Bayesian seminar

MLE & MAP & EM algorithm

염은지

목차

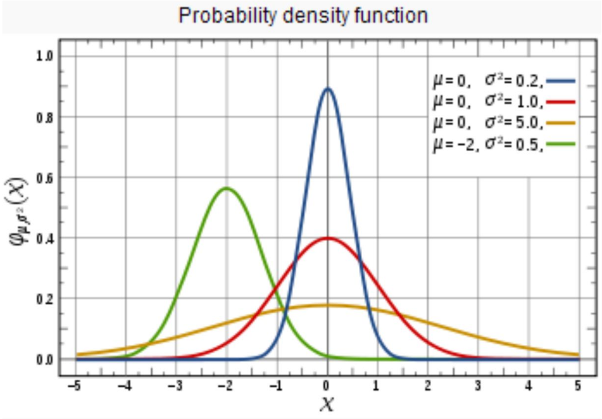
- Bayes Rule (Bayes Equation)
  - 조건부 확률 recap
- parameter estimation (모수 추정)
  - Maximum Likelihood Estimation(MLE)
  - Maximum A Posterior(MAP)
- EM(Expectation-Maximization)

Bayesian learning

- 우리가 관심을 갖는 값(모델,클래스)이 확률분포에 의해 좌우된다는 생각에서 시작
- 관측된 데이터와 함께 확률(모수)을 추론함으로써, 관심을 갖는 값에 대한 optimal한 decision을 내릴 수 있다는 것을 기초로 하는 learning

정규 분포 (Gaussian Distribution)

- 증명 : https://angeloyeo.github.io/2020/09/14/normal_distribution_derivation.html



$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (-\infty < x < \infty)$$

베이즈 정리

- P(H|E)
  - 사후 확률 (posterior)
  - 사건 E가 발생 후 갱신 된 사건 H의 확률
- P(H)
  - 사전 확률 (prior)
  - 사건 E가 발생하기 전에 가지고 있던 사건 H의 확률
- P(E|H)
  - 가능도 (likelihood)
  - 사건 H가 발생한 경우 사건 E의 확률
- P(E)
  - 정규화 상수 (normalizing constant) 또는 증거 (evidence)
  - 확률의 크기 조정
- 베이즈 정리는 근본적으로 사전확률과 사후확률 사이의 관계를 나타내는 정리
- 여기서… E = 표본 / H = 모수

$$\underbrace{\boxed{P(H|E)}}_{\substack{\text{사후 확률} \\ \text{(posterior)}}} = \frac{P(E|H)\overbrace{\boxed{P(H)}}^{\substack{\text{사전 확률} \\ \text{(prior)}}}}{P(E)}$$

베이즈 정리

- 베이즈 정리는 사건 E 가 발생함으로써 사건 H 의 확률이 어떻게 변화하는지를 표현한 정리
    - 사건 E 가 진실이라는 것을 알게 됨으로써
- 따라서 베이즈 정리는 새로운 정보가 기존의 추론에 어떻게 영향을 미치는지를 나타냄

베이즈 정리

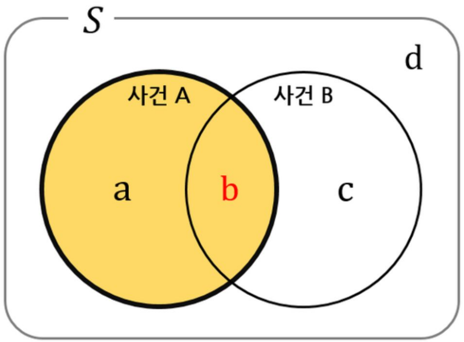
- 개념 recap

조건부 확률은 '**한 사건이 일어났다는 전제 하에서 다른 사건이 일어날 확률**' 입니다.

한 사건 A가 발생했다는 전제 하에서 다른 사건 B가 발생할 확률을 '조건부 확률'이라고 한다.

$$P(B|A)=\frac{P(A\cap B)}{P(A)} \quad (\text{단, } P(A)>0)$$

위에서 정의한 조건부 확률을 벤다이어그램으로 나타내면 다음 그림과 같습니다.



