Probabilistic Machine Learning:

8. Mixture Models & EM

ESC 2024 Spring Session 2주차

Contents

1. K-means Clustering

Introduction to K-means Clustering before GM and EM.

Since EM algorithm is analogous to the process of K-means Clustering.

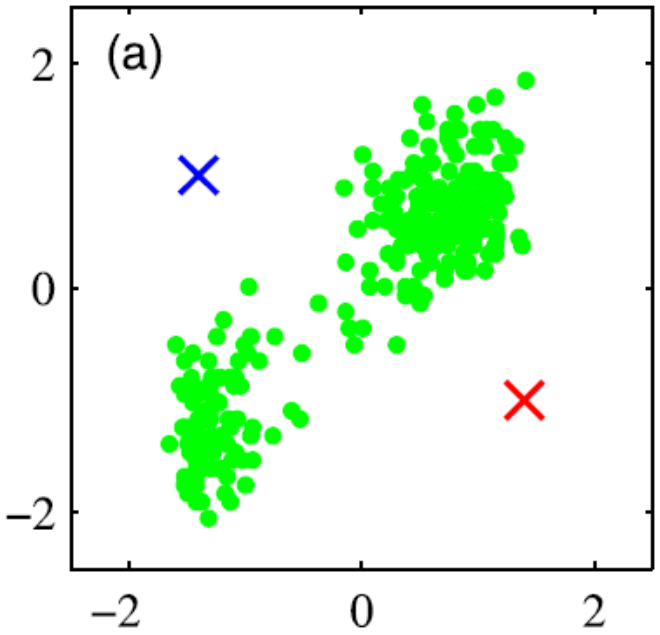
Let's begin to the question of how can we assign cluster to data points?

K-means Clustering

How to identify groups, or clusters, of data points ?

Our goal is to partition the

N observations of D-dimensional variable x to K clusters



Initialize cluster centers

And let's define the Energy function called

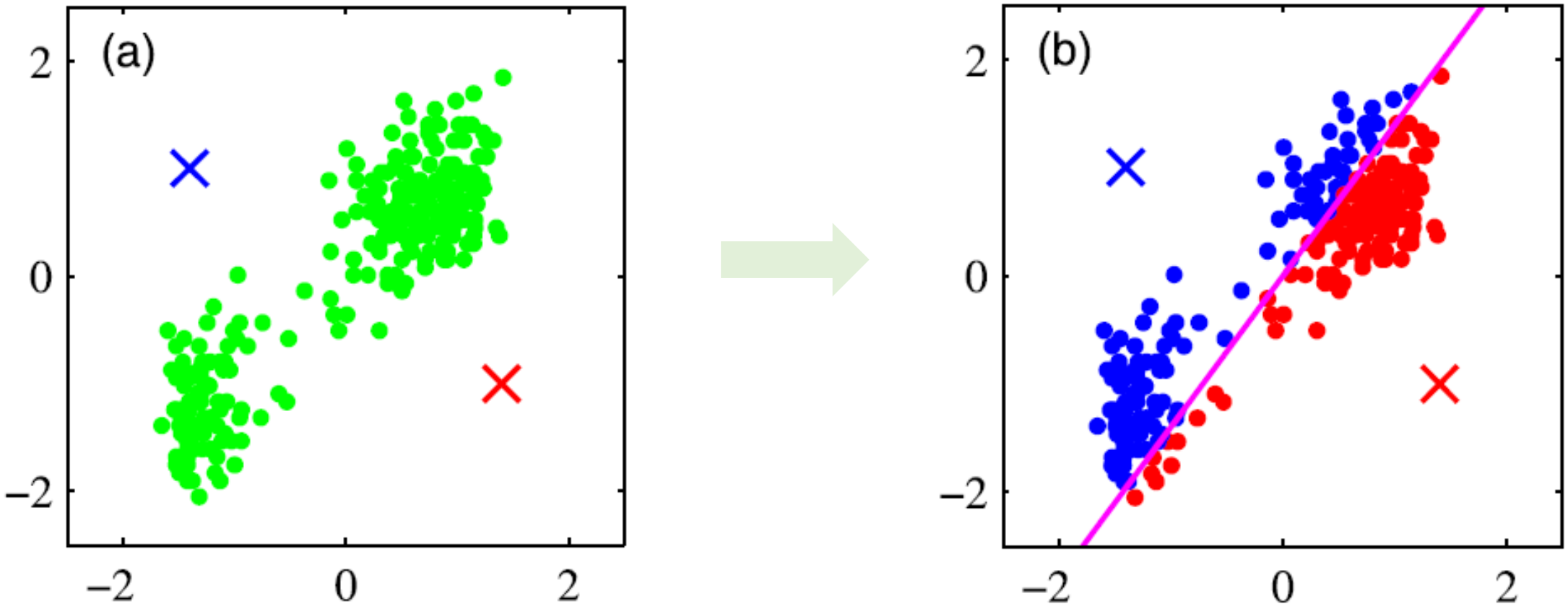
Distortion(or heterogeneity)

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2 \qquad r_{nk} = \begin{cases} 1 \text{ if } k = \text{argmin}_j ||\mathbf{x_n} - \mu_j||^2 \\ 0 \text{ otherwise.} \end{cases}$$

And assign cluster that minimizes distortion to each data points

K-means Clustering

E step : assign cluster that minimizes distortion to each data points



K-means Clustering

M step : re-compute each cluster center

Take a derivative to Distortion w.r.t μ_k

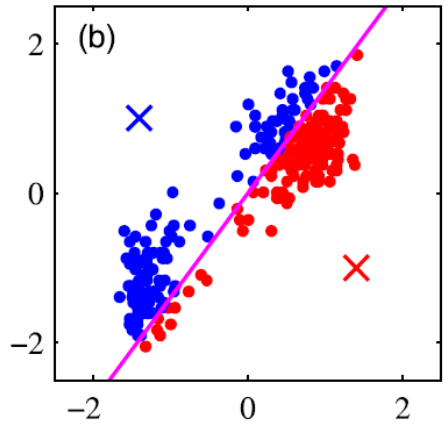
$$2\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \mu_k) = 0$$

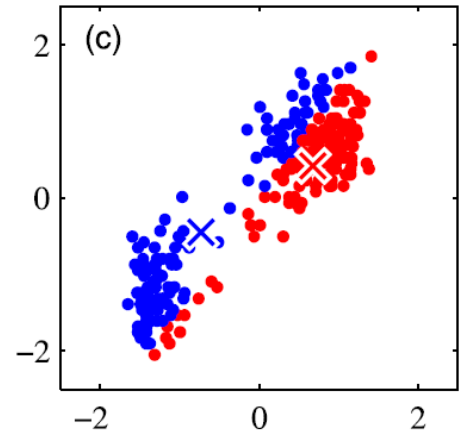
Then we obtain new center of cluster

$$\mu_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

It is form of a weighted average.

