Xudong LIU

7lxd@tongji.edu.cn

School of Software Engineering

Tongji University

Nov 28, 2015

Clustering & EM Algorithm

ABSTRACT

Clustering is one of the most important tools in data mining. It can cluster data record without any prior information. Expectation–Maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step.

Keywords: Data Mining,Clustering, Expectation -Maximization Algorithm, Centroids,MSE,Likelihood,Local optimal,Evaluation.

INTRODUCTION

EM-Algorithm:

The EM algorithm is used to find (locally) maximum likelihood parameters of a statistical model in cases where the equations cannot be solved directly. Typically these models involve latent variables in addition to unknown parameters and known data observations. That is, either there are missing values among the data, or the model can be formulated more simply by assuming the existence of additional unobserved data points. For example, a mixture model can be described more simply by assuming that each observed data point has a corresponding unobserved data point, or latent variable, specifying the mixture component that each data point belongs to.

Finding a maximum likelihood solution typically requires taking the derivatives of the likelihood function with respect to all the unknown values — viz. the parameters and the latent variables — and simultaneously solving the resulting equations. In statistical models with latent variables, this usually is not possible. Instead, the result is typically a set of interlocking equations in which the solution to the parameters requires the values of the latent variables and vice versa, but substituting one set of equations into the other produces an unsolvable equation.

EVALUATION:

Clustering evaluation is the wary to measure how good the cluster algorithm clustering the set of data to clusters is without any class information. Generally speaking, there are two types of cluster evaluation techniques, which are based on external validation indices and internal validation indices.

- External validation: Based on previous knowledge about data which we called the ground truth.
- Internal validation: Without any previous knowledges, based on the clusters' internal structure.

ALGORITHM

2.1 EM-Algorithm:

E-step

Operate a Kalman filter or a minimum-variance smoother designed with current parameter estimates to obtain updated state estimates.

M-step

Use the filtered or smoothed state estimates within maximum-likelihood calculations to obtain updated parameter estimates.

Suppose that a Kalman filter or minimum-variance smoother operates on noisy measurements of a single-input-single-output system. An updated measurement noise variance estimate can be obtained from the maximum likelihood calculation.

$$\hat{\sigma}_v^2 = \frac{1}{N}\sum_{k=1}^{N}(z_k - \hat{x}_k)^2$$

Figure 2.1

where Xk are scalar output estimates calculated by a filter or a smoother from N scalar measurements Zk. Similarly, for a first-order auto-regressive process, an updated process noise variance estimate can be calculated by

$$\hat{\sigma}_w^2 = \frac{1}{N}\sum_{k=1}^{N}(\hat{x}_{k+1} - \hat{F}\hat{x}_k)^2$$

Figure 2.2

where Xk and Xk+1 are scalar state estimates calculated by a filter or a smoother. The updated model coefficient estimate is obtained via

$$\hat{F} = \frac{\sum_{k=1}^{N}(\hat{x}_{k+1} - \hat{F}\hat{x}_k)}{\sum_{k=1}^{N}\hat{x}_k^2}.$$

Figure 2.3

The convergence of parameter estimates such as those above are well studied.

