EM Algorithm - 2021년 P-sat 논문스터디 2회차

권남택, 오정민, 유경민

2021년 1월 26일

EM Algorithm에 대한 전반적인 내용을 소개하는 자료입니다.

1. MLE (Maximum Likelihood Estimation)

mle는 가능도(likelihood) 함수 $\mathcal{L}(\theta)$ 를 최대로하는 모수 θ를 추정치로 선택하는 방법이다. 가능도는 주어진 표본에서 가장 가능한(likely) 모수를 추정하는 척도이다. 확률변수 X가 θ에 대해 $P(X)$의 확률분포를 가지고, 표본으로 x가 나왔을 때, 가능도 함수는 다음과 같이 표현된다.

$$P(X = x|\theta) = P(x_1, x_2, ...|\theta) = P(x_1|\theta) \times P(x_2|\theta) \times ...P(x_n|\theta) = \mathcal{L}(\theta|x)$$

1.1 다변량 정규분포 MLE

- **일변량 정규분포 확률밀도함수 :** $f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$
- **다변량 정규분포 확률밀도함수 :** $f(\mathbf{x}|\mu, \sigma) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\}$
- likelihood function

$$\begin{aligned}
l(\mu, \Sigma|\mathbf{x}) &= \ln \prod_{i=1}^{n} f(\mathbf{x_i}|\mu, \Sigma) \\
&= \ln \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x_i}-\mu)^T\Sigma^{-1}(\mathbf{x_i}-\mu)\right\} \\
&= -\frac{n}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\left\{(\mathbf{x_i}-\mu)^T\Sigma^{-1}(\mathbf{x_i}-\mu)\right\} + c \\
&\propto -\frac{n}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}((\mathbf{x_i}-\mu)^T\Sigma^{-1}(\mathbf{x_i}-\mu)) \quad (\because (\mathbf{x_i}-\mu)^T\Sigma^{-1}(\mathbf{x_i}-\mu)\ is\ scalar) \\
&= -\frac{n}{2}\ln|\Sigma| - \frac{1}{2}\mathrm{tr}(\Sigma^{-1}\sum_{i=1}^{n}(\mathbf{x_i}-\mu)(\mathbf{x_i}-\mu)^T)
\end{aligned}$$

- mle

$$\frac{d}{d\mu} l(\mu, \Sigma)$$

$$= -\frac{1}{2} \times 2 \times (-1) \times \sum_{i=1}^{n} \Sigma^{-1}(\mathbf{x_i} - \mu) = 0$$

$$\rightarrow \sum_{i=1}^{n} \mathbf{x_i} - n\mu = 0$$

$$\rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x_i}$$

$$\frac{d}{d\Sigma^{-1}} l(\mu, \Sigma) = \frac{n}{2}\Sigma - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x_i} - \mu)(\mathbf{x_i} - \mu)^T (-\mathbf{ln}|\Sigma| = \mathbf{ln}|\Sigma^{-1}|)$$

$$\rightarrow \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x_i} - \mu)(\mathbf{x_i} - \mu)^T$$

** 미분 참고 $\frac{\partial \mathbf{w}^\mathsf{T} A \mathbf{w}^\mathsf{t}}{\partial \mathbf{w}} = 2Aw$ $\frac{\partial \mathbf{tr}(\mathbf{x}^t A \mathbf{x})}{\partial A} = (\mathbf{x}\mathbf{x}^\mathsf{T})^T = \mathbf{x}\mathbf{x}^T$,

1.2 다항분포 MLE

- 다항분포 : 여러개의 값을 가질 수 있는 독립 확률변수들에 대한 확률분포 (일반적으로 다항분포는 확률 변수 값이 횟수인데, 경우에 따라 독립시행에서 나타나는 값 자체를 가르키기도 한다. 엄밀하게 말하면 이는 categorical 분포인데, 이번에 우리가 다항분포로 다루는 경우는 이것이다.)

ex. X = (0,0,1,0,0,0) → 6개의 값 나타날 수 있는데 그 중 세번째 값이 나타남 $\sum_k x_k = 1, P(X|\mu) = \prod_{k=1}^{K} \mu_k \quad such\ that\ \mu_k \geq 0, \sum_k \mu_k = 1$

- likelihood function N개의 관측값 가진 데이터 집합에서 가능도 함수 는 다음과 같이 정의된다.

$$\mathcal{L}(\mu|\mathcal{D}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{\sum_{n=1}^{N} x_{nk}} = \prod_{k=1}^{K} \mu_k^{m_k},\ m_k = \sum_{n=1}^{N} x_{nk}$$

- MLE MLE를 구하기 위하여 max $\mathcal{L}(\mu|\mathcal{D})\ \ subject\ \ to\ \mu_k \geq 0, \sum_k \mu_k = 1$ 식을 풀어야하는데, 이는 제약식을 가진 최대화 문제이므로 라그랑주 승수를 이용하여 제약식이 없는 간단한 형태로 만들어 최대화를 한다.

$$l(\mu, m, \lambda) = \sum_{k=1}^{K} m_k \mathbf{ln}\, \mu_k + \lambda(\sum_{k=1}^{K} \mu_k - 1)$$

$$\frac{dl}{d\mu_k} = \frac{m_k}{\mu_k} + \lambda = 0 \rightarrow \mu_k = -\frac{m_k}{\lambda}$$

$$since \sum_{k=1}^{K} \mu_k = 1,\ \sum_{k=1}^{K} -\frac{m_k}{\lambda} = 1 \rightarrow \sum_{k=1}^{K} m_k = N = -\lambda$$

$$\therefore \mu_k = \frac{m_k}{N} = \frac{\sum_{k=1}^{K} x_{nk}}{N}$$

1.3 MLE의 한계

왼쪽 그림과 같이 샘플들이 3개의 다른 정규분포로부터 추출되었을때, 기존의 하나의 정규 분포로는 해당 분포를 잘 설명하지 못한다. 따라서 3개의 정규 분포를 혼합하여 해당 샘플들에게 맞춘 분포를 만들어낼 필요가 있다.