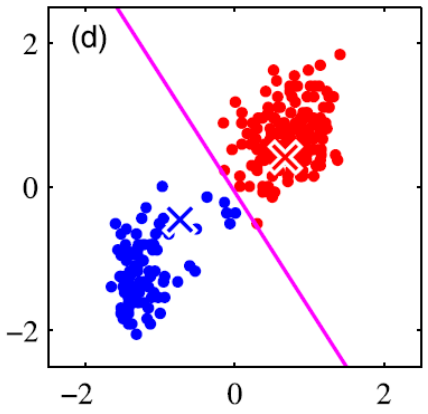
The denominator is equal to the # of points to cluster K

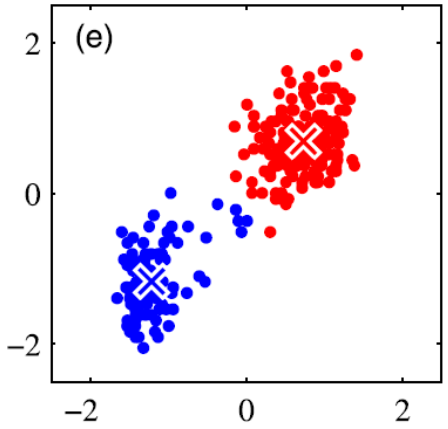




K-means Clustering

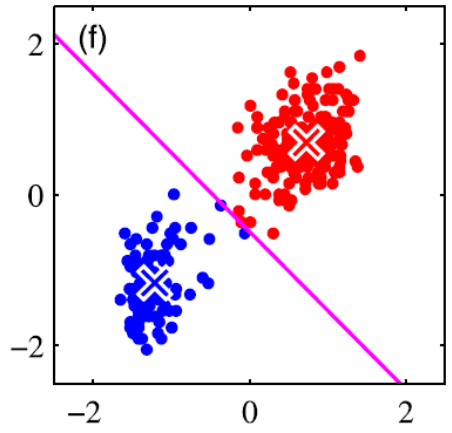
And check for the convergence of either the parameters or the distortion

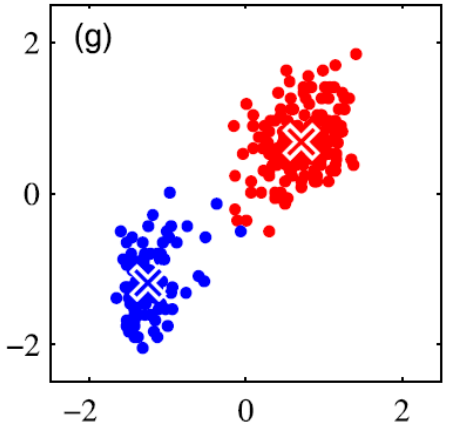


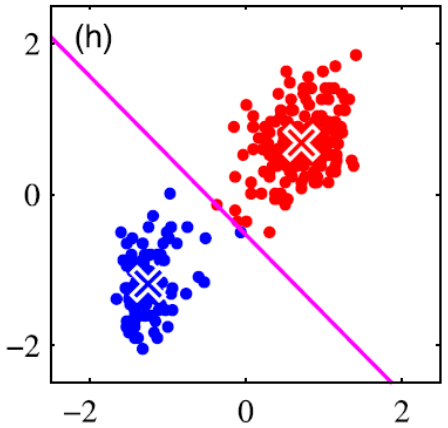


Until the cluster center does not move

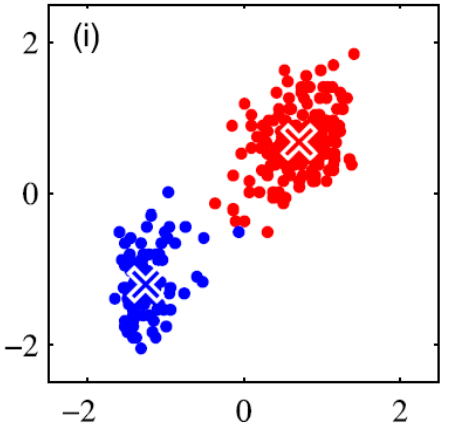
$$|\mu_{k+1} - \mu_k| \leq \epsilon$$







Converged



K-means Clustering (Appendix)

We can also derive an on-line stochastic algorithm

By applying the Robbins-Monro procedure

$$\mu_k^{new} = \mu_k^{old} + \eta_n(x_n - \mu_k^{old})$$

K-means Clustering (Appendix)

We can generalize the K-means algorithm

by introducing a more general dissimilarity…

$$\tilde{J} = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\mathcal{V}(\mathbf{x}_n, \mu_k)$$

It is called K-medoids algorithm.

Medoid means a point having smallest dissimilarity in the cluster

What will be the time complexity of K-means Clustering?

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}||x_n - \mu_k||^2$$

E-step : $O(nk)$

$$\mu_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

M-step : $O(n)$

However, time complexity of

K-medoids in M-step $O(n^2)$

Since we have to consider ∀data points

when we assign medoids

2. GM and EM

Linear combination decomposes complex entire object into sum of simple part.

Mixture of Gaussian does similar thing,

we can approximate complex distribution to sum of simple gaussian basis

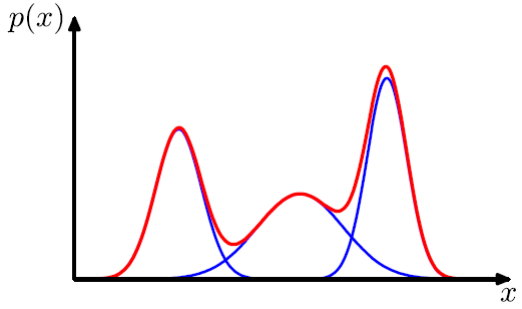
1. Mixture of Gaussians
2. Expectation-Maximization Method
    - E-step
    - M-step

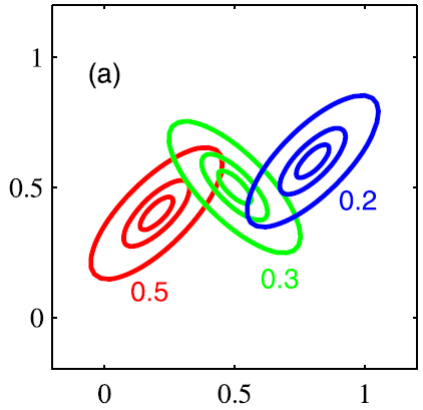
Mixture of Gaussians

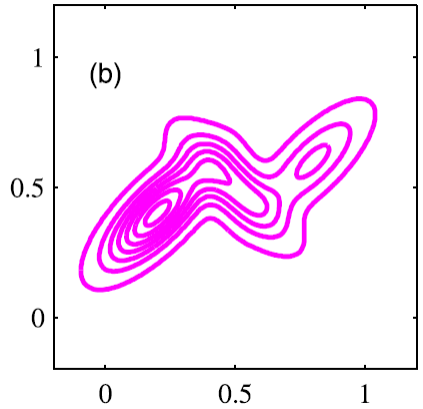
While the simple Gaussian distribution has some important analytical properties

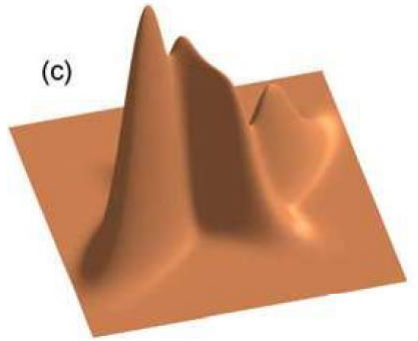
It suffers from significant limitations !!!