사건 A가 발생했다는 것이 전제해야 하므로 **a, b 영역**이 되어야 합니다. 이 a, b 영역 중에 **사건 B가 발생**한 것이기 때문에 **b영역**이 되어야 겠죠. 이것을 식으로 나타내면 다음과 같습니다.

$$P(B|A)=\frac{n(A\cap B)}{n(A)}=\frac{b}{a+b}$$

위 벤다이어그램에서 N=a+b+c+d 라 하면,

$$P(B|A)=\frac{b}{a+b}=\frac{\frac{b}{N}}{\frac{a+b}{N}}=\frac{P(A\cap B)}{P(A)}$$

P(B)는 조건부 확률로 표기해보면 P(B|S)는 사건 B의 확률을 구할 때 관심의 대상이 표본공간(S)이였다 S에서 B가 발생할 확률을 보았기 때문에 아래 사진 처럼 표기한다.

하지만 **P(B|A)에서는 표본공간(S)에서 사건 A로 축소 되었기 때문에 그 조건하에서 사건 B의 발생확률**이다.

- $$P(B)=P(B\mid S)=\frac{n(B)}{n(S)}$$
- 관심의 대상이 표본공간 S에서 사상 B 의 발생 확률

베이즈 정리

- 개념 recap

어떤 사건 A와 B가 있을 때, 아래와 같은 조건부확률을 정의할 수 있습니다.

$$P(A|B) = \frac{P(B \cap A)}{P(B)}$$

위 수식의 분자에 확률의 곱셈공식을 적용합시다.

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

위 등식이 베이즈정리입니다. $P(A|B)$ 를 구하고 싶은데, 직접 구하는 것이 어려운 대신 $P(A)$ 와 $P(B|A)$를 구하는 것은 상대적으로 쉽다면 위 등식은 쓸모가 있을겁니다.

곱셈법칙 (multiplication rule, 또는 곱셈공식)

조건부확률의 정의를 변형하면 아래와 같습니다.

$$P(A \cap B) = P(A)P(B|A)$$

이 수식의 의미를 생각해봅시다. A와 B가 둘다 발생한 확률은 A가 발행하고, A가 발생한 상황에서 B가 발생하면 된다 라고 해석할 수 있습니다.

이번에는 집합의 수를 3개로 확장해봅시다. 세 집합의 교집합은 아래와 같이 둘, 그리고 하나의 교집합으로 표현할 수 있습니다.

$$P(A \cap B \cap C) = P([A \cap B] \cap C)$$

위 곱셈법칙을 이용하여 변형합시다.

$$P(A \cap B \cap C) = P(A \cap B)P(C|A \cap B)$$

$P(A \cap B)$는 다시 아래와 같이 변형할 수 있습니다.

$$P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B)$$

베이즈 정리

- 개념 recap

서로 배반인 k개의 사건들 $A_1, A_2, \ldots, A_k$가 있다고 합시다. 이 사건들이 표본공간 S를 분할하고 있다고 합시다.

어떤 사건 B가 있다고 할 때, 아래 등식이 성립합니다. 위 조건과 상관없이 이건 그냥 당연히 성립합니다.

$P(B) = P(B \cap S)$

집합 A들이 표본공간을 분할하고 있으므로, 아래와 같이 변형가능합니다.

$P(B) = P(B \cap (A_1 \cup \ldots \cup A_k))$

분배법칙을 사용합시다.

$P(B) = P((B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_k))$

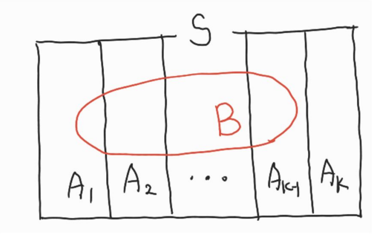
괄호안의 각 집합들은 배반이므로 아래 등식이 성립합니다.

