4. 데이터 기반 추론 1

임채영
서울대학교 통계학과

이번 강의에서 다룰 내용

- 데이터 기반 (통계적) 추론 - 통계적 추정의 개념 (점추정, 구간추정, Bootstrap)
- 통계적 추정방법 (적률추정, 최대가능도법)
- EM 알고리즘

통계적 추론 - 통계적 추정의 개념

통계적 추론

- 통계적 추론(Statistical Inference)은 표본으로부터의 정보를 이용하여 모집단에 관한 추측이나 결론을 이끌어내는 과정을 전반을 말한다.
  - 통계적 추정(Statistical Estimation)
  - 유의성 검정(Significance test, 또는 가설 검정(Hypothesis test))
- 관측 또는 실험에 의해 획득한 데이터를 모집단(유한 또는 무한)의 표본으로 본다.

통계적 추정(Statistical Estimation)

- 표본으로부터 모집단의 관심 정보에 대한 추측값과 오차를 제시
- 모집단의 관심 정보는 평균, 분산 과 같은 특정한 값(모수)일 수도 있고, 분포함수와 같은 복잡한 형태를 가질 수도 있다.
- 여기서는 모수에 대한 추정(Parameter Estimation)을 얘기하고 크게 두 가지로 나누어 생각한다.

- 점추정 (point estimation), 구간추정 (interval estimation)

점추정

점추정(Point Estimation)이란 표본으로부터 계산한 모수의 추정값(추정량, Estimator)을 제시하는 추정방식이다.

점추정의 예

- 모평균의 추정량 : 표본평균 $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- 모분산의 추정량 : 표본분산 $\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

추정량의 평가

추정량을 평가하는 기준

(1) 편향 (Bias, 편의): Bias $(\hat{\theta})$ = $E(\hat{\theta}) - \theta$

- 불편추정량 (Unbiased estimator)

  $E(\hat{\theta}) = \theta$를 만족하는 추정량 $\hat{\theta}$

  예)표본평균과 표본분산은 각각 모평균과 모분산의 불편추정량이다.

(2) 표준오차: 추정량의 표준편차 $SE(\hat{\theta})$

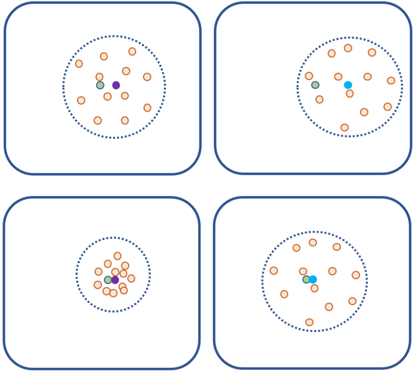
예) $X_1, \cdots, X_n \overset{i.i.d.}{\sim} (\mu, \sigma^2)$,

$$SE(\hat{\mu}) = SD(\hat{\mu}) = \sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$$

평균제곱오차(Mean Squared Error, MSE)

(3) Bias와 SE를 동시에 고려한 평가 기준

$$
\begin{aligned}
MSE(\hat{\theta}) &= Var(\hat{\theta}) + (Bias(\hat{\theta}))^2 \\
&= (SE(\hat{\theta}))^2 + (Bias(\hat{\theta}))^2
\end{aligned}
$$



구간추정

- 구간추정(Interval Estimation)이란 모수의 추정값을 구간으로 제공하는 추정방식이다.
- 구간추정의 일반적 방법: 신뢰구간 (Confidence Interval, CI)
- 신뢰수준 (Confidence level)이 $100(1-\alpha)\%$인 신뢰구간 (L, U)는 다음을 만족한다.

$$P(L \le \theta \le U) = 1 - \alpha$$

- L, U는 표본으로부터 구해짐. 즉, $L \equiv L(X_1, \cdots, X_n)$, $U \equiv U(X_1, \cdots, X_n)$
- 따라서 (L, U) 는 확률 변수로 이루어진 구간 (random interval)

- $1 - \alpha$ 는 포함확률(coverage probability)이라고 부름

신뢰구간 예시

모분산 σ^2를 알 때 정규모집단의 모평균 μ의 구간추정

- 신뢰수준 $100(1-\alpha)$%인 μ의 신뢰구간은 다음과 같다.