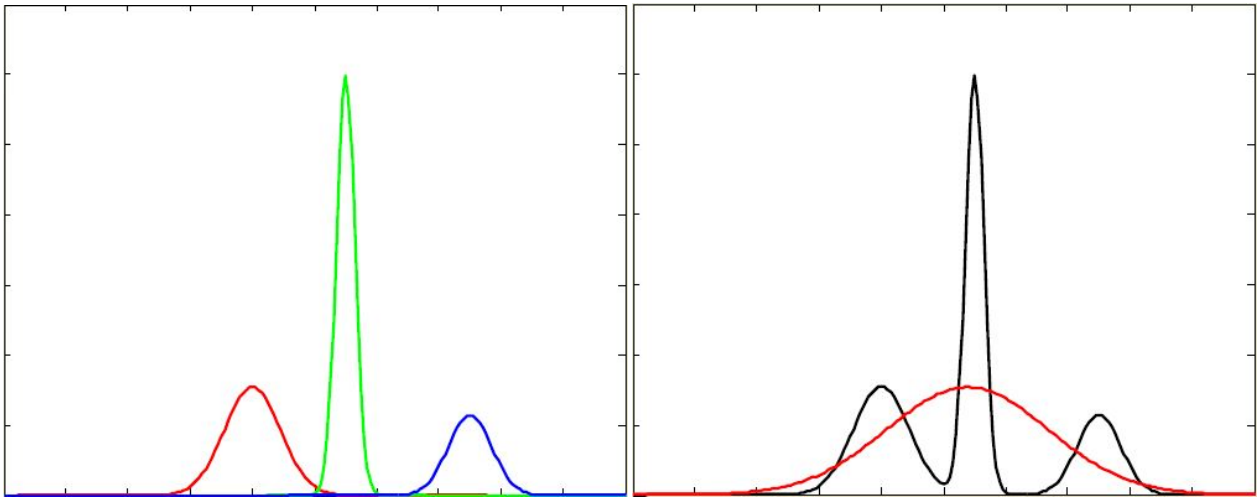


Figure 1: Mixture Distribution

따라서 우리는 혼합분포 $P(x) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(x|\mu_k, \sigma_k)$ 를 정의한다. 여기서 π_k는 mixing coefficient, $\mathcal{N}(x|\mu_k, \sigma_k)$는 mixture component라고 합니다. mixing coefficient π_k K개의 mixture component 중 k번째의 mixture component가 선택될 확률이다. 즉, π_k는 확률의 성질 $\sum_{k=1}^{K} \pi_k = 1, 0 \leq \pi_k \leq 1$을 만족하는 weight의 기능을 한다. 어떤 mixture component에 속하는지에 대하여 K가지의 선택지가 있고, 각각의 확률이 π_k이기에 이를 다항분포로 생각할 수 있다. 따라서 우리는 다항분포를 따르는 새로운 확률 변수 Z를 정의하여 위에서 언급한 혼합분포를 $P(x) = \sum_{k=1}^{K} P(z_k)P(x|z_k)$로 다시 적을 수 있다. Z는 실제로 볼 수 없지만 필요해서 가정하는 변수인 잠재변수이고, z_k는 k번째 mixture component에 속하는지 여부에 대한 것으로 다음과 같은 성질을 갖는다.

$$z_k \in \{0, 1\}, \sum_k z_k = 1, P(z_k = 1) = \pi_k, \sum_k \pi_k = 1, 0 \leq \pi_k \leq 1 P(Z) = \prod_{k=1}^{K} \pi_k$$

$$P(X|z_k = 1) = \mathcal{N}(x|\mu_k, \Sigma_k) \rightarrow P(X|Z) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k}$$

만약 우리가 각 관측치들이 어떤 mixture component에 속하는지 알고 있다면, 다음과 같은 likelihood function을 가진다.

$$\ln P(X|\pi, \mu, \Sigma) = \ln \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k} = \sum_{n=1}^{N} \sum_{k=1}^{K} \ln \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_k}$$

하지만 우리가 각 관측치들이 어떤 mixture component에 속하는지에 대해 알 수 없는 경우 mle를 구하기는 쉽지 않다. 이 경우 가능한 모든 잠재변수 값에 대해 분포를 고려해주어야하고, marginalize 하는 과정에서 로그 함수 내에 $\sum$에 대한 수식이 포함된다. 따라서 mle로 최대화 되는 지점을 찾기위한 계산이 어려워진다.

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

mle로 계산 가능하려면 잠재변수 Z가 관찰된 경우 밖에 없으므로 maximization 단계 이전에 관측치가 주어졌을 때, 관측치가 k번째 component에 속할 확률($p(z_k = 1|\mathbf{x})$)에 대해 추정하는 expectation 단계가 필요하다.