$P(B) = P(B \cap A_1) + P(B \cap A_2) + \ldots + P(B \cap A_k)$

우변의 각 항에 곱셈공식을 적용합시다.

$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_n)P(B|A_k)$

여기서 아래 그림을 통해 위 식의 의미를 한번 생각하고 넘어갑시다.



일부러 교집합이 없는 경우가 생기게 그렸습니다. 교집합이 없는 항이 있어도 상관없습니다. 그냥 해당 확률이 0이 되는겁니다.

위 식에 시그마기호를 적용하면 아래와 같습니다.

$P(B) = \sum_{i=1}^{k} P(A_i)P(B|A_i)$

베이즈 정리

- 개념 recap

변형

표본공간 S는 사건 A에 의해 둘로 나뉩니다. A 와 A^c 입니다. 따라서 위 등식 분모에 전확률공식을 적용할 수 있습니다.

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A)+P(A^c)P(B|A^c)}$$

일반화

사건 $A_1, A_2, \ldots, A_n$ 이 표본공간을 분할하고 있다고 합시다. 또 다른 사건 B가 있을 때, 아래 등식이 성립합니다.

$$P(A_j|B) = \frac{P(B \cap A_j)}{P(B)}$$

곱셈공식을 적용합시다.

$$P(A_j|B) = \frac{P(A_j)P(B|A_j)}{P(B)}$$

전확률공식을 적용합시다.

$$P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum_{i=1}^{k} P(A_i)P(B|A_i)}$$

여기서 A_j 는 뭐든 될 수 있습니다. A_1 이건 A_2 이건, 뭐든 성립합니다.

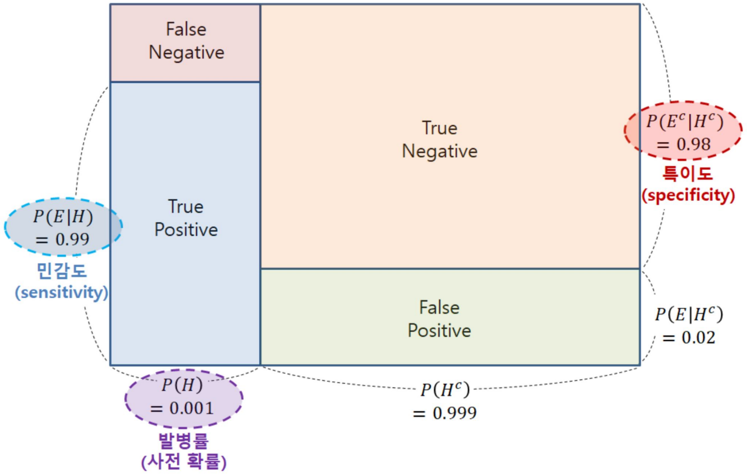
베이즈 정리

- 질병 A의 발병률은 0.1%로 알려져있다. 이 질병이 실제로 있을 때 질병이 있다고 검진할 확률(민감도)은 99%, 질병이 없을 때 없다고 실제로 질병이 없다고 검진할 확률(특이도)는 98%라고 하자.
- 만약 어떤 사람이 질병에 걸렸다고 검진받았을 때, 이 사람이 정말로 질병에 걸렸을 확률은?

✓ H : True이다. 실제로 병이 있다. ✓ $P(H) = 0.001$

✓ E : Positive로 측정되었다. 진단 O ✓ $P(E|H) = 0.99$

✓ $P(E^c|H^c) = 0.98$

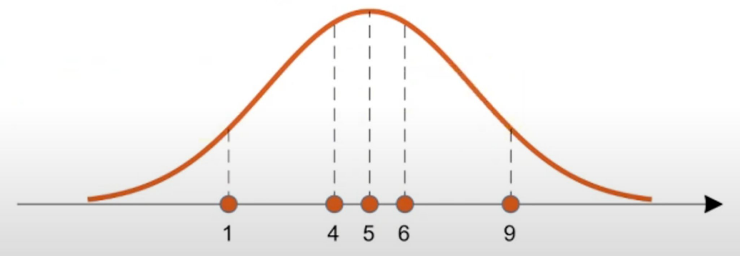


$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|H^c)P(H^c)} = \frac{0.001 \times 0.99}{0.001 \times 0.99 + 0.02 \times 0.999} = 0.047$$ (소숫점 세자리까지 반올림)