When it comes to modeling real data sets









Whereas a linear superposition of several Gaussians gives a better characterization…

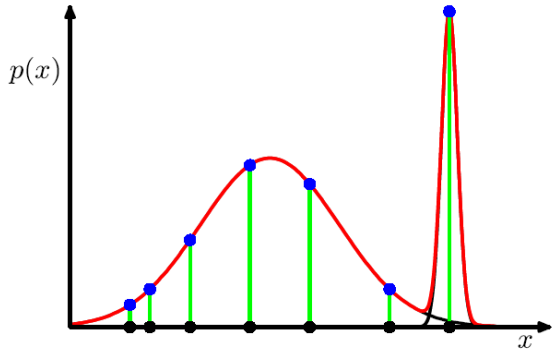
Mixture of Gaussians

linear superposition of several Gaussians gives a better characterization…

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_\mathbf{k}, \mathbf{\Sigma_k})$$

$\boldsymbol{\pi_k}$ **is mixing Coefficients.**

$\sum \boldsymbol{\pi_k} = \mathbf{1}, \boldsymbol{where}\ \mathbf{0} \leq \boldsymbol{\pi_k} \leq 1$



Our log-likelihood Function is

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln\{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x_n}|\mu_\mathbf{k}, \mathbf{\Sigma_k})\}$$

But, only using likelihood can cause "collapse" (severe overfitting)

Apply Bayes' Rule!

Mixture of Gaussians

From the sum and product rules, the marginal density is given by

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k)p(\mathbf{x}|k)$$

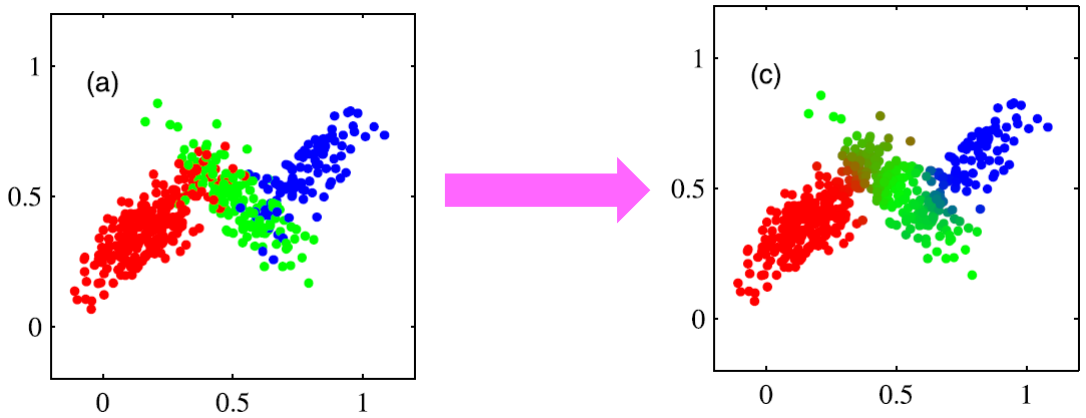
Assume the prior $p(k) = \pi_k$, the likelihood $p(x|k) = N(x|\mu_k, \Sigma_k)$

then the posteriori is ...

(r_k is called responsibility)

$$r_k(\mathbf{x}) \equiv p(k|\mathbf{x})$$

$$= \frac{p(k)p(\mathbf{x}|k)}{\sum_l p(l)p(\mathbf{x}|l)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_l \pi_l \mathcal{N}(\mathbf{x}|\mu_1, \mathbf{\Sigma_1})}$$



How do we optimize this Bayesian GM model?

Try to MLE

Mixture of Gaussians

Take a derivative to log-likelihood

$$\ln p(\mathbf{X}|\pi,\mu,\Sigma) = \sum_{n=1}^{N} \ln\{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x_n}|\mu_\mathbf{k},\mathbf{\Sigma_k})\}$$

w.r.t μ_k

Responsibility occurs

$$0 = -\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k,\Sigma_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\mu_j,\Sigma_j)} \Sigma_k^{-1}(\mathbf{x}_n - \mu_k)$$

$$\mu_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n$$

Where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

= Sum of responsibility

∀data points

With respect to covariance

$$\Sigma_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \mu_k^{\text{new}})(\mathbf{x}_n - \mu_k^{\text{new}})^{\mathrm{T}}$$

From Lagrange Multiplier $\ln p(\mathbf{X}|\pi,\mu,\mathbf{\Sigma}) + \lambda(\sum_{k=1}^{K}\pi_k - 1)$

$$\pi_k^{\text{new}} = \frac{N_k}{N}$$

(By derivative w.r.t π_k)

Identifiability Problem occurs

K-component mixture will have a

total of K! equivalent solutions of assigning

K sets of parameters to K components

Mixture of Gaussians

The log of the likelihood function

no longer has a closed-form analytical solution

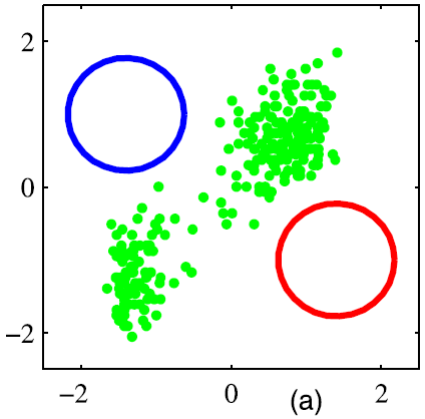
Iterative numerical optimization techniques only can be used
One approach is to apply gradient-based optimization techniques

However we now consider an

alternative approach known as the EM algorithm

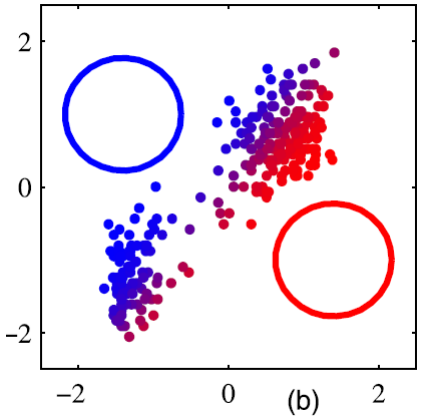
Expectation-Maximization Method

E step



We first choose some initial values

for the means μ_k, covariances Σ_k, and mixing coefficients π_k.

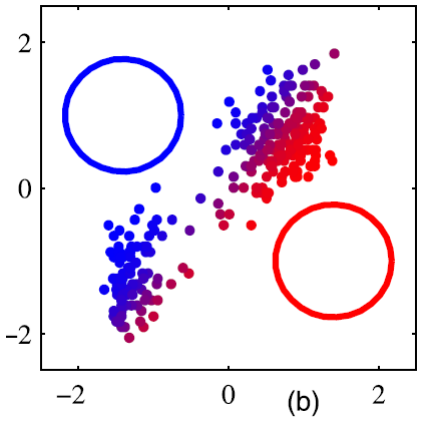


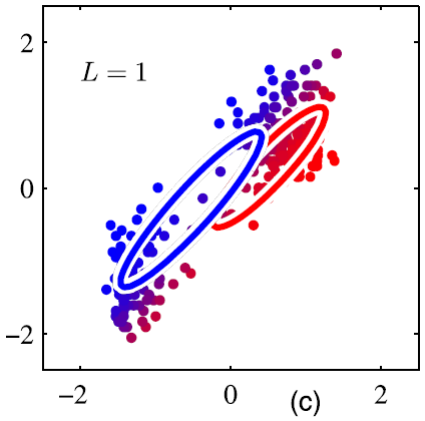
Evaluate the responsibilities $\gamma_k(\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_l \pi_l \mathcal{N}(\mathbf{x}|\mu_1, \boldsymbol{\Sigma}_1)}$

