

EM 알고리즘

정의

- 개요

1. 겉으로 드러나지 않은 잠재변수까지 고려하여 최대우도추정을 실시하는 알고리즘

- 유도

1. 관측된 개체의 벡터와, 잠재변수를 담고 있는 벡터를 각각

1) $X = \begin{bmatrix} X_1 \\ \dots \\ X_n \end{bmatrix}, Z = \begin{bmatrix} Z_1 \\ \dots \\ Z_n \end{bmatrix}$ 으로 표기하자.

2. X_i 는 $\theta \in \Omega$ 일 때 pdf $f(x; \theta)$ 를 가지며 X_i 와 Z_i 는 서로 독립을 가정한다.

(1) 이 때,

- $g(x; \theta) = \prod_{i=1}^n f(x_i; \theta)$ 인 결합 pdf,
- $h(x, z; \theta) = \prod_{i=1}^n f(x_i; \theta)f(z_i; \theta)$ 인 X 와 Z 의 결합 pdf

(2) $k(z|\theta, X) = \frac{h(x, z; \theta)}{g(x; \theta)}$ 이다.

정의

- 유도

3. 이들을 우도 함수로 변환하고, 각각을 다음과 같이 정의하면

1) $L(\theta; x) = \prod_{i=1}^n f(X_i; \theta)$, $L^c(\theta; x, z) = \prod_{i=1}^n h(X_i, Z_i; \theta)$

2) 이제 목표는 관측된 우도함수 $L(\theta; x)$ 를 최대화 하는 것이다.

3) 이 때

$$\begin{aligned} (1) \log L(\theta; x) &= \int \log L(\theta; x) k(z|\theta_0, X) dz \\ &= \int \log \left(\frac{h(x, z; \theta)}{k(z|\theta, X)} \right) k(z|\theta_0, X) dz \\ &= \int \log[h(x, z; \theta)] k(z|\theta_0, X) dz - \int \log[k(z|\theta, X)] dz \log[g(x; \theta)] \\ &= E_{\theta_0}[\log L^c(\theta; x, z)] - E_{\theta_0}[\log k(z|\theta, X)] \end{aligned}$$

4. $Q(\theta|\theta_0, X) = E_{\theta_0}[\log L^c(\theta; x, z)]$ 로 정의하고, 함수 $Q(\theta|\theta_0, X)$ 를 최대화하는 것이 목적이 된다.

정의

- Expectation단계와 Maximization단계

1. 앞서 구한 $Q(\theta|\hat{\theta}, X) = E_{\hat{\theta}}[\log L^c(\theta; x, z)|\hat{\theta}, X]$ 를 정의하는 것을 E단계라고 한다.

1) 즉. 잠재변수와 관측된 변수의 결합우도함수가 최대화 될 때

(1) $k(z|\hat{\theta}, X) = \frac{h(x, z; \hat{\theta})}{g(x; \hat{\theta})}$ 에서 잠재변수 K 의 우도함수 $k(z|\hat{\theta}, X)$ 가 최대화 되므로

(2) 우선 이 기댓값을 정의하는 것을 목적으로 한다.

2) 이 때, 이 기댓값을 받치는 함수는 $k(z|\theta, X)$ 가 된다.

$$(1) Q(\theta|\hat{\theta}, X) = \int \log[h(x, z; \hat{\theta})] k(z|\hat{\theta}, X) dz$$

정의

- Expectation단계와 Maximization단계

2. $Q(\theta|\theta_0, X)$ 를 최대화하는 단계를 Maximization 이라고 한다.

1) $\hat{\theta}^{(m+1)} = \operatorname{argmax}[Q(\theta|\hat{\theta}, X)]$

2) 이 때, 이 우도함수를 최대화하는 방법으로 최대우도추정(MLE)가 고려될 수 있다.

3. Expectation을 통해 로그우도의 기댓값 함수를 도출하고,

Maximization을 통해 각 추정량의 최대우도 추정량을 갱신해서 구하는 것을 EM알고리즘이라 한다.

예제

• EM알고리즘을 구하는 예

1. X_1, \dots, X_n 을 pdf $f(\mathbf{x} - \boldsymbol{\theta})$ 를 공통으로 갖는 확률변수라고 하자.
2. 또한 Z_1, \dots, Z_n 을 중도절단 관측값이라고 하고, 단지 X_i 와 Z_i 가 서로 독립이라는 것만 알려졌고, $Z_i > a$ 라는 것만 알려졌다고 가정하자.