Maximum Likelihood Estimation (MLE)

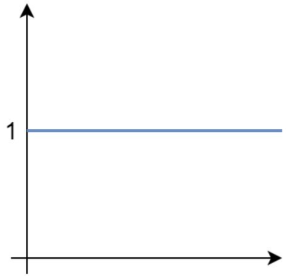
- Likelihood
  - 특정 사건(결과)이 일어날 확률
  - 데이터가 특정 분포로부터 만들어졌을(generate) 확률
  - 각 데이터 샘플에서 후보 분포에 대한 높이를 다 곱한 것
- MLE
  - 여러가지 모수 후보군에 대해서 높이를 다 곱한것(Likelihood)이 최대가 되게 하는 경우의 모수(Parameter) 분포가 우리가 데이터를 얻었을 것으로 추정되는 분포!

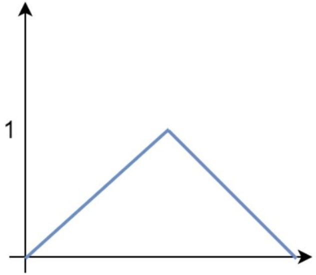
$$\Rightarrow \prod_{i=1}^{m} p(x^{(i)}|\theta)$$



Maximum Likelihood Estimation (MLE)

- maximum likelihood 기법은 머신러닝에서 모수를 추정할 때 가장 자주 쓰이는 개념 중에 하나
- 예시
  - X = (1, 1, 1, 1, 1)
  - 왼쪽 분포의 데이터 X에 대한 likelihood가 더 높다





Maximum Likelihood Estimation (MLE)

- 좀더 의미적으로 설명하자면?
  - 실험을 진행할 때 있는 dataset의 결과에 맞추도록, 즉 해당 결과가 발생할 확률이 최대가 되도록 확률을 정하는 것 -> 분류 데이터로 매핑하면 supervised learning의 classification 학습에 적용될 수 있음
- 이외에도 unsupervised learning ML과 DL에서도 모두 해석 가능
- detail : https://hyeongminlee.github.io/post/bnn002_mle_map/

Maximum Likelihood Estimation (MLE)

- 특정 결과는 독립적으로일어남 (independent)
  - 특정한 사건에 대한 여러개의 결과들이 연관이 있지는 않음
  - 예시) 주사위 던지기는 독립 시행

$$L(\theta) = p(X|\theta)$$

likelihood의 표현식. θ의 parameter를 가지는 분포.

$$L(\theta) = p(X|\theta) = \prod_{n=1}^{N} p(x_n|\theta)$$

likelihood 계산식

$$\mathrm{p}(\mathrm{x}_\mathrm{n} \mid \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(x_n-\mu)^2}{2\sigma^2}\}$$

데이터 x_n이 θ=(μ, σ)의 parameter를 가지는 정규분포를 따를 확률

Maximum Likelihood Estimation (MLE)

- 목적
  - 데이터 X가 θ의 parameter를 가지는 distribution을 따르려면 이 likelihood가 최대가 되는 distribution을 찾아야 함
- 과정
  - 계산의 편의를 위해 likelihood에 log와 -를(optional) 취하고 그 값이 최소가 되는 값을 구함으로써 maximum likelihood를 만들어주는 값을 구함
  - likelihood를 최대화하는 (-log likelihood를 최소화하는) θ 값을 찾을 찾으면 됨
  - log likelihood 식을 미분하고, 이 식이 0이 되는 값(최소값)을 찾음

$$E(\theta) = -\ln L(\theta) = -\sum_{n=1}^{N} \ln p(x_n|\theta)$$

log likelihood

$$\frac{\partial}{\partial\theta}E(\theta) = -\frac{\partial}{\partial\theta}\sum_{n=1}^{N} \ln p(x_n|\theta) = -\sum_{n=1}^{N} \frac{\frac{\partial}{\partial\theta}p(x_n|\theta)}{p(x_n|\theta)} \overset{!}{=} 0$$