using the current parameter values

Expectation-Maximization Method

M step





Re-estimate the parameters using the current responsibilities

$$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

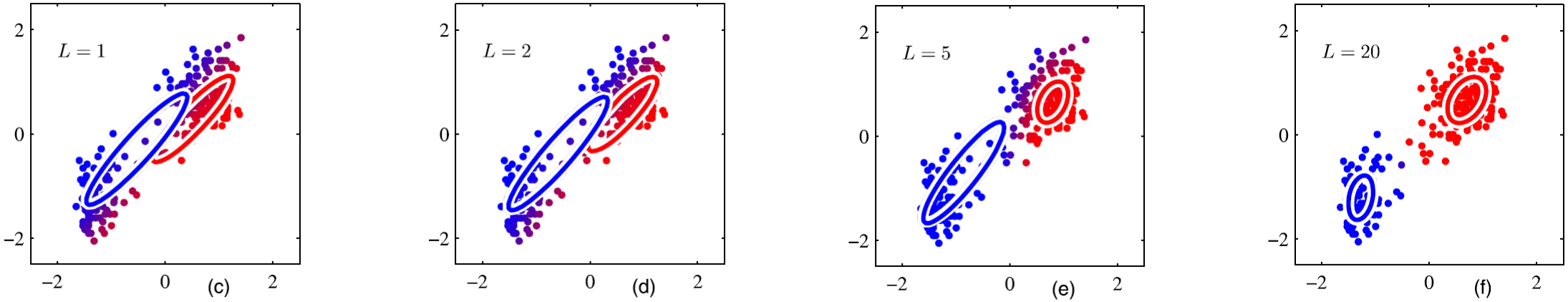
$$\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \mu_k^{\text{new}})(\mathbf{x}_n - \mu_k^{\text{new}})^{\text{T}}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N} \qquad \textbf{where} \quad N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

Expectation-Maximization Method

Evaluate the log likelihood

$$\ln p(\mathbf{X}|\pi,\mu,\Sigma) = \sum_{n=1}^{N} \ln\{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x_n}|\mu_\mathbf{k},\mathbf{\Sigma_k})\}$$



Converged

And check for the convergence of either the parameters or the log-likelihood

If the convergence criterion is not satisfied return to E-step

3

An Alternative View of EM

Gaussian mixtures revisited
Relations to K-means
Mixtures of Bernoulli distributions

The General EM Algorithm

EM Algorithm을 사용하는 목적

: 잠재 변수(latent variables)를 갖는 모델의 MLE 해를 구하기 위함

→ Z에 대해 사전에 알려진 정보는 없으며, 잠재 변수 Z와 입력 변수 X에 대한 조건부 분포 형태로 문제를 풀어야 한다.

EM Algorithm을 사용할 수 있는 조건

1) 잠재 변수가 존재해야 한다.

2) 잠재 변수가 관찰되지 않았을 경우에는 모수 추정이 불가능한 상태이나, 관찰된 이후에는 추정이 가능해진다.

→ 혼합 분포가 EM을 적용하기 좋은 예시

(잠재 변수의 역할이 혼합 분포 내의 한 분포를 선택하는 것을 나타내는 랜덤 변수이므로)

The General EM Algorithm

잠재 변수가 주어진 모델의 MLE에 대한 이해

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\}. \qquad (9.29)$$

· 결합 확률 분포 $p(X,Z|\theta)$가 지수족 분포를 따른다고 해도, 이에 대한 주변 확률 분포 $p(X|\theta)$는 지수족 분포가 아닐 수 있다.

· 가우시안 함수를 합한 식에 로그를 취한 함수는 간단한 이차 형식(quadratic)의 함수가 아니다.

→ 복잡한 형태의 MLE 풀이를 갖게 된다. (EM Algorithm이 필요)

The General EM Algorithm

Complete Data vs Incomplete Data

1) Complete Data : $\{X, Z\}$

: 모든 관찰 데이터에 대해 잠재 변수 Z가 주어진 모델

: log-likelihood function은 $\ln p(X, Z|\theta)$이며, 각각에 대해 likelihood를 구하면 되므로 어렵지 않음

2) Incomplete Data : Actual observed data X

: 가능한 모든 잠재 변수의 값에 대한 분포를 고려해야 한다. (log likelihood function : $\ln p(X|\theta) = \ln \sum_Z p(X, Z|\theta)$

: 로그 함수 내에 합에 대한 수식이 존재하므로, 최대화 시키는 지점을 찾기가 매우 어렵다.

결국, 계산을 잘 처리하기 위해서는 잠재 변수 Z가 관찰되어야 하지만 일반적으로 실제 값을 얻을 수 없다.

따라서 MLE를 풀기 위해서는 제공된 샘플을 통해 잠재 변수를 추정하여 값이 제공되었다고 생각할 수 있다.

Likelihood function에서 잠재 변수의 값을 정확히 알기 어렵기 때문에, 평균값을 활용한다.

: $E[\ln p(X, Z|\theta)]$를 사용 (Expectation of Log-likelihood)

The General EM Algorithm

Likelihood function의 기대값을 조금 더 전개해보면,

$$E_Z[\ln p(\mathbf{X}, \mathbf{Z}|\theta)] = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta) \ln p(\mathbf{X}, \mathbf{Z}|\theta)$$

이제 이 기대값을 최대로 만드는 파라미터 값을 추론해야 하는데, 여기서 EM Algorithm을 활용하게 된다.

· **E-step**

주어진 파라미터 θ^{old}는 고정되어 있고, 이를 활용해 잠재변수 Z의 확률값 $p(Z|X, \theta^{old})$를 얻어야 한다.

이제 다시 $p(Z|X, \theta^{old})$를 활용하여 complete-data에서 log likelihood의 expectation을 서술하면 다음과 같다.

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}). \qquad (9.30)$$

· **M-step**

예측된 Z의 기대값을 최대화하는 새로운 파라미터 θ^{new} 를 구한다.

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}). \qquad (9.31)$$

· 이 과정을 반복하여, 결과값이 수렴할 경우 종료한다.

Gaussian Mixtures Revisited

잠재 변수를 이용한 가우시안 혼합 분포(GMM)의 재해석

· Complete 데이터에 대한 (log) – likelihood function

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}} \qquad (9.35)$$

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\} . \qquad (9.36)$$

1) 로그함수가 가우시안 분포에 바로 적용이 가능하다. (파라미터의 MLE 값 계산이 쉽게 가능해짐)

2) $z_k \in 0,1$ 이므로 식 전개에는 영향을 주지 못하고, 하나의 가우시안 함수에 대해 처리되는 것처럼 동작함.

3) 각각의 k에 대해 MLE를 구하게 된다면, 다음과 같은 결과를 얻을 수 있다.

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} z_{nk} \qquad (9.37)$$

Gaussian Mixtures Revisited

· Incomplete 데이터에 대한 로그 가능도 함수

$$\ln p(\mathbf{X}|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(\mathbf{x}_n|\mu_k, \Sigma_k) \right\}$$

로그 함수 내에 합의 형태로 식이 존재하므로 MLE 처리가 매우 어렵다.

따라서 EM에서 사용했던, 로그 가능도 함수의 평균값을 정의한 후 이를 최대화하는 방식을 전개해야 한다.