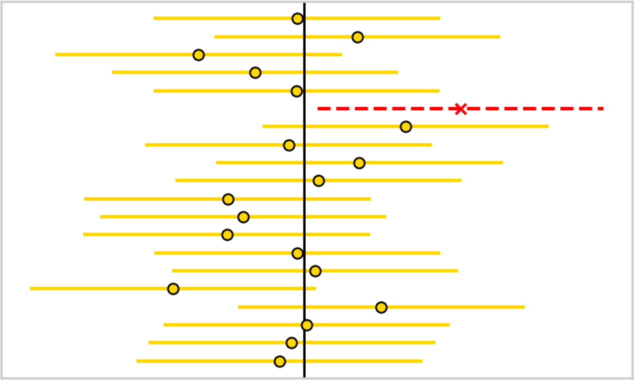
$$\left(\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\ \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

- $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$로부터

$$\begin{aligned} 1-\alpha &= P\left(-Z_{1-\frac{\alpha}{2}} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le Z_{1-\frac{\alpha}{2}}\right) \\ &= P\left(\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

신뢰구간의 의미

- μ의 100(1-α)% 신뢰구간 : 100번의 표본 추출을 통해 얻어진 100개의 신뢰구간 $\left(\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$ 에서, 100(1-α)% 개 정도의 신뢰구간이 모평균을 포함할 것으로 기대함



오차의 한계와 표본의 크기

- 오차의 한계(Margin of Error)는 최대 허용오차로 이를 통해 오차의 한계가 일정수준 이하가 되게 하기 위해 필요한 표본의 수를 계산할 수 있다.
- 모분산을 알고 있는 경우의 정규모집단의 모평균 신뢰구간으로부터 구한 오차의 한계는 $Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$이다.
- 이때, 오차의 한계를 d 이하로 하는 n는 다음의 식을 통해 찾는다.

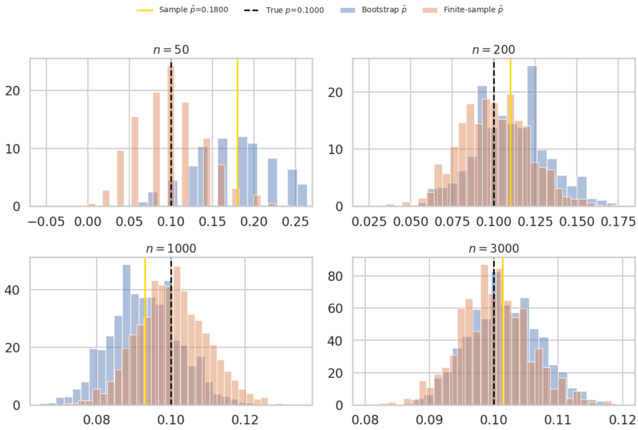
$$
\begin{aligned}
Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} &\leq d, \\
\left(\frac{Z_{1-\frac{\alpha}{2}} \sigma}{d} \right)^2 &\leq n
\end{aligned}
$$

붓스트랩

- 붓스트랩(Bootstrap)은 1979년 Efron에 의해 제안된 방법으로 하나의 표본으로부터 추정량의 불확실성을 추정하는 방법이다.
- 외부의 도움(모집단의 정보나 이론적 분포 가정)없이 스스로 문제를 해결(가지고 있는 표본으로)한다는 의미로도 해석할 수 있음 (Pull oneself up by one's bootstraps)
- 한 개의 표본을 모집단으로 생각하고 표본에서 많은 수의 표본을 재추출(Resampling)하여 여러 개의 추정값을 구해서 추정량의 경험적 분포를 근사적으로 구하는 방법
- 재추출때는 복원추출을 사용함