위 식을 만족하는 θ 값을 찾아야 한다. 뒤 식이 저 형태인 이유는 ln(f(x))를 x에 대해 미분하면 f'(x)/f(x) 형태가 되기 때문이다.

Maximum Likelihood Estimation (MLE)

- 정규 분포 예시
  - 미분식을 0으로 만드는 parameter θ=(μ, σ)을 찾아야 함
  - 각각 평균(μ)와 표준편차(σ)에 대해 편미분 함으로써 maximum likelihood를 만들어주는 평균과 분산을 도출할 수 있음
  - detail : https://angeloyeo.github.io/2020/07/17/MLE.html

$$p(x_n|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{||x_n-\mu||^2}{2\sigma^2}}$$

정규분포의 p 값

$$\begin{aligned}
\frac{\partial}{\partial\mu}E(\mu,\sigma) &= -\sum_{n=1}^{N}\frac{\frac{\partial}{\partial\mu}p(x_n|\mu,\sigma)}{p(x_n|\mu,\sigma)} \\
&= -\sum_{n=1}^{N}-\frac{2(x_n-\mu)}{2\sigma^2} \\
&= \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu) \\
&= \frac{1}{\sigma^2}\left(\sum_{n=1}^{N}x_n - N\mu\right)
\end{aligned}$$

log likelihood를 미분하는 식.

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N}x_n$$

maximum likelihood mu 값

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n-\hat{\mu})^2$$

도출된 분산값

Maximum A Posterior (MAP)

- 목적
    - posterior를 최대로 하는 θ를 찾는 것
- 이미 데이터 x의 분포를 알고 있다고 가정 (P(x)는 상수)
- posterior를 최대로 하는 θ를 찾는 것은 likelihood와 prior를 곱한 값을 최대로 만드는 θ를 찾는 것과 같음
- 마찬가지로 log를 취하여 최적화 가능

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

posterior의 계산식

$$\hat{\theta}_{MAP} = argmax_{\theta} p(\theta|x) = argmax_{\theta} \frac{p(x|\theta)p(\theta)}{p(x)} = argmax_{\theta} p(x|\theta)p(\theta)$$

p(xtheta)p(theta)를 최대로 하는 theta값을 찾는 것이 MAP estimation

$$Posterior = \frac{Likelihood \times Prior}{Normalization\ Factor}$$

Maximum A Posterior (MAP)

- 수식과 예시 (정규 분포)

$$
\begin{aligned}
\hat{\theta}_{\mathrm{MAP}} &:= \arg\max_{\theta} p(\theta \mid \mathcal{D}) \\
&= \arg\max_{\theta} \frac{p(\mathcal{D} \mid \theta) p(\theta)}{p(\mathcal{D})} \\
&= \arg\max_{\theta} p(\mathcal{D} \mid \theta) p(\theta) \\
&= \arg\max_{\theta} \left[\log p(\mathcal{D} \mid \theta) + \log p(\theta)\right] .
\end{aligned}
$$

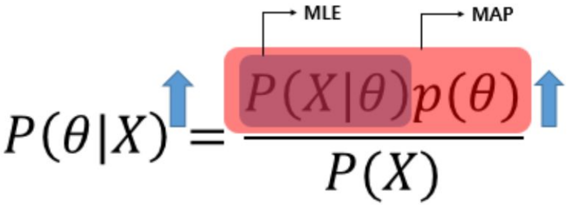
$$
\begin{aligned}
\mathcal{L}_{MAP}(\theta) &= \log p(\mathcal{D} \mid \theta) + \log p(\theta) \\
&\propto \left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{1}{2\alpha^2}\mu^2\right]
\end{aligned}
$$