따라서 잠재 변수 Z의 posterior를 통해 Z를 먼저 추론해야 한다. (→ Next page)

Z에 대한 posterior distribution : $$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_{nk}} . \qquad (9.38)$$

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi^{z_k}$$

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} N(\mathbf{x}|\mu_k, \Sigma_k)^{z_k}$$

Gaussian Mixtures Revisited

Log likelihood function의 기대값을 전개하면 다음과 같고,

$$E_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi)] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\{\ln \pi_k + \ln N(\mathbf{x}_n|\mu_{\mathbf{k}}, \Sigma_k)\}$$

얻어낸 식으로부터 MLE를 수행하여 파라미터를 구하면 동일한 결과를 얻을 수 있다.

$$\mu_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n$$

$$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

$$\pi_k = \frac{N_k}{N}$$

Gaussian Mixtures Revisited

Cf. Incomplete Data의 log likelihood function 기대값 전개 과정

$$E_{\mathbf{Z}}[\ln p(\mathbf{X},\mathbf{Z}|\theta)] = E_{\mathbf{Z}}[\sum_n \ln p(\mathbf{x}_n,\mathbf{z}_n|\theta)] = \sum_n E_{\mathbf{Z}}[\ln p(\mathbf{x}_n,\mathbf{z}_n|\theta)]$$

$$= \sum_{n=1}^{N} E_{\mathbf{Z}}[\ln(p(\mathbf{z}_n)p(\mathbf{x}_n|\mathbf{z}_n,\theta))] = \sum_{n=1}^{N} E_{\mathbf{Z}}[\ln[\prod_{k=1}^{K}(\pi_k N(\mathbf{x}_n|\mu_k,\Sigma_k))^{z_{nk}}]]$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} E_z[(z_{nk})\ln(\pi_k N(\mathbf{x}|\mu_k,\Sigma_k))] = \sum_{n=1}^{N}\sum_{k=1}^{K} E_z[z_{nk}]\ln(\pi_k N(\mathbf{x}_n|\mu_k,\Sigma_k))$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\{\ln\pi_k + \ln N(\mathbf{x}_n|\mu_\mathbf{k},\Sigma_k)\}$$

이때,

$$E[X+Y] = E[X] + E[Y]$$

$$p(\mathbf{x}_n|\mathbf{z}_n) = \prod_{k=1}^{K} N(\mathbf{x}_n|\mu_k,\Sigma_k)^{z_n k}$$

$$E[z_{nk}] = \frac{\sum_{\mathbf{z}_n} z_{nk}\prod_{k'}[\pi_{k'}N(\mathbf{x}_n|\mu_{k'},\Sigma_{k'})]^{z_{nk'}}}{\sum_{\mathbf{z}_n}\prod_j[\pi_j N(\mathbf{x}_n|\mu_j,\Sigma_j)]^{z_{nj}}}$$

$$= \frac{\pi_k N(\mathbf{x}_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N(\mathbf{x}_n|\mu_j,\Sigma_j)} = \gamma(z_{nk})$$

Relation to K-means

K-means vs EM Algorithm

· K-means

: hard assignment of data points to clusters (하나의 샘플이 오로지 하나의 클러스터에만 속하는 구조)

: 각 클러스터에 대한 분산도를 고려하지 않음 (유클라디안 방식의 평균과의 거리만 고려함)

· EM Algorithm

: soft assignment based on the posterior probabilities (각각의 클러스터에 대해 샘플이 속할 확률 값을 표현)

Relation to K-means

K-means 알고리즘은 GMM 알고리즘의 특별한 경우를 나타낸다.

: 각 Gaussian component들의 분산 값이 서로 동일하고 독립적이라고 가정하자. (eI) 그러면 x에 대한 확률 분포는,

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi\epsilon)^{1/2}} \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x} - \boldsymbol{\mu}_k\|^2\right\}. \qquad (9.41)$$

여기서 K개의 가우시안 분포를 고려하면 responsibility 함수는,

$$\gamma(z_{nk}) = \frac{\pi_k \exp\left\{-\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2/2\epsilon\right\}}{\sum_j \pi_j \exp\left\{-\|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2/2\epsilon\right\}}. \qquad (9.42)$$

이제 분산에서 $e \to 0$이라고 하면,

1) 평균만 의미가 있고 분산 값은 의미가 없어지게 된다.

2) responsibility 함수는 결국 binary indicator가 된다.

Log likelihood function의 기대값 역시 K-means에서 정의한 distortion measure와 계수를 제외하고 동일하게 계산된다.

$$E_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi)] \to -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\mathbf{x}_n - \mu_k\|^2 + const$$

Mixtures of Bernoulli Distributions

이산 분포인 베르누이 분포에 대한 혼합 분포 알아보기 (앞서 나왔던 모델들의 x는 모두 연속 변수)

D차원의 이항 변수 x_i를 가정$(i = 1, \ldots, D)$하고, 이를 베르누이 분포로 나타내면 다음과 같다.

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{i=1}^{D} \mu_i^{x_i}(1-\mu_i)^{(1-x_i)} \tag{9.44}$$

$$\text{where } \mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}} \text{ and } \boldsymbol{\mu} = (\mu_1, \ldots, \mu_D)^{\mathrm{T}}$$

이 때의 평균과 공분산은,

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} \tag{9.45}$$

$$\mathrm{cov}[\mathbf{x}] = \mathrm{diag}\{\mu_i(1-\mu_i)\}. \tag{9.46}$$

이제 여기서 K개의 베르누이 분포가 사용된다고 생각하면, 혼합 분포의 형태와 평균, 분산은 다음과 같다.

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\boldsymbol{\mu}_k) \tag{9.47}$$

$$\text{where } \boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}, \boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$$

$$\mathbb{E}[\mathbf{x}] = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k \tag{9.49}$$

$$\mathrm{cov}[\mathbf{x}] = \sum_{k=1}^{K} \pi_k \left\{\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^{\mathrm{T}}\right\} - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^{\mathrm{T}} \tag{9.50}$$

$$\text{where } \boldsymbol{\Sigma}_k = \mathrm{diag}\left\{\mu_{ki}(1-\mu_{ki})\right\}$$

Mixtures of Bernoulli Distributions

Data set $X = \{x_1, \ldots, x_N\}$이 주어질 경우의 log likelihood function,

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k) \right\}. \qquad (9.51)$$

마찬가지로 로그함수 안에 합의 형태를 한 식이 포함되어 있으므로, EM Algorithm을 도입해야 한다. (closed form이 아니므로)

EM Algorithm을 위해, K개의 상태를 나타내는 잠재 변수 $z = (z_1, \ldots, z_K)^T$를 도입해야 한다.

잠재 변수를 도입하면, 조건부 분포는 다음처럼 간단하게 기술할 수 있다.

$$p(\mathbf{x}|\mathbf{z}, \mu) = \prod_{k=1}^{K} p(\mathbf{x}|\mu_k)^{z_k}$$

(이때, 잠재 변수의 사전 확률 분포는 $p(z|\pi) = \prod_{k=1}^{K} \pi_k^{z_k}$)

Mixtures of Bernoulli Distributions

이제 EM Algorithm을 사용하기 위해 log likelihood function을 만들어야 한다.

$$\ln p(\mathbf{X}, \mathbf{Z}|\mu, \pi) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \left\{ \ln \pi_k + \sum_{i=1}^{D} [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki})] \right\}$$

기대값을 계산하면(E – step),

$$E_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\mu, \pi)] = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}) \left\{ \ln \pi_k + \sum_{i=1}^{D} [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki})] \right\}$$

$$\gamma(z_{nk}) = E[z_{nk}] = \frac{\sum_{\mathbf{z}_n} z_{nk} \prod_{k'} [\pi_{k'} p(\mathbf{x}_n|\mu_{k'})]^{z_{nk'}}}{\sum_{\mathbf{z}_n} \prod_j [\pi_j p(\mathbf{x}_n|\mu_j)]^{z_{nj}}} = \frac{\pi_k p(\mathbf{x}_n|\mu_k)}{\sum_{j=1}^{K} \pi_j p(\mathbf{x}_n|\mu_j)}$$