EXPERIMENT

3.1 Local Optimal:

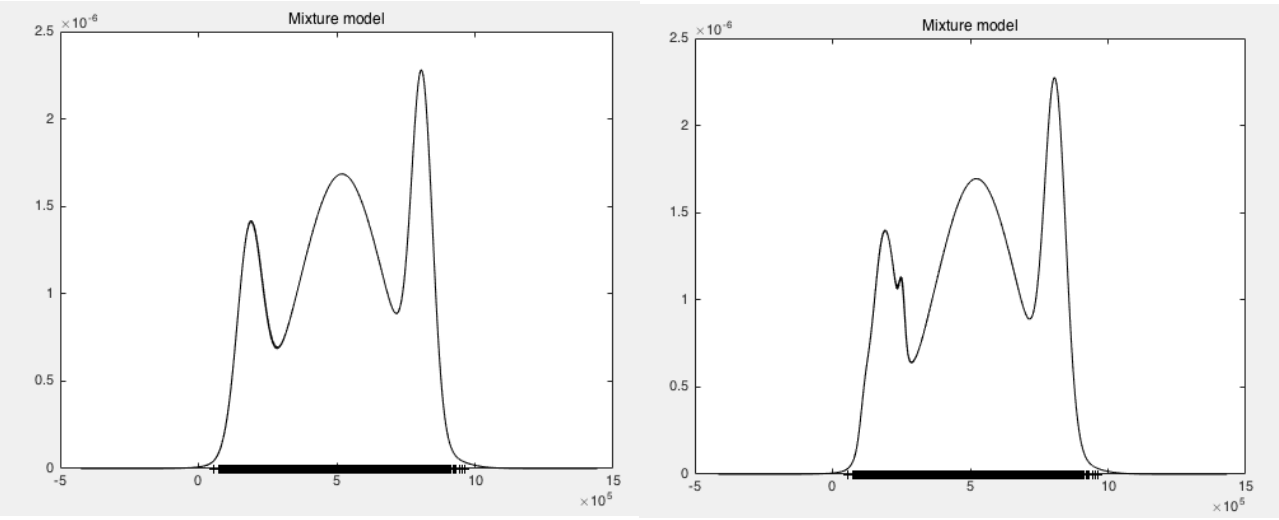
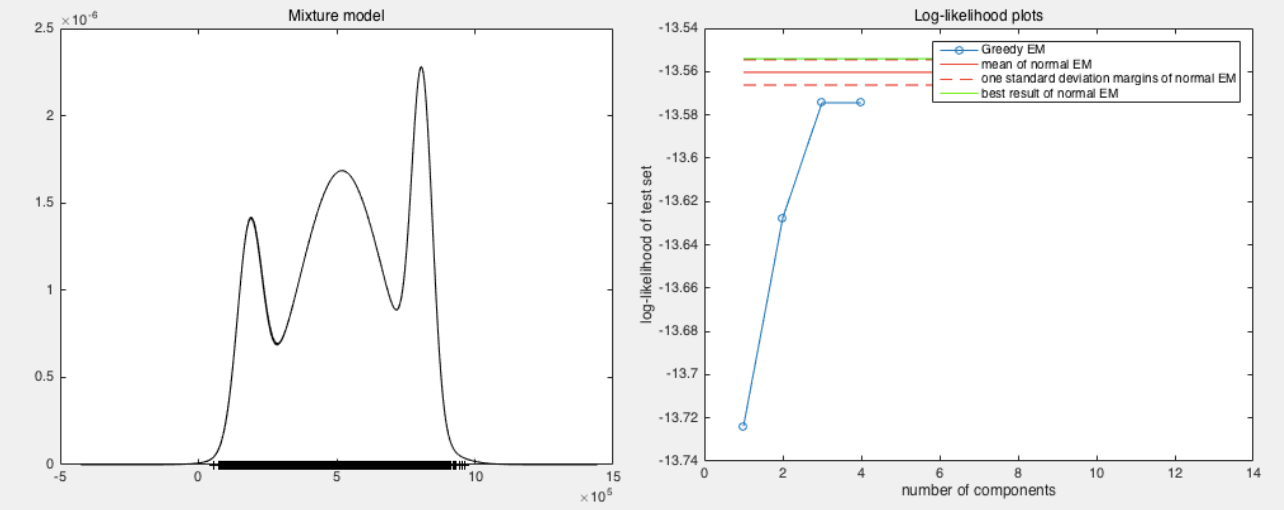


Figure 3.1

From the two pictures above, we can see that the EM-Algorithm is the local optimal algorithm. It may have many local optimal points and it won't find the best of all the local optimal points. The algorithm will stopped if it arrived one of the local optimal points.