Then, it follows from solving $\frac{\partial \mathcal{L}_{\mathrm{MAP}}}{\partial \mu} = 0$ that

$$
\hat{\mu}_{\mathrm{MAP}} = \frac{1}{\left(n + \frac{1}{\alpha^2}\right)} \sum_{i=1}^{n} x_i .
$$

MLE vs MAP

- 수식을 통한 간단 비교
  - P(x)는 상수이므로 제외 가능
  - 좌항 값을 알 수 없기 때문에 베이즈 정리를 통해 우항을 maximize 하는 방법으로 θ를 찾음
  - MAP는 우항 전체를 maximize
  - MLE는 P(θ)도 제외시키고 Likelihood term만 maximize



MLE vs MAP

- 수식을 통한 간단 비교

$$
\begin{aligned}
\theta_{MLE} &= \arg\max_{\theta} \log P(X|\theta) \\
&= \arg\max_{\theta} \log \prod_i P(x_i|\theta) \\
&= \arg\max_{\theta} \sum_i \log P(x_i|\theta)
\end{aligned}
$$

$$
\begin{aligned}
\theta_{MAP} &= \arg\max_{\theta} P(X|\theta)P(\theta) \\
&= \arg\max_{\theta} \log P(X|\theta) + \log P(\theta) \\
&= \arg\max_{\theta} \log \prod_i P(x_i|\theta) + \log P(\theta) \\
&= \arg\max_{\theta} \sum_i \log P(x_i|\theta) + \log P(\theta)
\end{aligned}
$$

MLE vs MAP

- prior가 uniform distribution일 경우
  - 예시
    - 분류문제의 경우, 클래스1과 클래스 2가 나타날 확률이 같다는 뜻
    - P(θ1)=P(θ2)=0.5
  - MAP는 MLE와 같아지게 되고, 따라서 MLE를 MAP의 특수한 경우로 생각할 수 있음

$$\begin{aligned}\theta_{MAP} &= \arg\max_{\theta} \sum_i \log P(x_i|\theta) + \log P(\theta) \\ &= \arg\max_{\theta} \sum_i \log P(x_i|\theta) + const \\ &= \arg\max_{\theta} \sum_i \log P(x_i|\theta) \\ &= \theta_{MLE}\end{aligned}$$

MLE vs MAP

가정 : 10만명에 대하여 성적 분포(100점 만점)를 구할 때, 10명만 샘플링 하여 parameter estimation을 수행하려고 한다.

- 10명이 약 70점을 평균값으로 뭉쳐있는데, 사실 대부분의 데이터는 '80점'정도를 중심으로 Gaussian을 이루는 경험(지식)을 가지고 있다면 어떻게 될까?
  - MLE는 10분의 데이터에만 의존하므로, 70점 근처로 결과를 얻음
  - MAP는 '80'점 근처라는 prior knowledge를 적용하여, 대략 70~80점 근처의 결과를 얻게됨

MLE vs MAP

- 어떤 사람의 통장잔고(x)를 보고, 그 사람이 게임을 하는 사람인지 안 하는 사람인지(theta)를 판단한다면?
    - MLE
        - 게임을 하는 사람들 중 그 통장잔고가 나올 확률과 게임을 하지 않는 사람들 중 그 통장잔고가 나올 확률을 비교(likelihood)하여 둘 중 더 높은 확률로 선택
        - 이 경우에는 게임을 하는 사람과 게임을 하지 않는 사람의 비율(prior)은 결정에 반영되지 않음
    - MAP
        - 통장잔고가 주어졌을 때, 그것이 게임을 하는 사람의 것일 확률과 게임을 하지 않는 사람의 것일 확률을 비교(posterior)하여 둘 중 더 높은 확률로 선택하는 것
        - 이 경우에는 게임을 하는 사람과 게임을 하지 않는 사람의 비율(prior)은 결정에 반영됨
- 우리가 알고 있는 사전 정보인 prior의 정보를 likelihood에 곱하여 반영함으로써 더 정확한 판단을 내리는 것이 MAP estimation
    - 우리가 알고있는사전정보?
        - 지식 / 경험 / 현실세계 반영