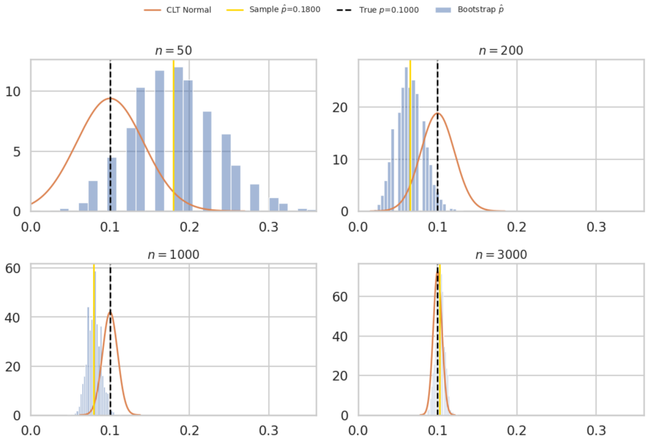
붓스트랩표본들로 구한 분포와 경험적 표본분포 비교

배터리 불량품 예시($p = 0.1$)에서 크기가 n개인 표본 1개에 대한 5000개의 붓스트랩 표본들로 구한 히스토그램과 크기가 n개인 표본을 여러 개(1000) 추출해서 $\hat{p}$의 히스토그램을 그려 비교



붓스트랩표본들로 구한 분포와 점근분포 비교

배터리 불량품 예시($p = 0.1$)에서 크기가 n개인 표본 1개에 대한 5000개의 붓스트랩 표본들로 구한 히스토그램과 $\hat{p}$의 중심극한정리에 의한 점근분포와 비교



붓스트랩표본으로 구한 신뢰구간

- 제품 불량률이 $p = 0.1$인 공장에서 제품 200개를 임의추출해서 계산한 표본비율은 0.13이라고 하자.
- CLT를 이용한 불량률 p의 95% 신뢰구간은 다음과 같다.
  $$\left(\hat{p} - Z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = (0.0834, 0.1766)$$
- 5000개의 붓스트랩 표본을 이용하여 2.5%, 97.5% 분위수로 구한 신뢰구간은 (0.0850, 0.1800)로 나온다. (표본 추출의 임의성으로 계산할때마다 약간씩 다른 값이 나올 수 있다)

통계적 추정 방법과 알고리즘

적률추정법

- 적률 (moment) :

- 확률변수 X의 거듭제곱의 기댓값

  k차 적률, $k - th$ moment, $m_k = E(X^k)$

- 표본적률 (sample moment):

- 랜덤 표본 $\{X_1, \cdots, X_n\}$으로부터

$$\hat{m}_k = \frac{1}{n}(X_1^k + \cdots + X_n^k)$$

- 1차 표본적률 = 표본평균

- 큰수의 법칙에 의해 $\hat{m}_k \to m_k$

- 적률추정법(Method of Moment Estimation)이란 적률에 대한 함수로 표현되는 모수에 대해, 해당 추정량을 표본적률에 대한 함수로 표현하는 방법이다.
- 적률추정법을 통한 모수의 추정량을 적률 추정량(Method of Moment Estimator, MME)이라 한다.
- 적률에 대한 함수로 나타나는 모수 $\eta = g(m_1, m_2, \cdots, m_k)$의 적률추정량 : $\hat{\eta} = g(\hat{m}_1, \hat{m}_2, \cdots, \hat{m}_k)$

적률추정량의 예

모분산 추정

- $X_1, \cdots, X_n \overset{i.i.d.}{\sim} (\mu, \sigma^2)$,

$$\sigma^2 = E(X_1^2) - (E(X_1))^2 \ = \ m_2 - m_1^2$$

$$\hat{\sigma^2}^{MME} = \hat{m}_2 - (\hat{m}_1)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X})^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$$

$$\hat{\sigma}^{MME} = \sqrt{\hat{m}_2 - (\hat{m}_1)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}$$

적률추정량의 비유일성

하나의 모수에 대한 적률추정량은 유일하지 않을수 있다.

- $X_1, \cdots, X_n \overset{i.i.d.}{\sim} Poisson(\lambda)$
- $E(X_i) = var(X_i) = \lambda$

  $\lambda = E(X_1) = m_1$ or $\lambda = Var(X_1) = m_2 - m_1^2$

  $\hat{\lambda}^{MME} = \hat{m}_1 = \bar{X}$ or $\hat{\lambda}^{MME} = \hat{m}_2 - (\hat{m}_1)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

최대 가능도 추정법

- 최대 가능도 추정법(Maximum Likelihood Estimation)이란 가능도(likelihood)를 최대가 되게 하는 모수 값을 추정량으로 구하는 방법으로 먼저 가능도를 설명한다.
- 가능도