3.2 Initialisation (greedy-EM):

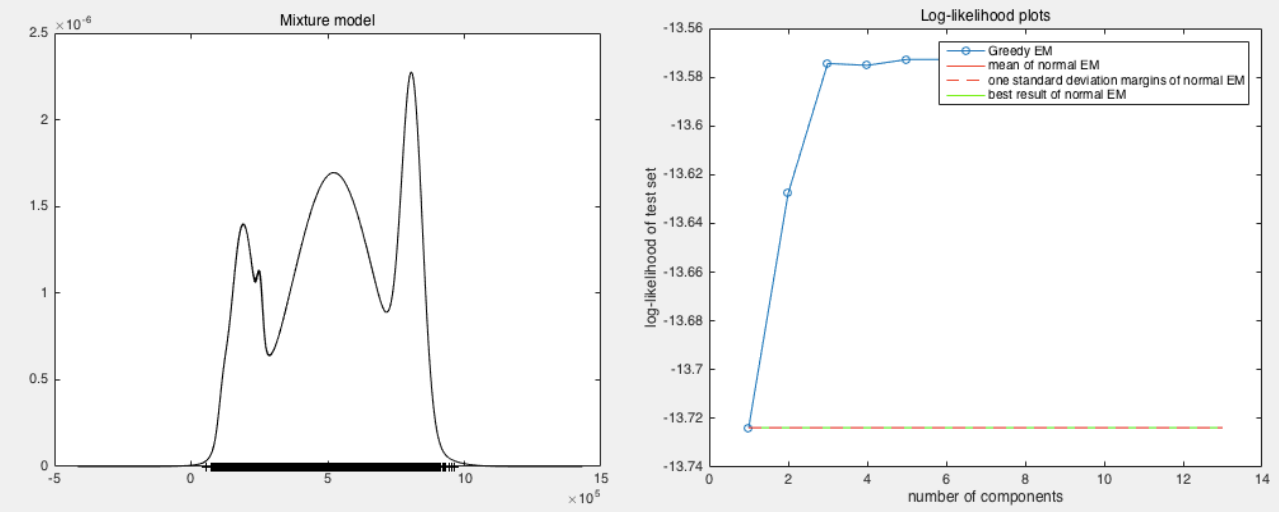
1.K-means init:



best_k = 3
log(likelihood) = -13.574598

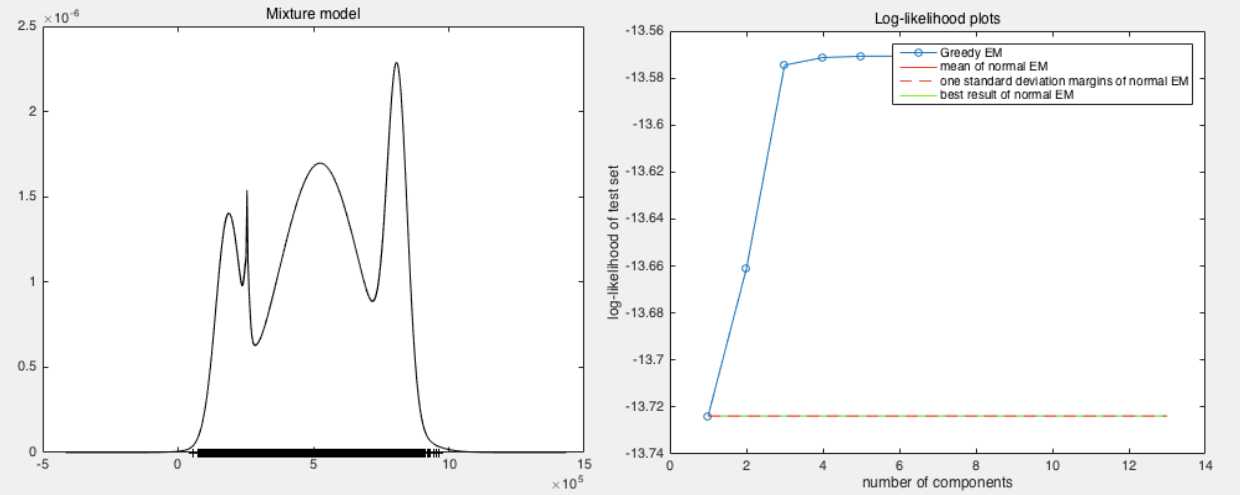
2. Random init:

1.



best_k = 5
log(likelihood) = -13.572747

2.