MLE vs MAP

- prior이 더해지게 되면 이는 MAP에서 일종의 regularizer의 역할을 하게됨
- Machine learning에서 흔히 볼 수 있는 issue : overfitting
  - Overfitting이란 어떠한 data를 설명하기 위한 model이 있다고 했을 때 likelihood가 1에 가까워지게 되는 상황
  - 우리는 이 값을 조금 잃더라도 overfitting을 막는 것이 더 중요함
- 예시
  - 동전 던지기를 예시로 봤을 때 실제로 앞면의 확률이 0.5라고 했을 때 일반적으로 시도 횟수가 많아질수록 MLE를 통해서 0.5에 수렴하게 될 것이지만, 시도 횟수가 적을 때는 0.5와는 다른 특정 확률을 추정하게 되고 이는 적은 data에 맞춰져서 overfitting이 발생하는 것과 동일한 상황으로 볼 수가 있음
- 이러한 상황을 막고자 belief로 prior를 함께 고려하는 것이고, 이를 overfitting을 막는다는 관점에서 regularizer로 볼 수 있음
- 충분히 많은 양의 sample이 존재한다면 prior의 의미가 희석되게 되므로 굳이 prior를 사용할 필요는 없음
  - 그러나 보통 sample의 양이 적은 상황이 대부분이므로 적절한 prior를 선택해서 사용
- 어떠한 prior도 없다면 MLE를 사용하기 좋은 상황이 됨

EM(Expectation-Maximization) Algorithm

- 상황
  - 모수에 대한 추정치를 구해야 하는 상황에서 MLE를 구하기 위한 완전한 정보가 없다!
- 정의
  - 관측되지 않는 잠재변수(Latent variable)가 존재하는 확률 모델에서 MLE나 MAP을 갖는 모수의 추정값을 찾는 반복적인 (iterative) 알고리즘
- 통계 모델의 수식을 정확히 풀 수 없을 때, MLE를 구하는데 사용
  - MLE를 통해 관측된 데이터에 알맞은 모델의 변수를 추정
- E단계와 M단계를 반복하며 새로운 모수 추정값과 기존의 모수 추정값의 차이가 매우 작아질 때까지(수렴) 반복
  - Expectation
  - Maximization

EM(Expectation-Maximization) Algorithm

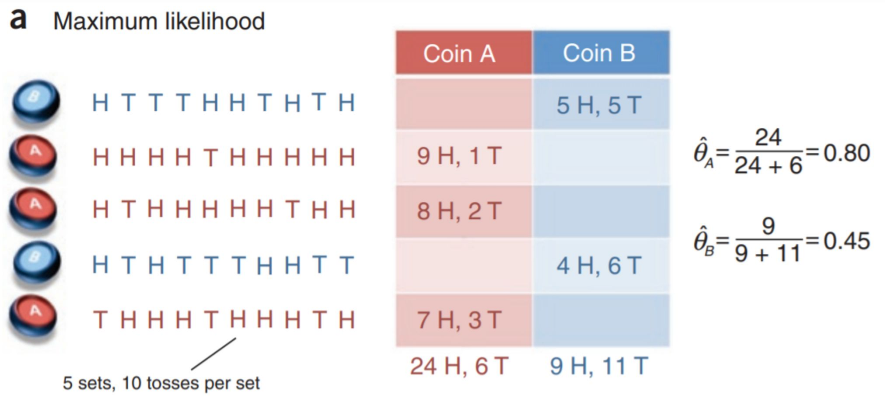
- 과정
  - 모수 초기값 세팅
  - Iter1
    - E-Step : 세팅한 값 기반으로(관찰되지 않은 결과를 포함한) Likelihood의 기대값 계산
    - M-Step : likelihood의 기대값을 최대화 하는 모수 추정값을 산출
  - Iter2
    - E-Step : 산출한 추정값 기반으로 Likelihood의 기대값 계산
    - M-Step : likelihood의 기대값을 최대화 하는 모수 추정값을 산출
  - … (반복)