정리하면,

- 잠재변수가 주어진 모델의 가능도 함수는 다음과 같음

$$\ln P(X|\theta) = \ln\{\sum_{Z} P(X, Z|\theta)\}$$

- 모든 관측치에 대해 잠재변수 Z가 주어진 경우 가능도 함수는 $\ln P(X, Z|\theta)$가 되고, MLE 계산이 가능하다.
- X에 대해서만 주어지고 잠재변수 Z가 주어지지 않은 경우 가능한 모든 잠재변수의 값에 대한 분포를 고려해야 하므로 가능도 함수는 $\ln\{\sum_{Z} P(X, Z|\theta)\}$으로 주어진다. 로그 내에 sum에 대한 수식이 존재하여 계산하기 어렵기에 Z에 대한 확률값을 추론하는 Expectation 과정을 거친 후, 얻어진 확률 값을 이용하여 가능도 함수를 구하는 방법이 고안되었다.

2. EM 알고리즘으로 하는 추정

EM 알고리즘의 목표는 잠재 변수를 가지고 있는 모델들의 최대 가능도 해를 찾는 것이다. 관측 데이터들을 행렬 X, 잠재변수들을 행렬 Z, 모든 모델 매개변수들의 집합을 θ로 지칭한다면 로그 가능도 함수를 다음과 같이 표현 가능하다. ($\{\mathbf{X}\}$를 불완전한 데이터 집합(imcomplete data set),$\{\mathbf{X}, \mathbf{Z}\}$를 완전한 데이터 집합(complete data set)이라 부르며 $q(\mathbf{Z})$는 Z에 대한 pdf이다.)

$$ln\, p(\mathbf{X}|\theta) \;=\; ln\,\{\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)\} \;=\; ln\,\{\sum_{\mathbf{Z}} q(\mathbf{Z})\frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})}\} \;\geq\; \sum_{\mathbf{Z}} q(\mathbf{Z})\, ln\, \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})}$$

위의 식에서 부등호의 경우 Jensen's inequality를 사용하면 된다. (ln 함수는 concave한 함수이고 concave한 함수에 대한 Jensen's inequality는 $\varphi(\frac{\sum a_i x_i}{\sum a_i}) \geq \frac{\sum a_i \varphi(x_i)}{\sum a_i}$)

2.1 Lower Bound

위에서 언급한 식에 조금 더 변형을 가해보자.

$$ln\, p(\mathbf{X}|\theta) \;\geq\; \sum_{\mathbf{Z}} q(\mathbf{Z})\, ln\, \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \;=\; \sum_{\mathbf{Z}} q(\mathbf{Z})\, ln\, \frac{p(\mathbf{Z}|\mathbf{X}, \theta)p(\mathbf{X}|\theta)}{q(\mathbf{Z})}$$

$$= \sum_{\mathbf{Z}} \{\, q(\mathbf{Z})\, ln\, \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} + q(\mathbf{Z})\, ln\, p(\mathbf{X}|\theta)\,\} = \; ln\, p(\mathbf{X}|\theta) \;+\; \sum_{\mathbf{Z}} \{\, q(\mathbf{Z})\, ln\, \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})}\}$$

$$= \; ln\, p(\mathbf{X}|\theta) \;-\; \sum_{\mathbf{Z}} \{\, q(\mathbf{Z})\, ln\, \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)}\}$$

여기서 주목해야할 부분은 바로 처음에 나온 식과 마지막에 나온 식이다. 우리는 결국 $ln\, p(\mathbf{X}|\theta)$에 대한 lower bound를 규정할 수 있게 되었다. 이를 토대로 보았을 때 결국 부등호의 요인이 되는 인자는 바로 $\sum_{\mathbf{Z}} \{\, q(\mathbf{Z})\, ln\frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X},\theta)}\}$ 라는 것을 알 수 있다. 이 값이 0이면 등호가 성립이 될 것이다.

2.2 KL divergence (쿨백-라이블러 발산)

위에서 우리는 부등호의 요인이 되는 인자가 $\sum_{Z}\{q(Z)\,ln\frac{q(Z)}{p(Z|X,\theta)}\}$임을 알 수 있었다. 이 식에 대해서 좀 더 알아보자. 해당 식은 쿨백-라이블러 발산 (Kullback-Leiber divegence)의 꼴이 되는데 KL divergence 식의 일반적인 형태는 다음과 같다.

$$KL(P||Q) = \sum_{i} P(i)\,ln(\frac{P(i)}{Q(i)})$$

위의 식에서 $P(i)$자리에 $q(Z)$, $Q(i)$자리에 $p(Z|X,\theta)$를 대입하면 $\sum_{Z}\{q(Z)ln\frac{q(Z)}{p(Z|X,\theta)}\}$식이 구해짐을 알 수 있다. 즉, $KL(q(Z)||p(Z|X,\theta))$이 된다.

KL divergence는 두 확률분포의 차이를 계산하는 데에 사용하는 함수이다. KL divergence값이 0에 가까울 수록 두 확률분포는 유사하다는 의미가 된다.

KL divergence의 특성에는 크게 2가지가 있는데, 1. $KL(P||Q) \geq 0$ 2. $KL(P||Q) \neq KL(Q||P)$(Non-symmetric)

다시 생각해보면 KL divergence의 첫번째 특성 때문에 우리는 $\sum_{Z}\{\,q(Z)\,ln\frac{q(Z)}{p(Z|X,\theta)}\} \geq 0$임을 알 수 있고 $ln\,p(X|\theta) - \sum_{Z}\{\,q(Z)\,ln\frac{q(Z)}{p(Z|X,\theta)}\}$가 $ln\,p(X|\theta)$의 lower bound가 맞다는 것을 알 수 있다.

