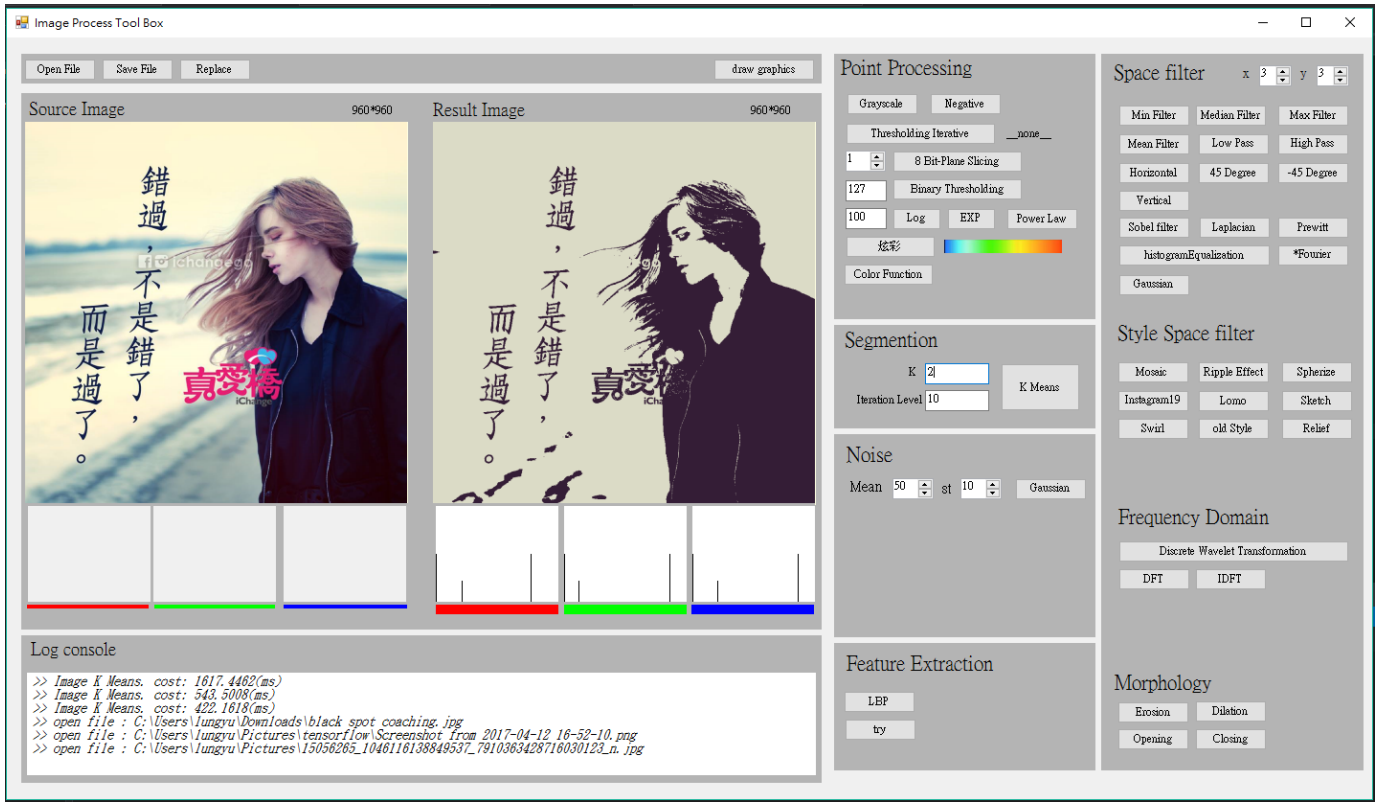
- Outlier detection

http://mirlab.org/jang/books/dcpr/

http://www.math.unipd.it/~aiolli/corsi/SI-0607/Lez17.211106.pdf

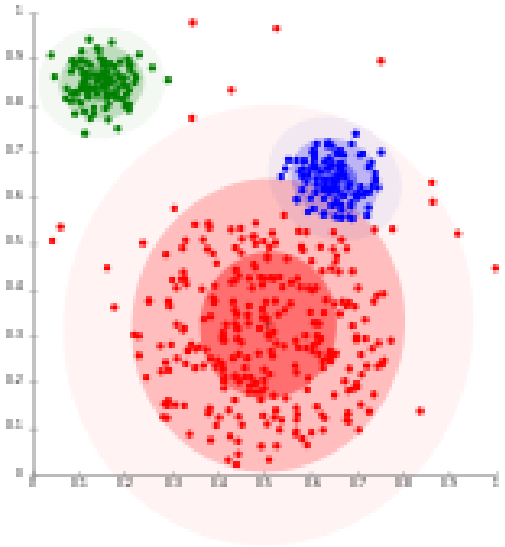
Partitioning method

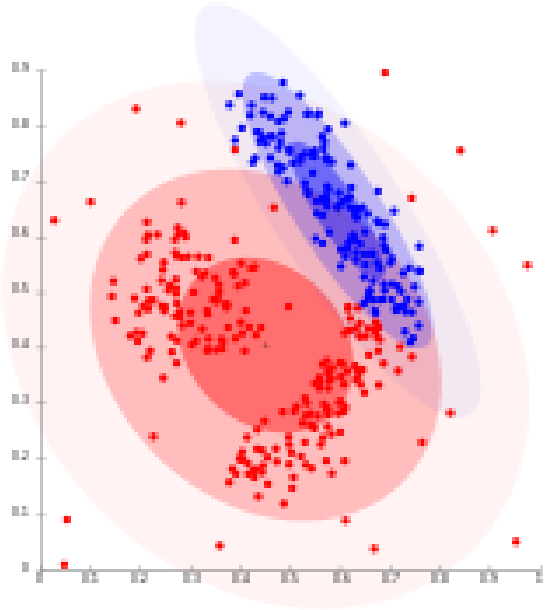
- Usually start with a random (partial)
- partitioning



Model-based method

- Instead of a Gaussian mixture





https://en.wikipedia.org/wiki/Cluster_analysis#Connectivity-based_clustering_.28hierarchical_clustering.29

Considerations

- Finding good seeds is even more critical for EM than for k-means (EM is prone to get stuck in local optima)

- Therefore (as in k-means) an initial assignment is often computed by another algorithm

Mean-Shift Clustering Algorithm

In [91]:
```
# The following bandwidth can be automatically detected using
bandwidth = estimate_bandwidth(X, quantile=0.2, n_samples=500)

ms = MeanShift(bandwidth=bandwidth, bin_seeding=True)
ms.fit(X)
labels = ms.labels_
cluster_centers = ms.cluster_centers_

labels_unique = np.unique(labels)
n_clusters_ = len(labels_unique)

print("number of estimated clusters : %d" % n_clusters_)
```

```
number of estimated clusters : 3
```

In [92]:
```
import matplotlib.pyplot as plt
from itertools import cycle

plt.figure(1)
plt.clf()

colors = cycle('bgrcmykbgrcmykbgrcmykbgrcmyk')
for k, col in zip(range(n_clusters_), colors):
    my_members = labels == k
    cluster_center = cluster_centers[k]
    plt.plot(X[my_members, 0], X[my_members, 1], col + '.')
    plt.plot(cluster_center[0], cluster_center[1], 'o', markerfacecolor=col,
             markeredgecolor='k', markersize=14)
plt.title('Estimated number of clusters: %d' % n_clusters_)
plt.show()
```