- 확률변수의 관측값이 주어졌을때 해당 확률의 정도
- 모수를 가지는 확률 분포의 경우, 모수값이 가능한 정도

- 일반적으로, 모수 θ에 대한 가능도함수는

$$L(\theta) = L(\theta|X_1, \cdots, X_n) = f(X_1, \cdots, X_n; \theta)$$

즉, 확률밀도함수에 확률변수의 관측값을 대입한 값

- 이산확률변수인 경우 가능도는 관측값이 주어졌을때의 해당 확률이 된다.
- 확률변수들이 독립적으로 관측되었을 경우 가능도는 다음과 같이 주변확률분포의 곱으로 표현

$$L(\theta) = L(\theta|X_1, \cdots, X_n) = \prod_{i=1}^{n} f(X_i; \theta)$$

최대가능도 추정량

- 주어진 관측값에서 모수의 가능도를 최대가 되게 하는 값을 최대가능도 추정량 (Maximum Likelihood Estimator, MLE)이다.

$$\hat{\theta} \quad = \quad \arg\max_{\theta\in\Omega} L(\theta|X_1, \cdots, X_n)$$

- 독립인 관측값들이 주어졌을 때, 로그변환을 시킬 경우 좀 더 다루기 편한 함수가 되기도 한다.

$\ell(\theta) = \log(L(\theta)) = \log(\prod f(X_i;\theta)) = \sum_{i=1}^{n} \log(f(X_i,\theta))$.

$$\hat{\theta} = \arg\max_{\theta\in\Omega} L(\theta) = \arg\max_{\theta\in\Omega} \ell(\theta)$$

가능도 계산의 예

- 공장 A의 한 생산라인의 불량률 (p)을 추정하기 위해 10개의 제품을 랜덤 추출하였다고 하자. 불량인 경우를 1로 할 경우, 10개의 제품의 불량정보는 $(0, 0, 0, 1, 0, 1, 0, 0, 0, 0)$이라고 하자.
- 각 제품의 불량 유무를 1또는 0을 가지는 확률변수 X로 봤을때 $X_1, \cdots, X_{10} \overset{i.i.d.}{\sim} Ber(p)$이고 이때의 가능도 함수는?

- 불량률 p에 대한 가능도: $L(p) = p^2(1-p)^8, 0 \le p \le 1$

  즉, 불량률이 p일때 $(0,0,0,1,0,1,0,0,0,0)$를 관측할 확률(정도)
- $L(0.5) = 0.5^2 0.5^8 = 0.00098$: 불량률이 $p = 0.5$일때 $(0,0,0,1,0,1,0,0,0,0)$를 관측할 확률(정도)
- $L(0.2) = 0.2^2 0.8^8 = 0.0067$: 불량률이 $p = 0.2$일때 $(0,0,0,1,0,1,0,0,0,0)$를 관측할 확률(정도)

MLE 의 성질

몇가지 가정하에,

- 일치성(Consistency):

$$\hat{\theta} \to \theta \text{ in probability}$$

- 점근적 정규성(Asymptotic Normality):

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, \mathcal{I}^{-1}(\theta)) \text{ in distribution},$$

  여기서 $\mathcal{I}(\theta) = E\left[\left(\frac{d}{d\theta}\log f(X;\theta)\right)^2\right]$는 피셔정보 (*Fisher Information*)라고 부른다.

- 즉, 점근적으로 (n이 커질때), MLE는 평균이 θ, 분산이 $\frac{1}{n\mathcal{I}(\theta)}$인 정규분포를 따른다.
- 점근분포를 이용하여 신뢰구간을 구할 수 있다.

최대가능도 추정량의 계산

- 보통 가능도가 최대가 되게 하는 θ를 찾기 위해
- $\dot{\ell}(\theta) = \partial\ell(\theta)/\partial\theta = 0$ (score function) 의 해 중에 최대가능도 추정량을 찾음.
- MLE가 closed-form으로 구해지는 경우도 있지만 모형이 복잡해지면 대부분 그렇지 않다.
- closed-form이 없는 경우, 수치적으로 최댓값을 찾는다- Newton-Rapson method, Fisher scoring method, EM/MM algorithm

Newton-Raphson 방법

- 수치적으로 $\dot{\ell}(\theta) = 0$의 해(root)를 찾는 방법 중, Newton-Raphson 방법이 있다.
- N-R 알고리즘은 테일러 전개를 통해 유도할수 있다.
- θ^*를 $f(\theta)$의 해라고 하자. 즉, $f(\theta^*) = 0$.