2.3 EM 알고리즘 (Expectation Maximization Algorithm)

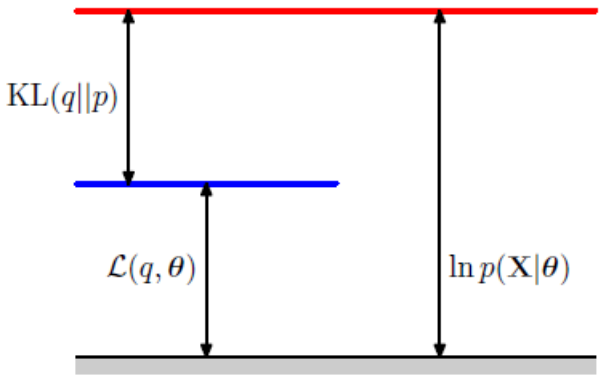
본격적인 EM 알고리즘을 다루기 전에, 표기의 간편화를 위해 다음과 같이 식을 정의하도록 하겠다.

$$\mathcal{L}(q,\theta) = \sum_{Z} q(Z)\,ln\,\frac{p(X,Z|\theta)}{q(Z)}$$
$$KL(q||p) = KL(q(Z)||p(Z|X,\theta))$$

이를 이용해서 정리해보면 결국 다음과 같은 식이 정리된다.

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

이를 보다 직관적으로 알기 쉽게 그림으로 표현하면 다음과 같다.



위의 그림을 보면 알 수 있듯이 결국 $KL(q||p)$를 최소화하는 q함수를 선택하면 이 q가 $\mathcal{L}(q,\theta)$를 최대화하게 된다. 우리가 결국 하고 싶은 것은 임의의 q함수를 적절하게 선택하여 $\mathcal{L}(q,\theta)$를 최대한 크게 만들고 싶은 것이고 최종적으로 이 $\mathcal{L}(q,\theta)$가 $ln\, p(\mathsf{X}|\theta)$와 같아지게 하고 싶은 것이다.

위의 KL divergence의 특성에서도 언급을 했지만 결국 $KL(q||p)$를 최소화 한다는 것은 $KL(q||p)$를 0으로 만들어 준다는 것이고 이는 $q=p$가 됨을 의미한다. 즉, $q(\mathsf{Z})=p(\mathsf{Z}|\mathsf{X},\theta)$가 되도록 만들어 주면 된다.

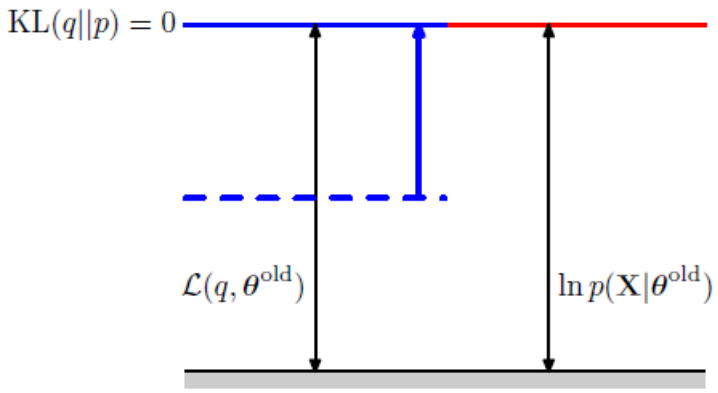
EM 알고리즘은 2단계의 과정을 반복적으로 수행하면서 MLE를 구하는 방법이다. 크게 E-Step 과 M-Step이 있다.

2.3.1 E-Step

E-step은 θ를 고정시킨 채로 $\mathcal{L}(q,\theta)$의 값을 $q(\mathsf{Z})$에 대해 최대화하는 단계이다.

먼저 파라미터 θ를 고정하고 이를 θ^{old}라 한다. 이 고정된 θ^{old}값을 활용하여 $\mathcal{L}(q,\theta^{old})$를 최대화하는 $q(\mathsf{Z})$를 찾으면 된다. 위에서도 언급했지만 결국 $q(\mathsf{Z})=p(\mathsf{Z}|\mathsf{X},\theta^{old})$를 구하면 된다.

이를 그림으로 표현하면 다음과 같다.



$q(\mathsf{Z})=p(\mathsf{Z}|\mathsf{X},\theta^{old})$가 되면 KL divergence 값이 0이 되어 $\mathcal{L}(q,\theta^{old})$와 로그 가능도 함수인 $ln\, p(\mathsf{X}|\theta^{old})$ 값이 같아짐을 알 수 있다.

2.3.2 M-Step

M-step은 $q(\mathsf{Z})$를 고정시킨 채로 $\mathcal{L}(q,\theta)$의 값을 θ에 대해 최대화하는 단계이다. 여기서 $q(\mathsf{Z})$값은 직전 E-step에서 구한 $p(\mathsf{Z}|\mathsf{X},\theta^{old})$가 된다. 이를 간단하게 전개해 보면 다음과 같다.