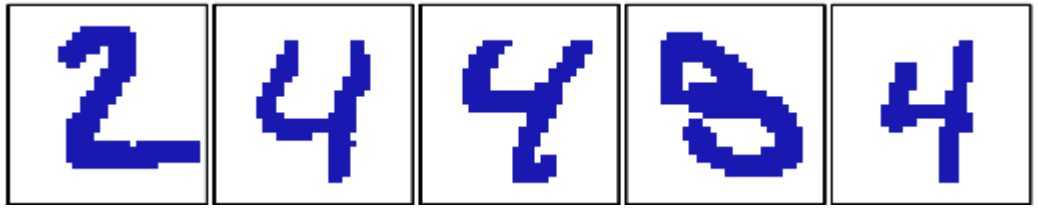
M – step에서의 추정에 대한 결과는 다음과 같다.

$$\mu_k = \bar{\mathbf{x}}_k$$

$$\pi_k = \frac{N_k}{N}$$

Mixtures of Bernoulli Distributions

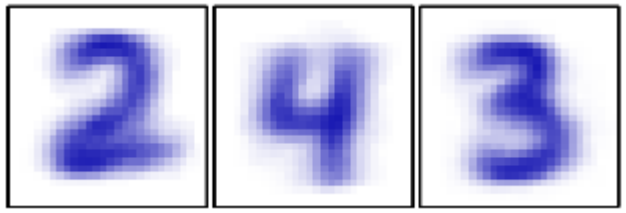
베르누이 혼합 분포 예제



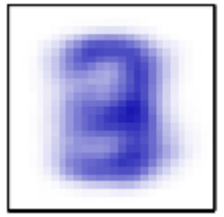
숫자 인식을 위한 5개의 데이터가 주어졌을 때,
2, 3, 4와 같은 값들을 결정할 수 있어야 한다.

EM 알고리즘을 이용하여 이를 분류해보자 (K=3을 가정) (2, 3, 4라는 숫자만을 갖고 있는 600개의 데이터를 활용)

1) 초기 $\pi_k = 1/K$로 시작하고, 이미지 벡터는 0과 1로 gray-scale을 표현 (둘을 구분하는 확률 값은 0.5로 정의)
2) EM을 수행하기 위해서는 해당 필드에 대한 확률의 기대값이 필요하므로, (0.25, 0.75) 사이의 범위에서 랜덤하게 선택 후 시작
3) Conjugate prior으로는 beta distribution을 사용



3개의 클러스터로 나누어진 후,
각각의 평균값으로 그림을 표현한 것



K=1인 경우
전체 샘플에 대한 평균값 (샘플 이미지의 평균 이미지)

EM for Bayesian Linear Regression

GMM과 같은 혼합 모델이 아닌 경우에 대한 EM 적용에 대해 – 베이지안 선형 회귀(Bayesian linear regression)를 예시로

· 목표 : 베이지안 선형 회귀에 사용되는 hyper-parameter α, β를 EM 알고리즘을 활용하여 예측
(보통 이 값은 사용자가 임의로 지정하지만, 주어진 샘플을 활용해 적절한 값을 선택하는 방법도 가능하기 때문)

· 모수로 사용했던 w가 잠재 변수의 역할을 수행하고, α, β가 파라미터로 사용된다.

· E-step : w에 대한 사후 분포를 구한다. (α, β는 고정)

· M-step : log likelihood function의 기대값을 최대화하는 α, β를 구한다.

$$\ln p(\mathbf{t}, \mathbf{w}|\alpha, \beta) = \ln p(\mathbf{t}|\mathbf{w}, \beta) + \ln p(\mathbf{w}|\alpha)$$

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} N(t_n|\mathbf{w}^T\phi(\mathbf{x}_n), \beta^{-1})$$

$$p(\mathbf{w}|\alpha) = N(\mathbf{w}|0, \alpha^{-1}\mathbf{I})$$

EM for Bayesian Linear Regression

최종적으로 log likelihood function의 기대값을 기술하면,

$$E[\ln p(\mathbf{t}, \mathbf{w}|\alpha, \beta)] = \frac{M}{2}\ln\left(\frac{\alpha}{2\pi}\right) - \frac{\alpha}{2}E[\mathbf{w}^T\mathbf{w}] + \frac{N}{2}\ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}\sum_{n=1}^{N}E[(t_n - \mathbf{w}^T\phi_n)^2]$$

이제 이 식을 미분하여 식을 얻어낼 수 있고, 수식을 전개하면 기존과 동일한 결과를 얻을 수 있다.

Summary :

- **E-Step**
  - responsibilities 계산 : 잠재변수 $\mathbf{w}$ 의 값을 결정한다.
  - $p(\mathbf{w}|\mathbf{t}, \alpha, \beta) = N(\mathbf{w}|\mathbf{m}, S)$
  - $m = \beta S\Phi^T\mathbf{t}$
  - $S^{-1} = \alpha\mathbf{I} + \beta\Phi^T\Phi$
- **M-Step**
  - $\alpha^{-1} = \frac{1}{M}(\mathbf{m}^T\mathbf{m} + Tr(S))$
  - $\beta^{-1} = \frac{1}{N}\sum_{n=1}^{N} t_n - \mathbf{m}^T\phi(\mathbf{x}_n)^2$

4

The EM Algorithm in General

The EM Algorithm in General

· EM 알고리즘은 잠재변수(latent variable)를 가지고 있는 확률 모델에서 MLE를 구하는 일반화된 기법

· 또한 변분 추론(variational inference)의 한 방식이다.

Variational Inference

: 추론의 목적은 데이터의 likelihood를 계산하는 것이고, 잠재변수의 posterior distribution $p(Z|X)$를 구하는 것이다.

: 하지만 정확한 분포를 알지 못하므로, 이를 최적화 문제로 바꾸어 최대한 비슷한 값을 구하고자 한다.

데이터 X에 대한 로그 주변부 확률분포는 다음과 같다.

$$\log p(\mathbf{X}) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z} - \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z}$$

$$= \mathcal{L}(q) + \mathrm{KL}(q(\mathbf{Z}) \| p(\mathbf{Z}|\mathbf{X}))$$

1) Z는 잠재 변수이고, $q(\cdot)$는 다룰 수 있는(tractable) Z를 확률변수로 갖는 임의의 확률분포이다.

2) $\mathcal{L}(q)$는 ELBO(Evidence Lower Bound) 또는 Variational Free Energy라고 한다.

3) 뒤의 항 KL은 분포 간의 유사도의 측도로 사용될 수 있는 KL-divergence이다.

The EM Algorithm in General

$$\mathcal{L}(q) + \mathrm{KL}(q(\mathbf{Z})\|p(\mathbf{Z}|\mathbf{X}))$$

$\log p(X)$가 고정된 상태에서 ELBO를 최대화하는 것은 KL-divergence의 최소화 문제와 동일하고,
KL-divergence가 최소화된다면 사후확률분포와 $q(Z)$가 유사해진다는 뜻이다. (알지 못하는 사후확률분포를 $q(Z)$를 통해서 추론)

이처럼 $q(\mathrm{Z})$를 도입하여 ELBO를 최대화함으로써 계산 불가능한 KL-divergence를 간접적으로 줄이는 방법을

"variational inference"라고 하며, 여기서 $q(\mathrm{Z})$를 variational distribution이라고 한다.

The EM Algorithm in General

잠재 변수 Z를 가진 모델의 관찰 데이터 X가 주어졌을 때, 이에 대한 결합 법칙은 다음과 같이 기술된다.

$$p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)$$

여기서 $p(X|\theta)$를 바로 구하는 것은 어렵지만, complete data의 likelihood function $p(X, Z|\theta)$는 쉽게 계산이 가능함

이제 잠재 변수 Z에 대한 함수이자, 다음을 만족하는 $q(Z)$를 도입한다.