$$0 = f(\theta^*) \approx f(\theta^{(r)}) + f'(\theta^{(r)})(\theta^* - \theta^{(r)})$$

- $\theta^* \approx \theta^{(r)} - \frac{f(\theta^{(r)})}{f'(\theta^{(r)})}$이 되므로

$$\theta_{new} = \theta_{old} - \frac{f(\theta_{old})}{f'(\theta_{old})}$$

- $f(\theta) = \dot{\ell}(\theta)$로 생각하면 다음의 반복알고리즘(iterative algorithm)이 유도된다

$$\hat{\theta}^{(r+1)} = \hat{\theta}^{(r)} + [-\ddot{\ell}(\hat{\theta}^{(r)})]^{-1}\dot{\ell}(\hat{\theta}^{(r)}),\ r = 0, 1, 2. \cdots$$

EM 알고리즘

- EM(Expectation-Maximization)알고리즘은 잠재변수(Latent variable)나 결측치 (Missing variable)가 있는 경우의 가능도 함수를 최대화 할때 (MLE를 구할때) 사용하는 알고리즘이다.
- 다양한 통계분석방법 - 혼합분포 모형(mixture model), 군집분석 (cluster analysis), 임의효과 모형 (random effects model)- 에서 사용된다.

EM 알고리즘을 통한 MLE 찾기

- 먼저 다음과 같은 기호를 사용한다.
- $\boldsymbol{Y}_o$:관측한 데이터
- $\boldsymbol{Y}_m$:결측치 또는 잠재변수
- $f(\boldsymbol{Y}_o, \boldsymbol{Y}_m|\boldsymbol{\theta})$: $(\boldsymbol{Y}_o, \boldsymbol{Y}_m)$의 결합확률밀도함수
- $f(\boldsymbol{Y}_o|\boldsymbol{\theta})$: $\boldsymbol{Y}_o$의 확률밀도함수.
- $f(\boldsymbol{Y}_m|\boldsymbol{Y}_o, \boldsymbol{\theta}) = \frac{f(\boldsymbol{Y}_o, \boldsymbol{Y}_m|\boldsymbol{\theta})}{f(\boldsymbol{Y}_o|\boldsymbol{\theta})}$: $\boldsymbol{Y}_m$의 조건부 확률밀도함수.

- MLE는 관측한 데이터로 구한 가능도 함수를 최대화 하는 모수를 찾는것이다. 즉,

$$\arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \log f(\boldsymbol{Y}_o|\boldsymbol{\theta})$$

- $\ell(\boldsymbol{\theta}) = \log f(\boldsymbol{Y}_o|\boldsymbol{\theta})$의 계산이 복잡한 경우

  $\ell(\boldsymbol{\theta}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{(r)})$이고, $\ell(\boldsymbol{\theta}_{(r)}) = Q(\boldsymbol{\theta}_{(r)}|\boldsymbol{\theta}_{(r)})$이면서 최대화를 하기 쉬운 Q를 찾아 Q(minorize함수)를 최대화하는 $\boldsymbol{\theta}$로 업데이트 하는 방법을 생각해볼수 있다.

E-Step and M-step

$\ell(\boldsymbol{\theta}) = \log f(\boldsymbol{Y}_o|\boldsymbol{\theta})$를 최대화하는 EM알고리즘은 반복알고리즘으로 (r)번째 스텝에서 $(r+1)$번째 스텝으로의 업데이트는 다음의 두 과정을 통해 진행된다.

(1) **E**xpectation step: $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{(r)})$ 찾기

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{(r)}) &= \int \log\left(f(\boldsymbol{Y}_o, \boldsymbol{Y}_m|\boldsymbol{\theta})\right) f(\boldsymbol{Y}_m|\boldsymbol{Y}_o, \boldsymbol{\theta}_{(r)}) d\boldsymbol{Y}_m \\ &= E\left(\log\left(f(\boldsymbol{Y}_o, \boldsymbol{Y}_m|\boldsymbol{\theta})\right)\right) \end{aligned}$$

- $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{(r)})$이 기댓값 ($E()$)로 표현되기 때문에 Expectation step이라고 부름

