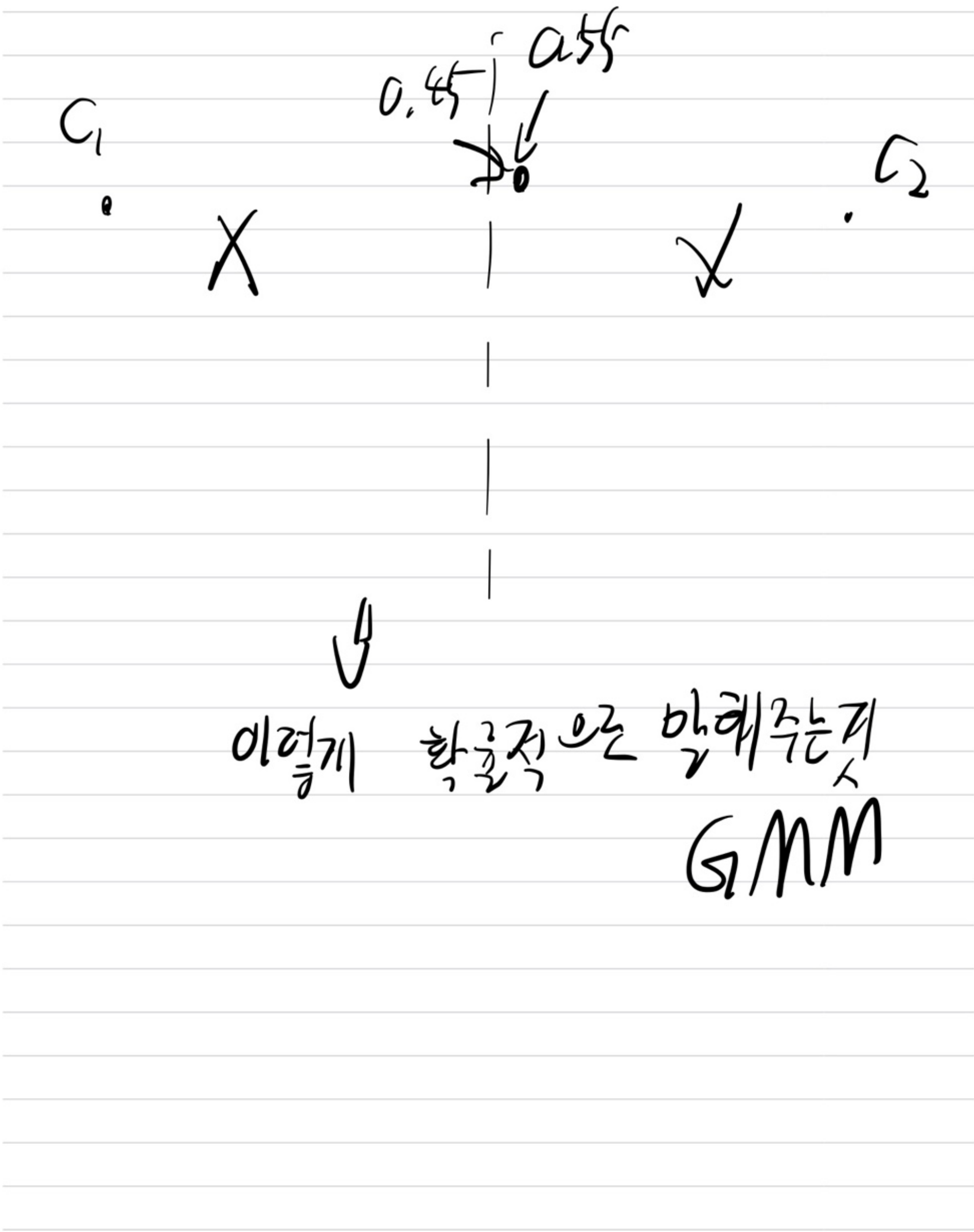
Gaussian Mixture Model

Introduction

- A probabilistic version of K-means



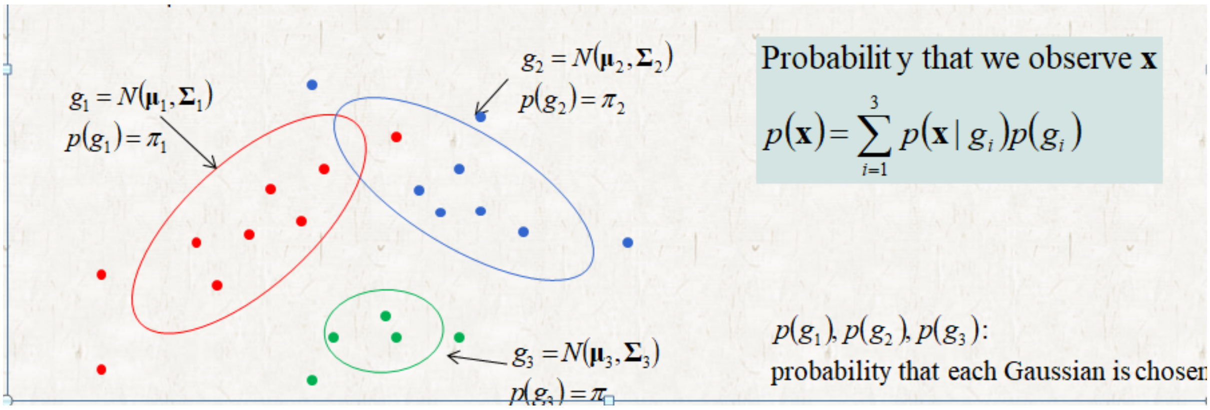
- Soft boundaries: a data point belong to all clusters. The degrees of belonging are different

- Clustering with any shape of ellipses
- Usually, solved by Expectation and Maximization Algorithm

Mixture of Gaussians

Example

- We have 3 Gaussians
- We know the parameters of Gaussians
- Randomly choose one of Gaussians, and generate a data
- Repeat this!