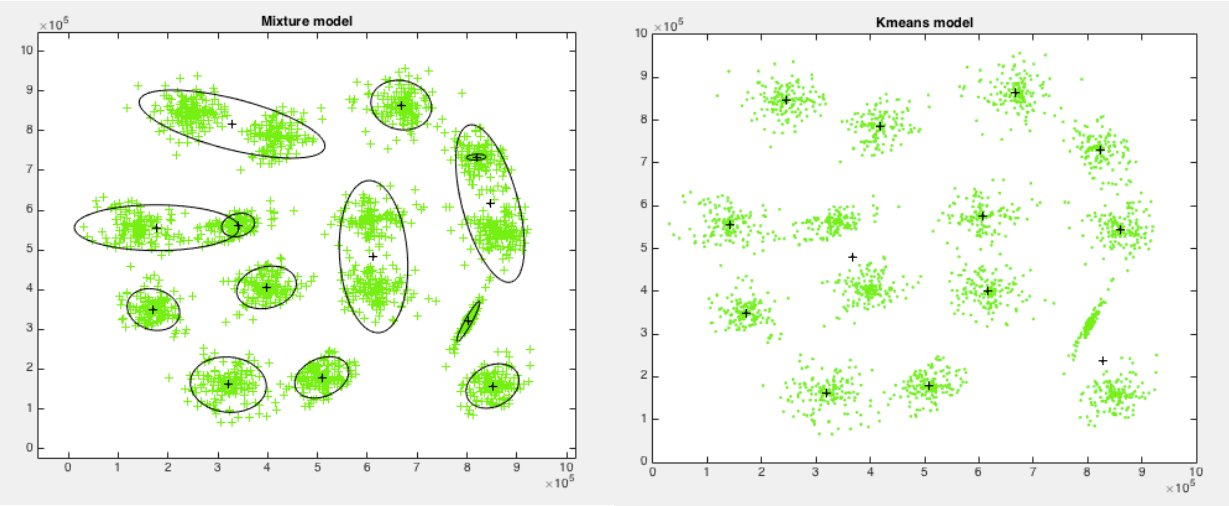
best_k = 5
log(likelihood) = -13.572747

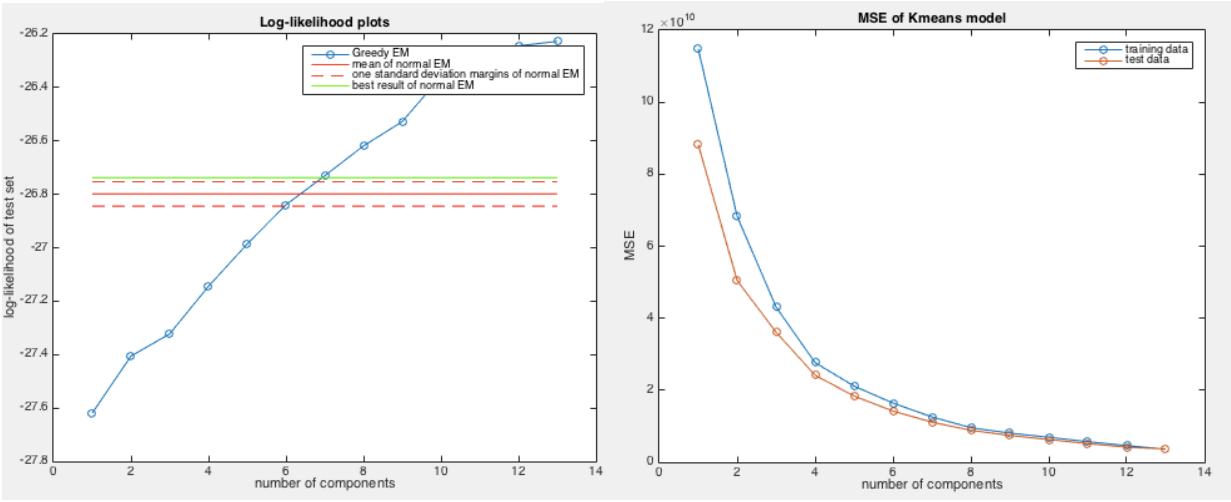
Conclusion:

From the two group experiments, it's easily to find that: K-means init will lead to one result as random init will lead to many different results. However, neither the K-means init or the random init, it won't guarantee that the result will be the best optimal.

3.3 EM Algorithm VS K-means Algorithm:

1.(DataSet:s1.txt)





EM-Algorithm:

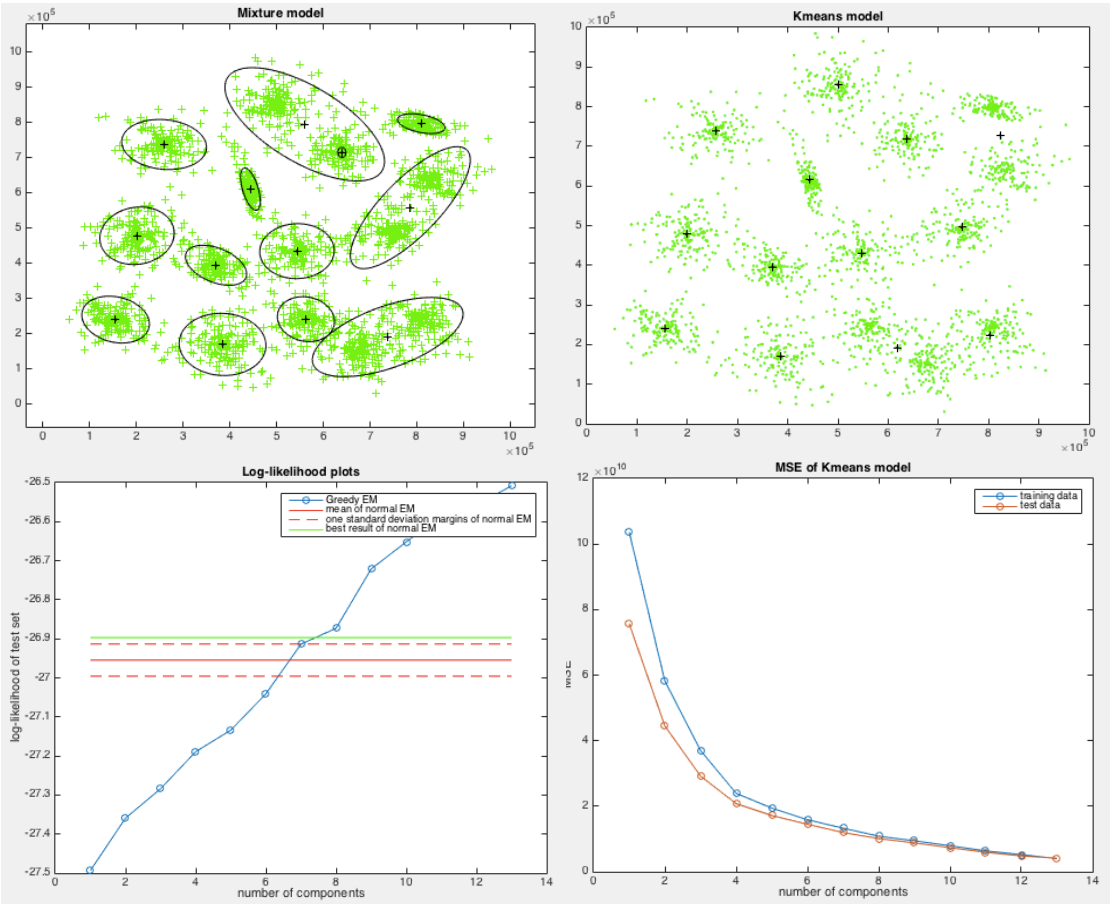
best_k = 13

log(likelihood) = -26.227021

K-means:

best_k = 13 min(MSE)

2.(DataSet:s2.txt)



EM-Algorithm:

best_k = 13

log(likelihood) = -26.508061

K-means:

best_k = 13 min(MSE)

Conclusion:

From the two experiments above, we can see that K-means algorithm and EM-algorithm have some differences but also have something in common. I will explain this as follows:

Similarities:

1. Both of them are local optimal algorithm
2. All need the parameter K.
3. Both of them are clustering data.

Differences:

1. Suited Data Structure:

   K-means: Continuous Euclidean distance

   EM-Algorithm: Gaussian

2. Evaluation Index:

   K-means: MSE(Mean Squared Error)

   EM-Algorithm: Log(likelihood)

3.4 Evaluation EM with External index and Internal index:

External index:

Jaccard coefficient: J = SS/(SS+SD+DS)

Use Jacccard index evaluate the EM-greedy algorithm:

J-index = **0.5906~0.6554**

Internal index:

Ball and Hall = SSW/m

Use Ball and Hall index evaluate the EM-greedy algorithm:

BH-index = **0.2301~0.2363**

REFERENCE:

[1] Data Set From Professor ZHAO: http://sse.tongji.edu.cn/zhaoqinpei/Datasets/

[2] Data Set From UCI Repo: http://archive.ics.uci.edu/ml/datasets/Iris

[3] *Internal versus External cluster validation indexes* from INTERNATIONAL JOURNAL OF COMPUTERS AND COMMUNICATIONS Issue 1, Volume 5, 2011 by Eréndira Rendón, Itzel Abundez, Alejandra Arizmendi and Elvia M. Quiroz.