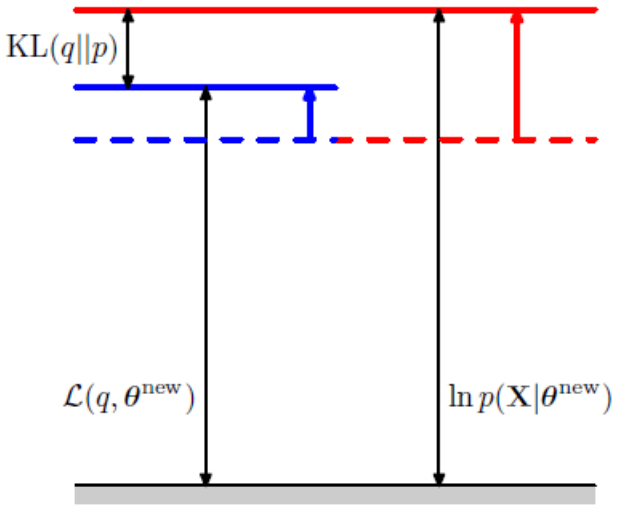
$$\mathcal{L}(q,\theta)=\sum_{\mathsf{Z}} q(\mathsf{Z})\, ln\, \frac{p(\mathsf{X},\mathsf{Z}|\theta)}{q(\mathsf{Z})}=\sum_{\mathsf{Z}} p(\mathsf{Z}|\mathsf{X},\theta^{old})\, ln\, p(\mathsf{X},\mathsf{Z}|\theta)-\sum_{\mathsf{Z}} p(\mathsf{Z}|\mathsf{X},\theta^{old})\, ln\, p(\mathsf{Z}|\mathsf{X},\theta^{old})$$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

여기서 상수는 분포 q의 엔트로피에 해당되어, θ에 독립적이기 때문에 상수로 표기하였다. $\mathcal{Q}(\theta,\theta^{old})$는 $ln\, p(\mathsf{X},\mathsf{Z}|\theta)$의 기대값임을 알 수 있다. 결국 M-Step에서 최대화되는 값은 완전 데이터 로그 가능도 함수의 기대값이다.

즉 수정된 매개변수 추정값 θ^{new}는 다음과 같다.

$$\theta^{new}=\underset{\theta}{\text{argmax}}\,\mathcal{Q}(\theta,\theta^{old})=\underset{\theta}{\text{argmax}}\,E_{q(Z)}(ln\, p(\mathsf{X},\mathsf{Z}|\theta))$$

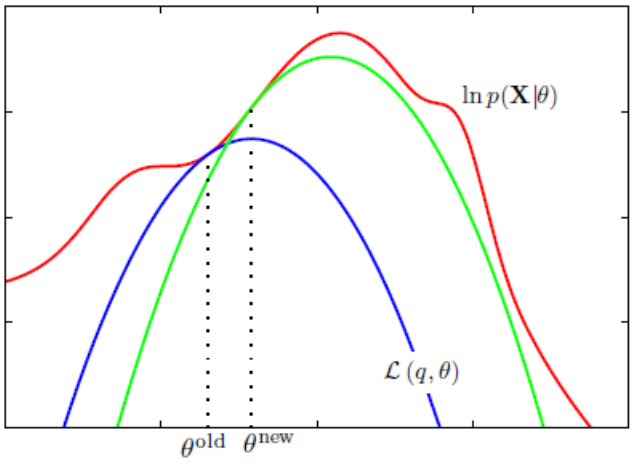
이를 그림으로 표현하면 다음과 같다.



분포 $q(\mathbf{Z})$는 새 값이 아닌 이전의 매개변수들을 바탕으로 결정되었고, M단계 동안 고정되어 있을 것이기 때문에 새로운 $p(\mathbf{Z}|\mathbf{X}, \theta^{new})$ 값과 다르게 된다. 따라서 KL divergence 값은 0 이 아닌게 된다. 따라서 로그 가능도 함수의 증가분은 하한의 증가분 보다 커진다.

2.3.3 EM 알고리즘 정리

다음의 그림과 함께 EM 알고리즘의 관한 내용을 요약해보면 크게 4단계로 분류가 가능하다.



1. θ^{old}의 초깃값을 설정한다.
2. E-Step $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$를 계산한다.

그 결과 로그 가능도(빨간선)와 접하는 하한값 $\mathcal{L}(q, \theta^{old})$를 얻게 된다(파란색)

3. M-Step 다음처럼 주어지는 θ^{new}를 계산한다.

$$\theta^{new} = \underset{\theta}{\text{argmax}}\, \mathcal{Q}(\theta, \theta^{old})$$

4. 수렴 조건을 만족할 때까지 2,3 단계를 반복한다.

이러한 과정을 반복하다 보면 알고리즘은 언젠가 local optimum으로 수렴하게 된다.

3. EM 알고리즘의 적용

EM 알고리즘이 가장 많이 쓰이는 예시들을 두 가지 정도 소개하고자 한다. 먼저 클러스터링에서 K-Means 알고리즘이다.

3.1 K-means 알고리즘

3.1.1 K-means vs GMM

먼저 명확히 해야할 것은 K-means 알고리즘은 가우시안 혼합모형(Gaussian Mixture Model, GMM)의 특이 케이스로 이해될 수 있다는 점이다.