Gaussian Mixture Model

Example

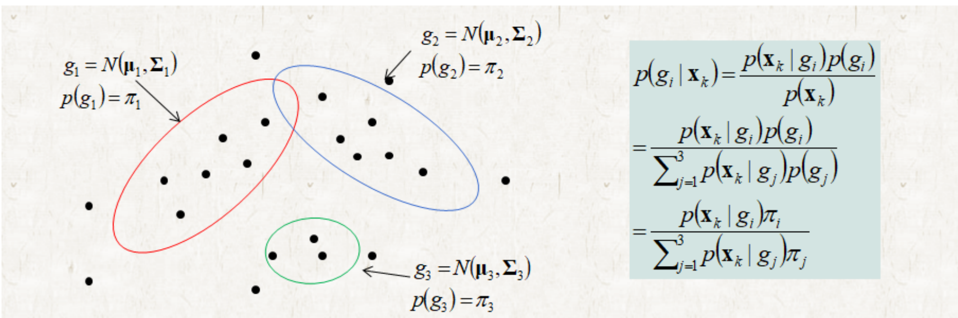
- We have data which are generated by 3 Gaussians
- (1)We don't know parameters and chosen prob of Gaussians
- (2)We don't know which Gaussian generates each data

  —>Find the Gaussians

Situation1

- We have data which are generated by 3 Gaussians
- We know parameters and chosen prob of Gaussians
- We don't know which Gaussian generates each data

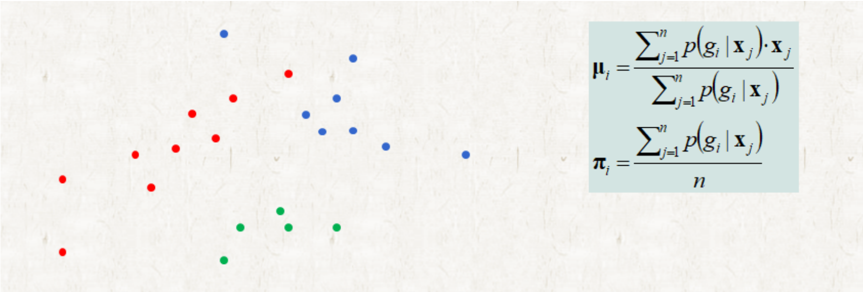
—>Guess which Gaussian generates a given data point=P(gi|x)