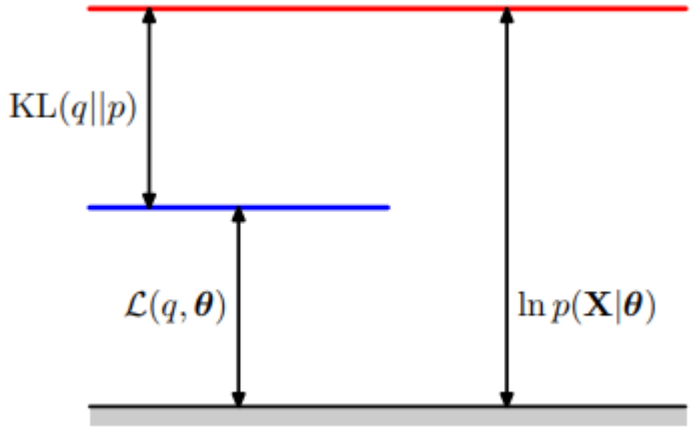
$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q\|p) \tag{9.70}$$

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \tag{9.71}$$

$$\mathrm{KL}(q\|p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}. \tag{9.72}$$

The EM Algorithm in General

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q||p) \qquad (9.70)$$



KL은 정의에 의해, $p = q$인 경우에만 0이고, 그 외에는 양수의 값을 갖게 된다.

따라서 $\mathcal{L}(q, \theta)$는 로그 가능도 함수의 lower bound

또한 $\mathcal{L}(q, \theta)$과 KL함수 모두에 q가 관여하고 있음을 알 수 있다. (q에 대해 상보적인 관계를 가짐)

KL을 최소화하는 q를 선택하면 이것이 $\mathcal{L}$을 최대화하기에, 임의의 q를 적절하게 선택하여 $\mathcal{L}$이 로그 가능도 함수와 같아지기를 원함

이때 KL은 $\mathrm{p} = q$일 때 해당 값이 0으로 최소가 되므로, $q(Z) = p(Z|X, \theta)$로 계산하면 된다.

여기서, 위에서 제시된 식은 2개의 텀으로 나누어져 있고, 이를 EM 알고리즘 각 2개의 단계와 연결지어 생각할 수 있다.

The EM Algorithm in General

· **E-step**

파라미터 θ를 특정 값으로 고정하고(θ^{old}), 이 값을 활용하여 Z의 사후 분포(responsibility)를 계산하는 과정이 필요했음

$\mathcal{L}$ 함수를 통해 생각해보면, $\mathcal{L}(q,\theta)$에서 θ는 고정된 값으로 생각할 수 있다.

따라서 $\mathcal{L}(q,\theta^{old})$의 값을 최대화하는 q함수를 선택하는 단계가 된다. ($q(Z) = p(Z|X,\theta^{old})$

· M-step

잠재 변수 Z를 고정시킨 다음, 새로운 파라미터 θ^{new}를 MLE를 이용하여 추론하는 단계

$\mathcal{L}$ 함수를 통해 생각해보면, $q(Z)$를 고정시킨 상태에서 새로운 파라미터를 추론하는 것과 같음

이때 최대화하는 함수는 $\mathcal{L}(q^{fixed},\theta)$가 되는데, 이를 확인하기 위해 $\mathcal{L}$ 함수를 전개해서 살펴보면 다음과 같다.

$$q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\theta^{old})$$

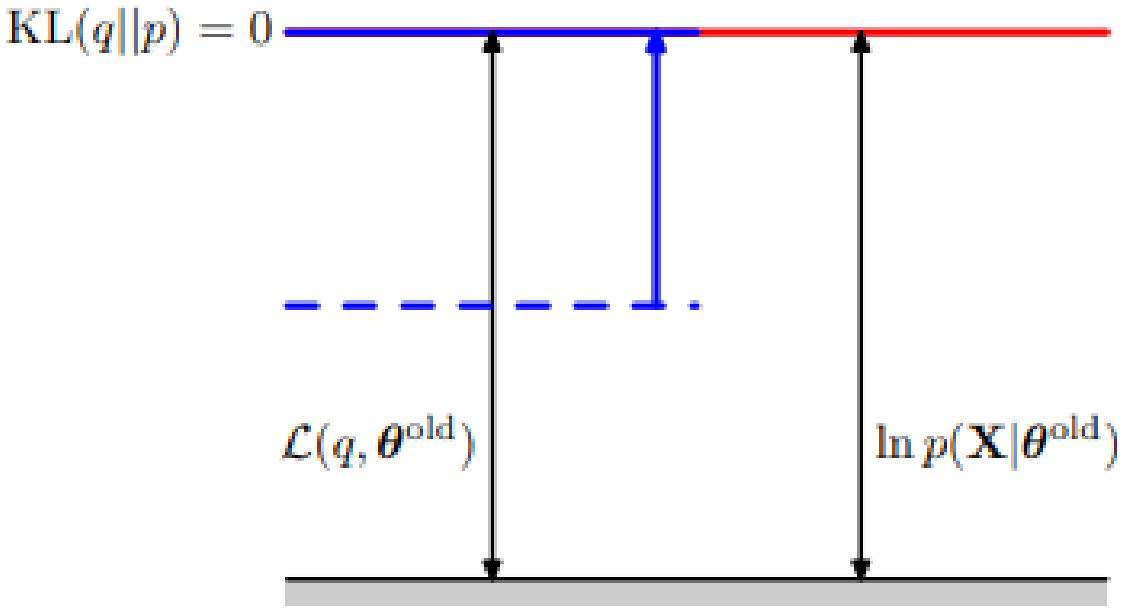
$$L(q,\theta) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{old}) \ln p(\mathbf{X},\mathbf{Z}|\theta) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{old}) \ln p(\mathbf{Z}|\mathbf{X},\theta^{old}) = Q(\theta,\theta^{old}) + const$$

· 이때 Q는 수식을 정리하기 위해 새롭게 정의한 함수인데, 자세히 보면 $\ln p(X,Z|\theta)$ 에 대한 기댓값이다.

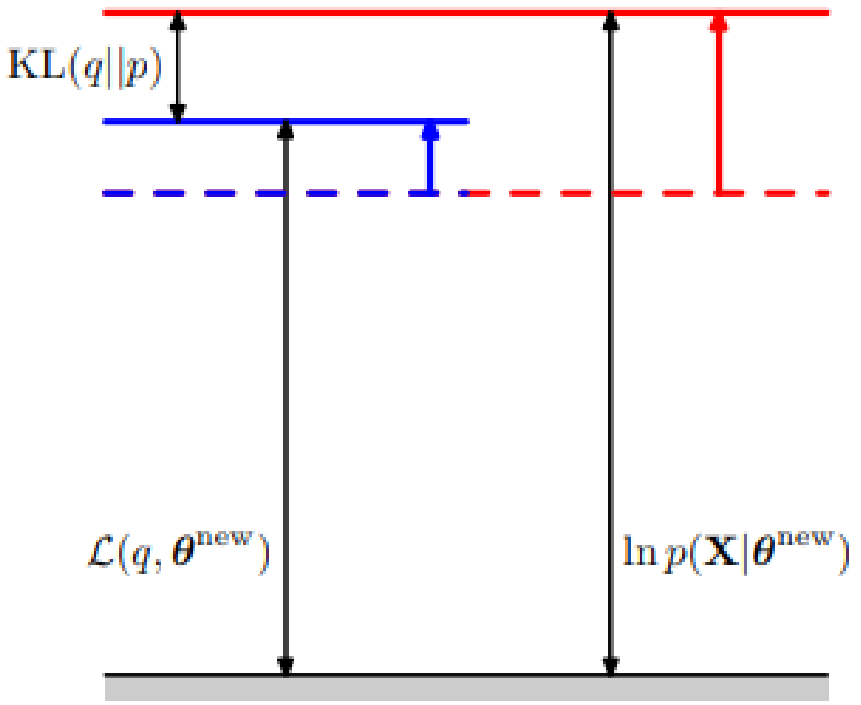
(앞 단원에서 likelihood function 대신 Q값을 사용할 수 있었던 이유)