(2) **M**aximization step: $\boldsymbol{\theta}_{(r+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{(r)})$

<참고> 문제에 따라 위 두 스텝을 정리하여 한번에 표시할수 있다. (혼합분포를 이용한 군집분석의 예)

E- step 유도 과정 [참]

$\ell(\boldsymbol{\theta}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{(r)})$이고, $\ell(\boldsymbol{\theta}_{(r)}) = Q(\boldsymbol{\theta}_{(r)}|\boldsymbol{\theta}_{(r)})$인 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{(r)})$ 찾기

$$(1)\ \ell(\boldsymbol{\theta}) = \log f(\boldsymbol{Y}_o|\boldsymbol{\theta}) = \log \int f(\boldsymbol{Y}_o, \boldsymbol{Y}_m|\boldsymbol{\theta}) d\boldsymbol{Y}_m$$

$$(2)\ \ell(\boldsymbol{\theta}_{(r)}) = \log f(\boldsymbol{Y}_o|\boldsymbol{\theta}_{(r)}) = \int \log f(\boldsymbol{Y}_o|\boldsymbol{\theta}_{(r)}) f(\boldsymbol{Y}_m|\boldsymbol{Y}_o, \boldsymbol{\theta}_{(r)}) d\boldsymbol{Y}_m$$

(1) 과 (2)를 이용하여 다음의 관계를 찾을수 있다.

$$\ell(\boldsymbol{\theta}) \geq \ell(\boldsymbol{\theta}_{(r)}) + \int \log\left(\frac{f(\boldsymbol{Y}_o, \boldsymbol{Y}_m|\boldsymbol{\theta})}{f(\boldsymbol{Y}_o, \boldsymbol{Y}_m|\boldsymbol{\theta}_{(r)})}\right) f(\boldsymbol{Y}_m|\boldsymbol{Y}_o, \boldsymbol{\theta}_{(r)}) d\boldsymbol{Y}_m =: Q_0\left(\boldsymbol{\theta}|\boldsymbol{\theta}_{(r)}\right)$$

$$\ell(\boldsymbol{\theta}_{(r)}) = Q_0\left(\boldsymbol{\theta}_{(r)}|\boldsymbol{\theta}_{(r)}\right)$$

$Q_0(\boldsymbol{\theta}|\boldsymbol{\theta}_{(r)})$에서 $\boldsymbol{\theta}$에 관련된 부분은

$$\int \log\left(f(\boldsymbol{Y}_o, \boldsymbol{Y}_m|\boldsymbol{\theta})\right) f(\boldsymbol{Y}_m|\boldsymbol{Y}_o, \boldsymbol{\theta}_{(r)}) d\boldsymbol{Y}_m =: Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{(r)})$$

EM 알고리즘의 예- 혼합분포를 이용한 군집분석

- EM알고리즘을 적용하는 예로 4장에서 다루었던 가우시안 혼합분포를 이용한 군집분석를 다시 소개한다.
- 두개의 군집으로 이루어져있는 데이터가 두 개의 구성원을 갖는 가우시안 혼합분포를 따른다고 가정하자.

즉, $X_1, \cdots, X_n \overset{i.i.d.}{\sim} f(x|\boldsymbol{\theta})$ with

$$f(x|\boldsymbol{\theta}) = w_0 \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(X_i-\mu_0)^2} + (1-w_0)\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(X_i-\mu_1)^2}$$

$\boldsymbol{\theta} = (\mu_0, \sigma_0^2, \mu_1, \sigma_1^2, w_0)$.

- 로그 가능도함수는 다음과 같다.

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log f(X_i|\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \log\left(w_0 \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(X_i-\mu_0)^2} + (1-w_0)\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(X_i-\mu_1)^2} \right)\end{aligned}$$

- 로그함수 내부에 복잡한 함수형태가 또 들어있어서 가능도함수를 최대가 되게 하는 $\boldsymbol{\theta}$ 를 찾기가 쉽지 않다.

$i-$번째 데이터 X_i가 어느 구성원으로부터 왔는지를 잠재 변수 K_i를 도입하여 표현할 수 있다.