Situation2

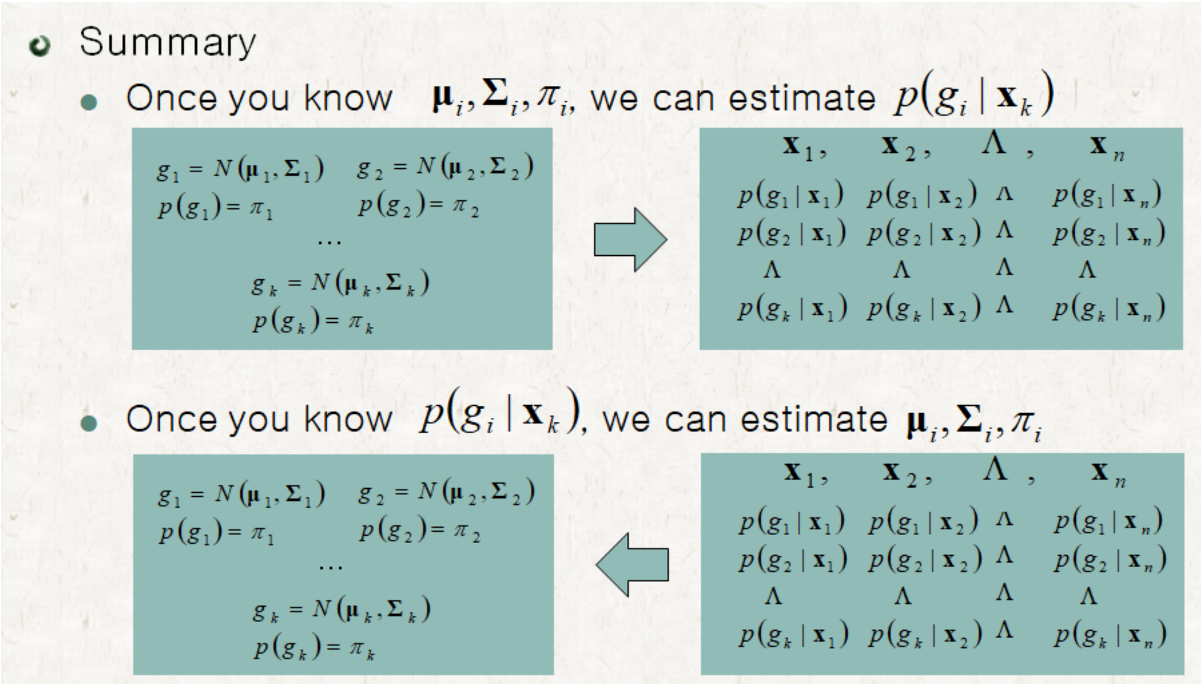
- We have data which are generated by 3 Gaussians
- We don't know parameters and chosen prob of Gaussians
- We Know which Gaussian generates each data

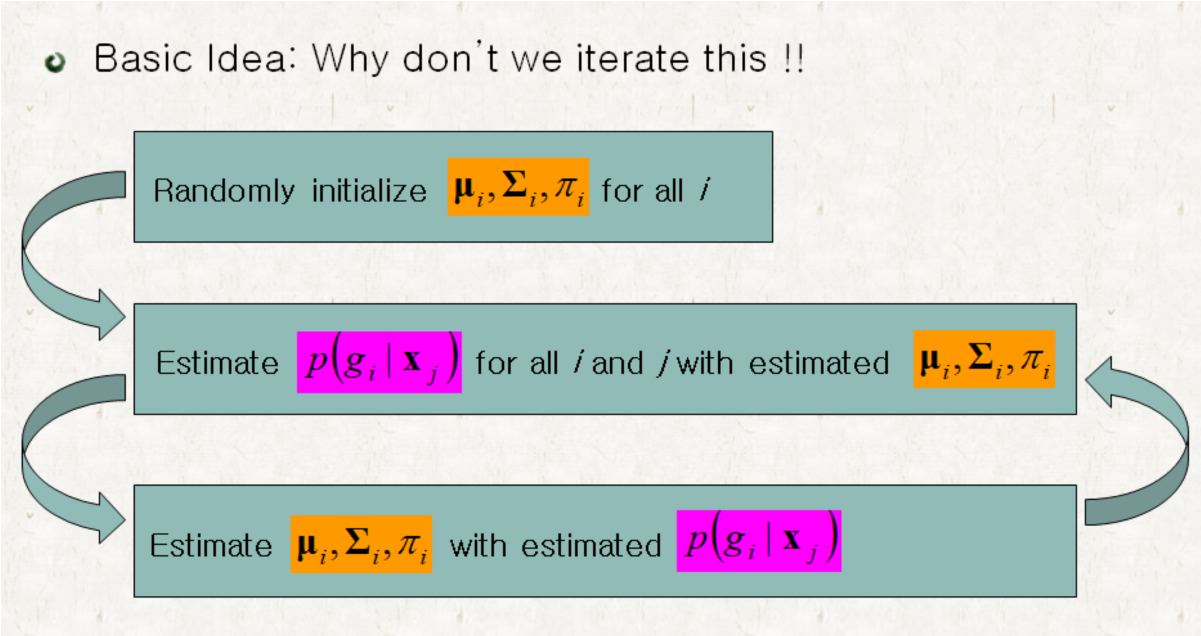
—> Guess the parameters and the chosen prob of Gaussians i:i번째 가우시안