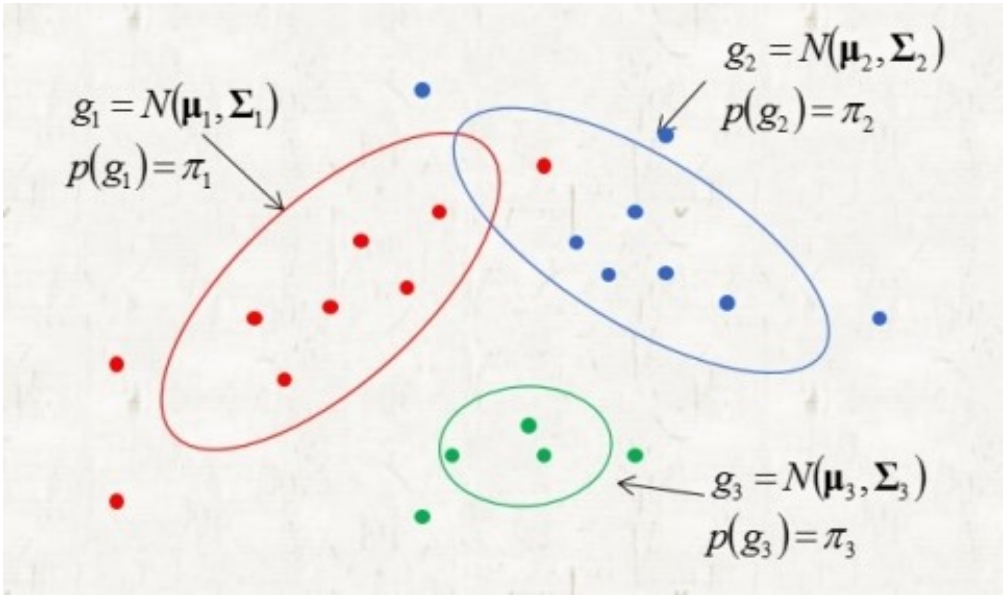


Figure 2: GMM Clustering

GMM은 저렇게 타원형의 군집을 찾아낸다. 타원형의 군집이 구성되는 이유는 각각의 군집별로 공분산 구조를 다르게 생각하기 때문이다. 만약 여기서 우리가 공분산 구조가 $\Sigma_k = \epsilon I$로 다 동일하다고 생각하면, 구(sphere)형의 구조를 찾아내기에 적합한 모델이 될 것이다. 결국 K-means는 GMM에서 공분산 구조가 단위행렬 I가 되는 경우라고 생각해도 된다.

또한 자세히 설명하기에는 시작이 부족하지만, GMM은 베이지안의 관점에서 문제를 바라보기 때문에, 경계값을 명확하게 어떤 군집으로 배정하는 것이 아니라 확률을 부여할 수 있게 된다. 따라서 GMM은 특정 클러스터로 배정하기 어려운 관측값에 확률을 부여함으로써 Soft Assignment를 가능하게 한다. 하지만 K-means는 soft assignment를 고려하지 않고, hard assignment만 고려하게 된다는 차이가 존재한다.

그렇다면 우리는 EM을 적용할 때, 가우시안 혼합분포에서 시작해서 문제를 간단하게 변형시켜 K-menas를 풀어내려 한다.

3.1.2 K-means by EM

Log-Likelihood for entire data

$$lnP(X|\pi, \mu, \Sigma) = \sum_{j=1}^{K} ln\{\sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)\}$$

EM Algorithm for GMM

$$\text{Expectation} \quad : \quad \gamma(z_{nk}) = \frac{P(z_k=1)P(x|z_k=1)}{\sum_{i=1}^{K} P(z_i=1)P(x|z_i=1)} = \frac{\pi_k N(x|\mu_k,\Sigma_k)}{\sum_{i=1}^{K}\pi_i N(x|\mu_i,\Sigma_i)}$$

$$\text{Maximization} \quad : \quad \hat{\mu_k} = \frac{\sum_{n=1}^{N}\gamma(z_{nk})x_n}{\sum_{n=1}^{N}\gamma(z_{nk})}, \qquad \hat{\Sigma_k} = \frac{\sum_{n=1}^{N}\gamma(z_{nk})(x_n-\hat{\mu_k})(x_n-\hat{\mu_k})^T}{\sum_{n=1}^{N}\gamma(z_{nk})}, \qquad \hat{\pi}_k = \frac{\sum_{n=1}^{N}\gamma(z_{nk})}{N}$$

이전에 정의한 다변량정규분포의 MLE와 다항분포의 MLE를 계산하는 방법을 적용하면 다음과 같이 각각의 step을 요약할 수 있다. 하지만 여기서 GMM을 K-means 문제로 단순화하게 되면, maximization step에서 공분산 Σ_k와 각각의 π_k를 정의하는데에 필요한 계산을 생략할 수 있고, 더불어서 expectation step에서 각 점이 클러스터에 속하는 확률값을 계산할 필요없이, 가장 가까운 중심의 클러스터에 hard assign되면 된다. 이를 수식으로 나타내자.

Loss function of K-means

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

다음과 같이 매우 간단하게 변할 수 있다. 여기서 r_{nk}값은 0또는 1의 값만 가지고, 이는 hard assign됨을 의미한다. Expectation step에서는 각각의 r_{nk}를 해당 데이터포인트에 가장 가까운 중심을 갖는 군집으로 할당하고, Maximization step에서는 새로 군집이 할당된 데이터포인트 사이에서 중심 μ_k를 업데이트한다.

$$\text{Expectation} \quad : \quad \text{Assign the data points to the nearest centroid}$$

$$\text{Maximization} \quad : \quad \mu_k = \frac{\sum_{n=1}^{N} r_{nk}x_n}{\sum_{n=1}^{N} r_{nk}}$$

Maximization step에서 새로운 μ_k는 $\frac{\text{assign된 데이터의 위치}}{assign\text{개수}}$가 되어서, 새롭게 assign된 데이터의 평균으로 μ_k가 이동함을 알 수 있다.