- $K_i = 0$이면 $X_i \sim N(\mu_0, \sigma_0^2)$, $K_i = 1$이면 $X_i \sim N(\mu_1, \sigma_1^2)$

  즉, K_i는 $i-$번째 데이터가 어느 군집에 속하는지 알려주는 label로 생각할 수 있다.
- 잠재변수 K_i는 $P(K_i = 0) = w_0$, $P(K_i = 1) = 1 - w_0$인 확률변수이다.
- EM알고리즘을 사용하여 $f(X_i|\theta)$보다 간단한 형태인 $f(X_i, K_i|\theta)$를 이용한 최대화 문제로 바꾸게 된다.

- $\boldsymbol{Y}_o = \boldsymbol{X} = (X_1, \cdots, X_n)$, $\boldsymbol{Y}_m = \boldsymbol{K} = (K_1, \cdots, K_n)$
- $f(\boldsymbol{Y}_o, \boldsymbol{Y}_m|\boldsymbol{\theta}) = f(\boldsymbol{X}, \boldsymbol{K}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(X_i, K_i|\boldsymbol{\theta})$

$$
\begin{aligned}
\log f(\boldsymbol{Y}_o, \boldsymbol{Y}_m|\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log f(X_i, K_i|\boldsymbol{\theta}) \\
&= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log(\sigma_{K_i}^2) - \sum_{i=1}^{n}\left(\frac{1}{2\sigma_{K_i}^2}(X_i - \mu_{K_i})^2\right) \\
&+ \left(n - \sum_{i=1}^{n} K_i\right)\log(w_0) + \left(\sum_{i=1}^{n} K_i\right)\log(1 - w_0)
\end{aligned}
$$

- 가우시안 혼합분포에서의 E-step을 위한 Q

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{(r)}) &= \int \log\left(f(\boldsymbol{Y}_o, \boldsymbol{Y}_m|\boldsymbol{\theta})\right) f(\boldsymbol{Y}_m|\boldsymbol{Y}_o, \boldsymbol{\theta}_{(r)}) d\boldsymbol{Y}_m \\
&= \int \left(\sum_{i=1}^{n} \log f(X_i, K_i|\boldsymbol{\theta})\right) f(\boldsymbol{K}|\boldsymbol{X}, \boldsymbol{\theta}_{(r)}) d\boldsymbol{K} \\
&= \sum_{i=1}^{n} \Big(\log f(X_i, K_i = 0|\boldsymbol{\theta}) P(K_i = 0|X_i, \boldsymbol{\theta}_{(r)}) \\
&\qquad + \log f(X_i, K_i = 1|\boldsymbol{\theta}) P(K_i = 1|X_i, \boldsymbol{\theta}_{(r)}) \Big)
\end{aligned}
$$

- 가우시안 혼합분포에서의 M-step을 위해 $\frac{\partial Q}{\partial \mu_0} = 0$, $\frac{\partial Q}{\partial \mu_1} = 0$, $\frac{\partial Q}{\partial \sigma_0^2} = 0$, $\frac{\partial Q}{\partial \sigma_1^2} = 0$, $\frac{\partial Q}{\partial w_0} = 0$을 풀면 다음과 같은 업데이트식이 구해지고

  가우시안 혼합분포에서의 $\boldsymbol{\theta}$를 추정하기위한 최종 EM알고리즘 이 된다.

$$
\begin{aligned}
\mu_{j(r+1)} &= \frac{\sum_{i=1}^{n} X_i \lambda_j(X_i, \boldsymbol{\theta}_{(r)})}{\sum_{i=1}^{n} \lambda_j(X_i, \boldsymbol{\theta}_{(r)})}, \; j = 0, 1. \\
\sigma^2_{j(r+1)} &= \frac{\sum_{i=1}^{n} (X_i - \mu_{j(r+1)})^2 \lambda_j(X_i, \boldsymbol{\theta}_{(r)})}{\sum_{i=1}^{n} \lambda_j(X_i, \boldsymbol{\theta}_{(r)})}, \; j = 0, 1 \\
w_{0(r+1)} &= \frac{1}{n} \sum_{i=1}^{n} \lambda_0(X_i, \boldsymbol{\theta}_{(r)}).
\end{aligned}
$$

여기서 $\lambda_j(X_i, \boldsymbol{\theta}_{(r)}) = P(K_i = j | X_i, \boldsymbol{\theta}_{(r)}), \; j = 0, 1.$

The End