만약 빨간 한 점이 xj 빨간 가우시안의 선택확률을 p(g1)이라 할 경우, P(g1|xj)=1 ,P(g2|xj)=0,P(g3|xj)=0 이다.

situation1,2를 종합해보면





1. Randomly initialize

$$\boldsymbol{\mu}_1^0, \boldsymbol{\Sigma}_1^0, \Lambda, \boldsymbol{\mu}_k^0, \boldsymbol{\Sigma}_k^0, \pi_1^0, \Lambda, \pi_k^0$$

2. Evaluate for all i and j

$$p(g_i \mid \mathbf{x}_j) = \frac{p(\mathbf{x}_j \mid g_i) \cdot \pi_i^t}{\sum_{c=1}^{k} p(\mathbf{x}_j \mid g_c) \cdot \pi_c^t} \quad \text{where} \quad p(\mathbf{x}_j \mid g_i) = N(\mathbf{x}_j \mid \boldsymbol{\mu}_i^t, \boldsymbol{\Sigma}_i^t)$$

3. Evaluate for all i

$$\boldsymbol{\mu}_i^{t+1} = \frac{\sum_{j=1}^{n} p(g_i \mid \mathbf{x}_j) \cdot \mathbf{x}_j}{\sum_{j=1}^{n} p(g_i \mid \mathbf{x}_j)} \qquad \boldsymbol{\Sigma}_i^{t+1} = \frac{\sum_{j=1}^{n} p(g_i \mid \mathbf{x}_j) \cdot (\mathbf{x}_j - \boldsymbol{\mu}_i^{t+1})^T \bullet (\mathbf{x}_j - \boldsymbol{\mu}_i^{t+1})}{\sum_{j=1}^{n} p(g_i \mid \mathbf{x}_j)}$$

$$\pi_i^{t+1} = \frac{1}{n} \sum_{j=1}^{n} p(g_i \mid \mathbf{x}_j) \qquad \text{where } \mathbf{x}_k = (x_{k1}, x_{k2}, \Lambda, x_{kd}),\ \boldsymbol{\mu}_i^t = (\mu_{i1}^t, \mu_{i2}^t, \Lambda, \mu_{id}^t)$$

4. Go back to Step 2, stop until parameters don't change

initialize

covariance:diagonal을 1로 나머진 0

파이:1/n

Comparison to K-means

1. Randomly initialize

$$\boldsymbol{\mu}_1^0, \boldsymbol{\Sigma}_1^0, \Lambda, \boldsymbol{\mu}_k^0, \boldsymbol{\Sigma}_k^0, p_1^0, \Lambda, p_k^0$$

2. Evaluate

$$p(g_i \mid \mathbf{x}_j) = \frac{p(\mathbf{x}_j \mid g_i)\cdot \pi_i^t}{\sum_{c=1}^{k} p(\mathbf{x}_j \mid g_c)\cdot \pi_c^t}$$

3. Evaluate

$$\boldsymbol{\mu}_i^{t+1} = \frac{\sum_{j=1}^{n} p(g_i \mid \mathbf{x}_j)\cdot \mathbf{x}_j}{\sum_{j=1}^{n} p(g_i \mid \mathbf{x}_j)}$$

$$\boldsymbol{\Sigma}_i^{t+1} = \frac{\sum_{j=1}^{n} p(g_i \mid \mathbf{x}_j)\cdot (\mathbf{x}_j - \boldsymbol{\mu}_i^{t+1})^T \bullet (\mathbf{x}_j - \boldsymbol{\mu}_i^{t+1})}{\sum_{j=1}^{n} p(g_i \mid \mathbf{x}_j)}$$