1) 관측된 우도함수와 완전한 우도함수를 구하면

$$(1) L(\theta; x) = [1 - F(a - \theta)]^{n_2} \prod_{i=1}^{n_1} f(X_i - \theta)$$

$$(2) L^c(\theta; x, z) = \prod_{i=1}^{n_1} f(X_i - \theta) \prod_{i=1}^{n_1} f(z_i - \theta)$$

2) 조건부 분포함수를 구하면

$$(1) \frac{\prod_{i=1}^{n_1} f(X_i - \hat{\theta}) \prod_{i=1}^{n_1} f(z_i - \hat{\theta})}{[1 - F(a - \hat{\theta})]^{n_2} \prod_{i=1}^{n_1} f(X_i - \hat{\theta})} = [1 - F(a - \hat{\theta})]^{-n_2} \prod_{i=1}^{n_1} f(z_i - \hat{\theta})$$

예제

- EM알고리즘을 구하는 예

3) 이 함수를 받침함수로 하는 완전한 우도함수의 기댓값을 구하면

$$\begin{aligned}(1) \quad Q(\theta|\theta_0, X) &= E_{\theta_0}(\log \prod_{i=1}^{n_1} f(X_i - \theta) \prod_{i=1}^{n_1} f(z_i - \theta)) \\ &= E_{\theta_0}(\sum_{i=1}^{n_1} \log f(x_i - \theta) + \sum_{i=1}^{n_1} \log f(z_i - \theta))\end{aligned}$$

$$\begin{aligned}(2) \quad &\text{이 때, 기댓값 } E_{\theta_0} \text{는 } z_i \text{에 대한 기댓값이므로} \\ &= \sum_{i=1}^{n_1} \log f(x_i - \theta) + E_{\theta_0}[\sum_{i=1}^{n_1} \log f(z_i - \theta)] \\ &= \sum_{i=1}^{n_1} \log f(x_i - \theta) + \int \sum_{i=1}^{n_1} \log f(z_i - \theta) [1 - F(a - \theta)]^{-n_2} \prod_{i=1}^{n_1} f(z_i - \theta) dz \\ &= \sum_{i=1}^{n_1} \log f(x_i - \theta) + \int \log f(Z - \theta) [1 - F(a - \theta)]^{-n_2} f(Z - \theta) dz\end{aligned}$$

예제

- EM알고리즘을 구하는 예

4) $Q(\theta|\theta_0, X)$ 를 최대화하기 위해 MLE 추정량을 구하면

$$\begin{aligned}(1) \frac{\partial Q(\theta|\theta_0, X)}{\partial \theta} &= \frac{1}{\partial \theta} \left\{ \sum_{i=1}^{n_1} \log f(x_i - \theta) + \int \log f(Z - \theta) [1 - F(a - \hat{\theta})]^{-n_2} f(Z - \hat{\theta}) dz \right\} \\ &= \sum_{i=1}^{n_1} \frac{f'(x_i - \theta)}{f(x_i - \theta)} + \int \frac{f'(Z - \theta)}{f(Z - \theta)} [1 - F(a - \hat{\theta})]^{-n_2} f(Z - \hat{\theta}) dz\end{aligned}$$

(2) $\frac{\partial Q(\theta|\theta_0, X)}{\partial \theta} = 0$ 으로 놓고, 각 모수에 대한 MLE 추정량을 구한다.

예제

- 정규분포

1. 위 논의에서 X 를 $N(\theta, 1)$ 인 정규분포를 따른다고 가정하면

$$(1) \quad \frac{\partial Q(\theta|\theta_0, X)}{\partial \theta} = - \left\{ \sum_{i=1}^{n_1} \frac{f'(x_i - \theta)}{f(x_i - \theta)} + n_2 \int \frac{f'(Z - \theta)}{f(Z - \theta)} [1 - F(a - \hat{\theta})]^{-n_2} f(Z - \hat{\theta}) dz \right\}$$

$$(2) \quad \text{위에서 } f(x - \theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right) \text{ 이고, } f'(x - \theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right) (x - \theta)(-1)$$

$$\text{- 따라서 } \frac{f'(x-\theta)}{f(x-\theta)} = \frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right) (x-\theta)(-1)}{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right)} = -(x - \theta)$$

(3) 다시 정리하면

$$\begin{aligned} \text{- } & - \left\{ \sum_{i=1}^{n_1} -(x_i - \theta) - n_2 \int (Z - \theta) [1 - F(a - \hat{\theta})]^{-1} f(Z - \hat{\theta}) dz \right\} \\ & = n_1(\bar{x} - \theta) + n_2 \int (Z - \hat{\theta}) \frac{1}{\sqrt{2\pi}} \frac{\exp\left(-\frac{(Z-\hat{\theta})^2}{2}\right)}{1 - F(a - \hat{\theta})} dz - n_2(\theta - \hat{\theta}) \\ & = n_1(\bar{x} - \theta) + \frac{n_2}{1 - \Phi(a - \hat{\theta})} \phi(a - \hat{\theta}) - n_2(\theta - \hat{\theta}) \end{aligned}$$

$$2. \quad n_1(\bar{x} - \theta) + \frac{n_2}{1 - \Phi(a - \hat{\theta})} \phi(a - \hat{\theta}) - n_2(\theta - \hat{\theta}) = 0 \text{ 으로 놓고 } \theta \text{에 대해 풀면}$$

$$\text{- } (n_1 + n_2)\theta = n_1\bar{x} + \frac{n_2}{1 - \Phi(a - \hat{\theta})} \phi(a - \hat{\theta}) + n_2\hat{\theta}$$

$$\text{- } N = (n_1 + n_2) \text{ 로 놓으면 } \theta = \frac{n_1}{n}\bar{x} + \frac{n_2}{n} \frac{\phi(a - \hat{\theta})}{1 - \Phi(a - \hat{\theta})} + \frac{n_2}{n}\hat{\theta}$$