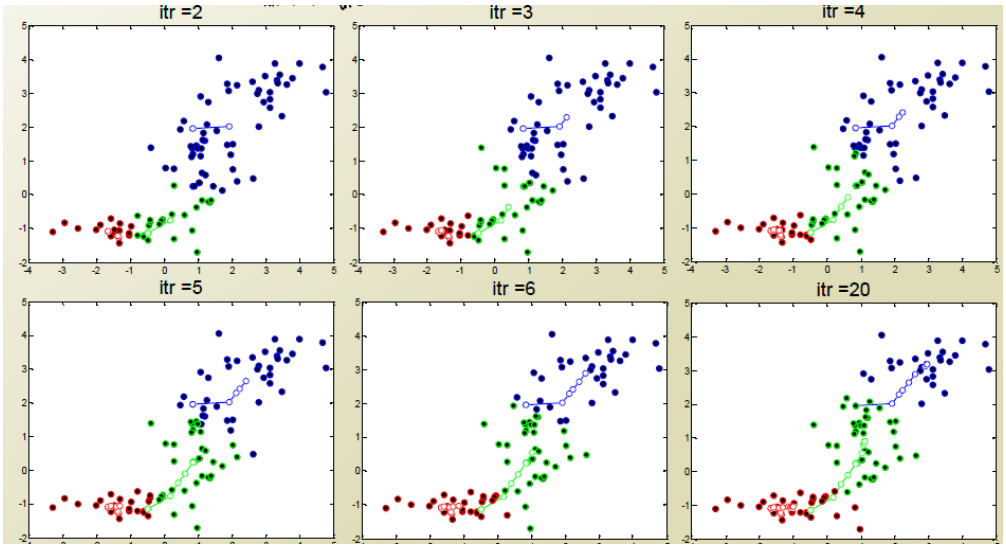


Figure 3: K-menas Clustering

이렇듯 우리가 흔히 사용하던 k-means 또한 EM으로 설명될 수 있음을 확인했다.

3.2 Latent Variable Modeling

우리가 익숙한 클러스터링에 대한 EM 예제를 살펴보았다. 하지만 이런 EM은 클러스터링 이외에도 잠재변수가 존재하는 경우 매우 유용하게 사용될 수 있다고 한다. 이를 위해서 다른 예시를 하나 더 가져왔다. 논문에 있는 예시는 ML추정으로 풀 수 있는 예시이기 때문에, 더 복잡한 예시를 가져와서 tex 치느라 힘들어 줄을뻔했다.

3.2.1 example - incomplete

남택이가 운영하는 베스킨라빈스는 민트초코, 딸기, 바닐라, 체리만 판다고 하자…즉 우리가 관측할 수 있는 변수 y는 다음과 같이 {민트초코, 딸기, 바닐라, 체리}로 이루어진다. 각각을 선택할 확률을 $(\frac{1}{2}+\frac{p}{4}, \frac{1}{4}(1-p), \frac{1}{4}(1-p), \frac{p}{4})$ 이라 하자. 그렇다면 하루동안의 매출을 이용해서 p값을 추정하고자 한다. 만약 하루동안 n명의 사람이 베라에 왔다면, 각 종류별로 팔린 아이스크림의 개수인 확률변수 Y는 다음과 같은 이항분포를 따른다.

$$Y \sim multi(n, \mathbf{p})$$
$$\mathbf{p} = (\frac{1}{2}+\frac{p}{4}, \frac{1}{4}(1-p), \frac{1}{4}(1-p), \frac{p}{4})$$

결국 확률변수 Y에 대한 모수로 θ는 최대가능도 추정을 통해 추정하는 것이 가능하다.

$$\begin{aligned} l(p) &= log\binom{n}{y_1\ y_2\ y_3\ y_4}(\frac{1}{2}+\frac{p}{4})^{y_1}(\frac{1-p}{4})^{y_2}(\frac{1-p}{4})^{y_3}(\frac{p}{4})^{y_4} \\ &= log\binom{n}{y_1\ y_2\ y_3\ y_4} + y_1 log(\frac{1}{2}+\frac{p}{4}) + y_2 log(\frac{1-p}{4}) + y_3 log(\frac{1-p}{4}) + y_4 log(\frac{p}{4}) \end{aligned}$$

이 경우는 상대적으로 미분값=0을 하면 간단하게 정리되지 않는다. p에 대한 이차식이 나오는데, 뭐 구하려면 근의공식 써서 analytical form을 찾을 수 있지만, 간단하지 않다. 따라서 우리는 EM으로 이 문제를 접근하고자 한다.

3.2.2 example - complete

우리는 이런 잠재변수까지 고려했을때 complete하다고 한다. 확률변수 $X = (X_1, X_2, X_3, X_4, X_5)$가 다섯가지 카테고리를 갖는 다항분포를 따른다고 하자.

$$X \sim multi(n, \mathbf{p})$$
$$\mathbf{p} = (\frac{1}{2},\ \frac{p}{4},\ \frac{1}{4}(1-p),\ \frac{1}{4}(1-p),\ \frac{p}{4})$$

그러면 확률변수 X와 Y는 다음과 같이 매핑될 수 있다.

$$Y = (Y_1, Y_2, Y_3, Y_4) = T(X) = (X_1 + X_2,\ X_3,\ X_4,\ X_5)$$

이를 일상의 언어로 표현하면, {민트초코, 딸기, 바닐라, 체리} 중에서 민트초코를 선택하는 잠재적인 요인에는 {민초단, 초코단} $\rightarrow$ {민트초코}라는 경로가 존재하는 것이다. 즉, 민트초코를 구매하는 고객중에, 이들이 민초에 대한 수요인지, 초코에 대한 수요인지까지 모델링하기 때문에 잠재변수까지 complete하게 표현하는 형태이다.