The EM Algorithm in General

Summary : EM Algorithm



E-step

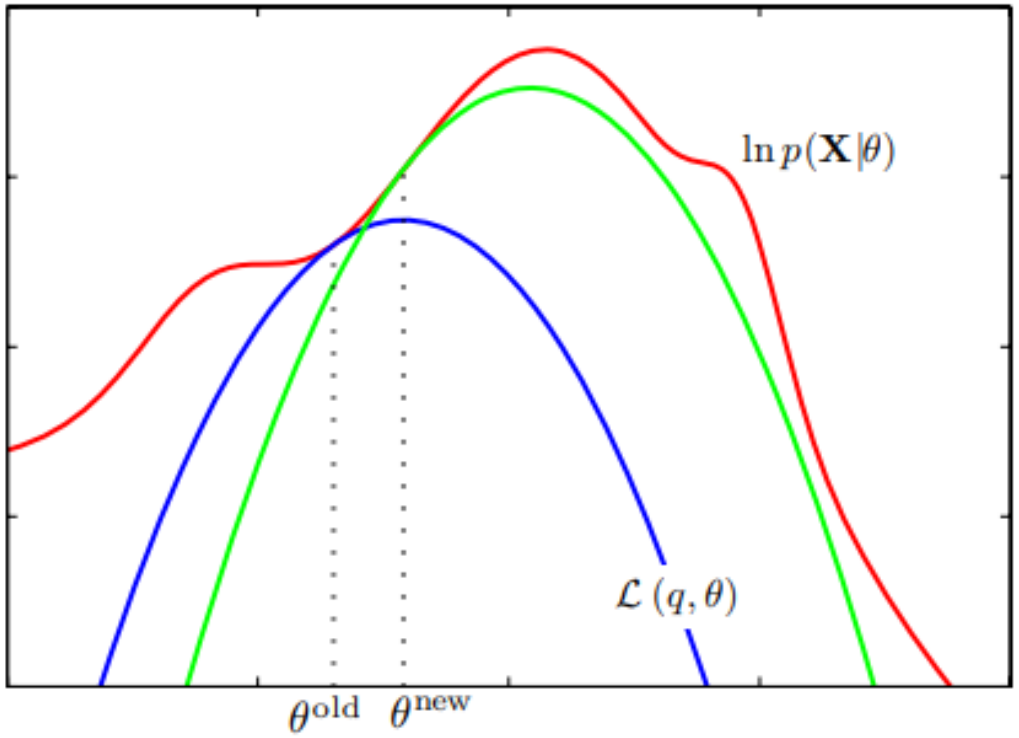


M-step

Review : EM Algorithm

EM 알고리즘은 반복적인 과정을 통해 MLE를 구하는 알고리즘

: q 함수를 도입하여 일반화된 EM 알고리즘을 도출할 수 있다.



E-step

1) 임의로 고정한 θ^{old}로부터 EM 알고리즘 시작

2) 해당 지점에서 likelihood function과 최대한 근사한 $\mathcal{L}$ 함수를 만든다.

M step

3) 얻어진 $\mathcal{L}$ 함수를 최대화하는 새로운 파라미터를 선정한다.

Then..

4) 수렴 조건을 만족할 때까지 반복

Case I : independent, identically distributed data set X

데이터 집단 X와 이에 대응되는 잠재변수 Z가 모두 서로 독립적인 경우

: E-step에서 사용될 Z에 관한 함수는 다음과 같이 기술할 수 있다.

$$p(\mathbf{Z}|\mathbf{X},\theta) = \frac{p(\mathbf{X},\mathbf{Z}|\theta)}{\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\theta)} = \frac{\prod_{n=1}^{N} p(\mathbf{x}_n,\mathbf{z}_n|\theta)}{\sum_{\mathbf{Z}} \prod_{n=1}^{N} p(\mathbf{x}_n,\mathbf{z}_n|\theta)} = \prod_{n=1}^{N} p(\mathbf{z}_n|\mathbf{x}_n,\theta)$$

Z의 사후 분포는 각각의 데이터들의 요소 z_n에 대응되는 확률 값의 곱으로 표현된다.
(GMM에서 각각의 데이터에 대해 z_n의 responsibility 값을 곱한 것과 같은 맥락)

또한 EM 알고리즘에 $p(\theta)$를 추가로 적용하여 베이지안 모델을 만들 수 있다.

$$\ln p(\theta|\mathbf{x}) = \ln p(\theta,\mathbf{x}) - \ln p(\mathbf{x})$$

$$\ln p(\theta|\mathbf{X}) = L(q,\theta) + KL(q\|p) + \ln p(\theta) - \ln p(\mathbf{X}) \geq L(q,\theta) + \ln p(\theta) - p(\mathbf{X})$$

여기서 $p(X)$는 상수 값이 되고, 식 자체는 크게 달라지지 않는다. 최대화를 진행하는 과정을 생각해보면,
새롭게 추가된 것은 $p(\theta)$이므로 q를 선택하는 것은 기존의 EM 알고리즘과 차이가 없음 (q는 $\mathcal{L}$과 KL 함수에서만 등장하므로)
M-step에서는 값이 바뀌므로 조금 달라지지만, 실제 식 적용에는 큰 변화가 없고 MAP와 같은 효과를 갖는다.

Case II : GEM (Generalized EM)

EM 알고리즘은 잠재 변수가 주어진 경우 MLE를 쉽게 적용할 수 있게 해주는 알고리즘이지만,
복잡한 모델이 주어졌을 때, E 단계와 M 단계를 처리하기 힘든 경우도 존재하며, 이를 위해 2가지 확장 EM이 존재한다.

GEM (Generalized EM)

: M-step은 E-step과 상관없이 별도의 방식으로 구현 가능

: E-step에서 얻어진 $\mathcal{L}$ 함수에 대해 이 함수값을 최대로 하는 파라미터를 꼭 구할 필요는 없음

: 단순하게 $\mathcal{L}$ 함수의 값이 더 커지는 방향으로의 파라미터를 선택하는 방법 (SGD)

Case III : Incremental EM

Incremental EM

: 데이터 자체가 분산화되어 있어서 배치(batch) 방식의 업데이트가 불가능한 경우에 사용

: 데이터를 스트리밍 방식으로 처리 가능하다

: E-step에서 여러 개의 데이터를 다루는 것이 아니라, 하나의 데이터를 다루는 것처럼 수정된다.

: E-step, M-step 모두 데이터 크기와 무관하게 고정된 계산량이 필요함

(한 번 데이터를 거치면 완료되는 것이 아니라, 데이터 입력에 의해 반복되면서 변경되므로)

$$\mu_k^{new} = \mu_k^{old} + \left(\frac{\gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})}{N_k^{new}} \right) (\mathbf{x}_n - \mu_k^{old})$$

$$N_k^{new} = N_k^{old} + \gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})$$

· 배치(batch) : 데이터를 실시간으로 처리하는 것이 아니라, 일괄적으로 모아서 처리하는 작업을 의미함

감사합니다