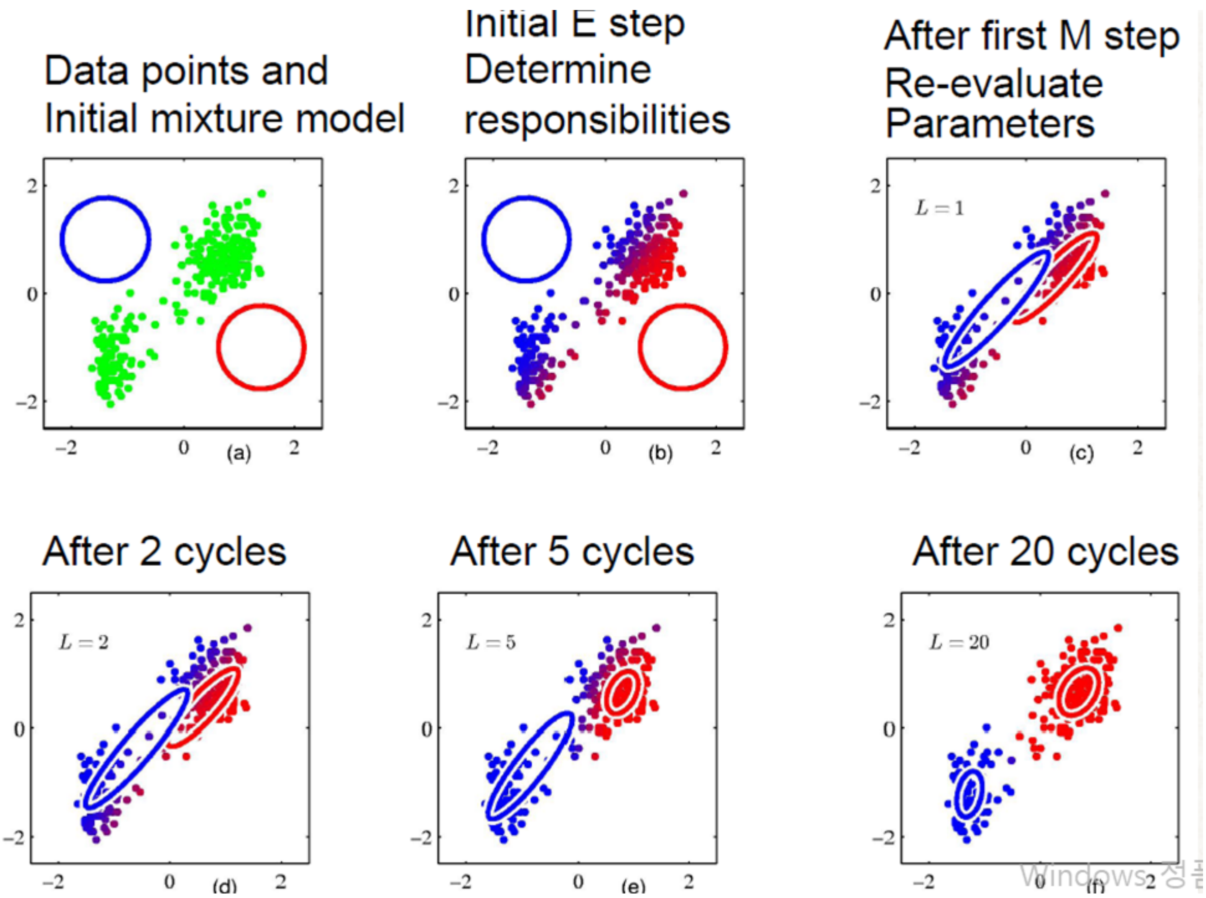
$$\pi_i^{t+1} = \frac{1}{n}\sum_{j=1}^{n} p(g_i \mid \mathbf{x}_j)$$

4. Go back to Step 2 until parameters don't change

1. Randomly choose seed points.
2. Assign each object to the nearest seed point.
3. Compute the centroids (mean point) of the current clusters.
4. Go back to Step 2 until parameters don't change

k-means is a hard version of GMM

GMM과정이 MLE를 최대화 하는 과정이다.



Practical Issues

- Takes more iterations than K-means.
- Each cycle requires significantly more comparison
- Common to run K-means first in order to find suitable initialization
- Covariance matrices can be initialized to covariances of clusters found by K-means.

—>K-means를 먼저 사용하여 중심점을 찾아놓고 그걸 initial point로 잡는다.

- Problem of Singularities(특이점 문제)

—>하나의 데이터에 가우시안 분포가 생겨서(∵ MLE때문에)—>그러므로 covariance 초기값을 넉넉하게 줘야한다.

- EM is not guaranteed to find global maximum of log likelihood function