이 X에 대한 가능도함수는 다음과 같이 정의될 수 있다.

$$L(\theta) = \Pi_{i=1}^{n}\binom{n}{x_1\ x_2\ x_3\ x_4\ x_5}(\frac{1}{2})^{x_1}\ (\frac{1}{4}p)^{x_2}\ (\frac{1-p}{4})^{x_3}\ (\frac{1-p}{4})^{x_4}\ (\frac{1}{4}p)^{x_5}$$

로그가능도는 다음과 같다.

$$l(p) = h(x,n) + (x_2 + x_5)log(\frac{1}{4}p) + (x_3 + x_4)log(\frac{1-p}{4})$$

이를 p에 대해 미분해서 최대 가능도 추정량을 구하면 다음과 같다.

$$\hat{p} = \frac{x_2 + x_5}{x_2 + x_3 + x_4 + x_5}$$

3.2.3 Expectation step

이제 em 알고리즘을 위해 다음을 거치자.

$$Q(p, p^{(k)}) = E[X \mid Y = y, p^{(k)}]$$

이때 X_3, X_4, X_5는 Y_2, Y_3, Y_4과 같다. 그래서 다음이 성립한다.

$$E[X_i \mid Y = y, p^{(k)}] = y_{i-1}, \quad where \ \ i = 3, 4, 5$$

반면에 Y_1은 관측할 수 있지만, Y_1을 안다고 해서 X_1, X_2를 알 수는 없다. 하지만 Y_1를 알고 있다면, 최소한 X_1, X_2가 이항분포를 따르게 된다. 데이터 y와 $p^{(k)}$가 주어졌을 때, X_1에 대한 조건부 pdf는 다음과 같다.

$$\begin{aligned} p\left(x_1 | \underline{y}, p^{(k)}\right) &= \frac{p\left(x_1, y | p^{(k)}\right)}{p\left(y | p^{(k)}\right)} \\ &= \frac{\frac{n!}{x_1!(y_1-x_1)!y_2!y_3!y_4!}\left(\frac{1}{2}\right)^{x_1}\left(\frac{1}{4}p^{(k)}\right)^{y_1-x_1}}{\frac{n!}{y_1!y_2!y_3!y_4!}\left(\frac{1}{2}+\frac{1}{4}p^{(k)}\right)^{y_1}} \\ &= \frac{y_1!}{x_1!\left(y_1-x_1\right)!}\left(\frac{\frac{1}{2}}{\frac{1}{2}+\frac{1}{4}p^{(k)}}\right)^{x_1}\left(\frac{\frac{1}{4}p^{(k)}}{\frac{1}{2}+\frac{1}{4}p^{(k)}}\right)^{y_1-x_1}. \end{aligned}$$

$X_2 = y_1 - X_1$이기 때문에, 다음과 같이 이항분포를 만들 수 있다.

$$\begin{aligned} X_1 | \underline{Y} = \underline{y}, p^{(k)} &\sim B\left(y_1, \frac{\frac{1}{2}}{\frac{1}{2}+\frac{1}{4}p^{(k)}}\right) \\ X_2 | \underline{Y} = \underline{y}, p^{(k)} &\sim B\left(y_1, \frac{\frac{1}{4}p^{(k)}}{\frac{1}{2}+\frac{1}{4}p^{(k)}}\right) \end{aligned}$$

그래서 각각의 기대값을 구하게 되면, 다음과 같다.

$$\begin{aligned} \mathbb{E}\left[X_1 | \underline{Y} = \underline{y}, p^{(k)}\right] &= y_1 \frac{\frac{1}{2}}{\frac{1}{2}+\frac{1}{4}p^{(k)}}, \\ \mathbb{E}\left[X_2 | \underline{Y} = \underline{y}, p^{(k)}\right] &= y_1 \frac{\frac{1}{4}p^{(k)}}{\frac{1}{2}+\frac{1}{4}p^{(k)}} \end{aligned}$$

이 값을 다시 대입하면,

$$
\begin{aligned}
Q\left(p, p^{(k)}\right) &= \mathbb{E}\left[l\left(\theta; \underline{X}\right) | \underline{Y} = \underline{y}, p^{(k)}\right] \\
&= \left(y_1 \frac{\frac{1}{4}p^{(k)}}{\frac{1}{2} + \frac{1}{4}p^{(k)}} + y_4\right) log\left(\frac{1}{4}p\right) + \left(y_2 + y_3\right) log\left(\frac{1-p}{4}\right) + h\left(\underline{y}, p^{(k)}\right)
\end{aligned}
$$

결국에 아까 X의 로그 가능도 함수에서, x들의 자리가 y로 대체된 것을 확인할 수 있다.

3.2.4 Maximization step

이에 대한 최대가능도 추정은 아까 식에 대입만 하면 된다.

$$
p^{(k+1)} = \frac{\left(y_1 \frac{\frac{1}{4}p^{(k)}}{\frac{1}{2} + \frac{1}{4}p^{(k)}} + y_4\right)}{y_1 \frac{\frac{1}{4}p^{(k)}}{\frac{1}{2} + \frac{1}{4}p^{(k)}} + y_2 + y_3 + y_4}
$$