EM(Expectation-Maximization) Algorithm

- 예시
  - 상황 : 동전 A/B 중에 하나를 선택하고, 해당 동전을 10번 던져서 앞면(H)과 뒷면(T)이 나온 결과를 차례대로 적는다 (5번 반복)
  - 목적 : 동전 A/B의 각각 앞면과 뒷면이 나올 확률은?

EM(Expectation-Maximization) Algorithm

- 예시 (State Known)
    - 각 차수마다 어떤 동전을 선택했는지 알고있는 경우 -> 한 차수에는 한 동전에 대한 가능성만 계산
    - MLE 사용

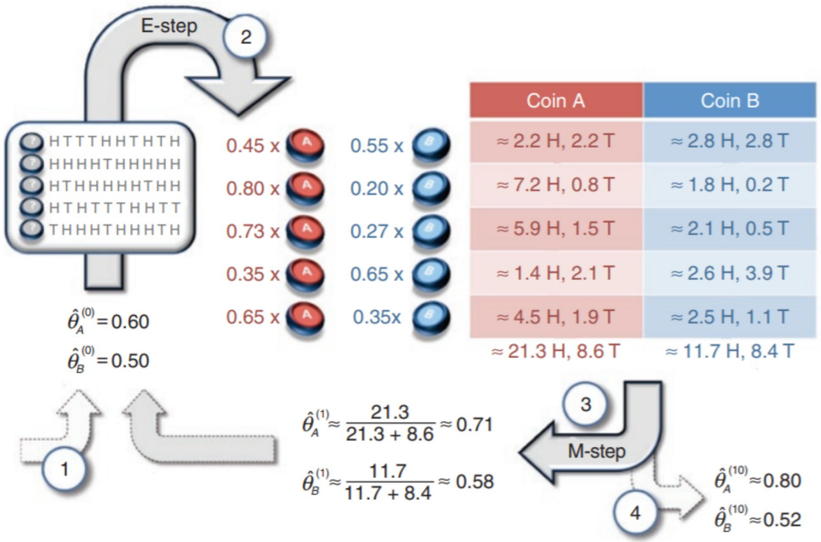


EM(Expectation-Maximization) Algorithm

- 예시 (State Unknown)
  - 각 차수마다 어떤 동전을 선택했는지 모르는 경우 -> 각 차수에 어떤 동전이 선택되었는지 추가로 추측해야 함
  - 사용된 동전의 수 : Hidden variable / Latent variable
  - 단계
    i. 추정값(각 동전 당 앞면이 나올 확률) 초기화
    ii. E-Step : Hidden variable의 responsibility 계산
        - responsibility? : 부담률, 전체 클러스터(A/B 동전 2개)에 대한 입력데이터(A or B 동전)의 클러스터 확률
        - state known의 경우는 responsibility가 0 or 1로 명확했음
    iii. M-Step
        - 추정값 업데이트

EM(Expectation-Maximization) Algorithm

- 예시 (State Unknown)
  - E-Step
    1. 1차수에 대해 계산 (아래 설명)
    2. 추정값이 A가 더 높으므로 H가 많이 나온 sequence에서는 A의 responsibility가 더 높음
    3. 모든 차수에 대해 실행



- 동전 A를 사용한 경우 $a = 0.6^5 * 0.4^5 = 7.96e - 05$
- 동전 B를 사용한 경우 $b = 0.5^5 * 0.5^5 = 9.76e - 05$
- 1차수에서 동전 A를 사용했을 비율(responsibility): $a/(a+b) = 0.45$
- 1차수에서 동전 B를 사용했을 비율(responsibility): $b/(a+b) = 0.55$

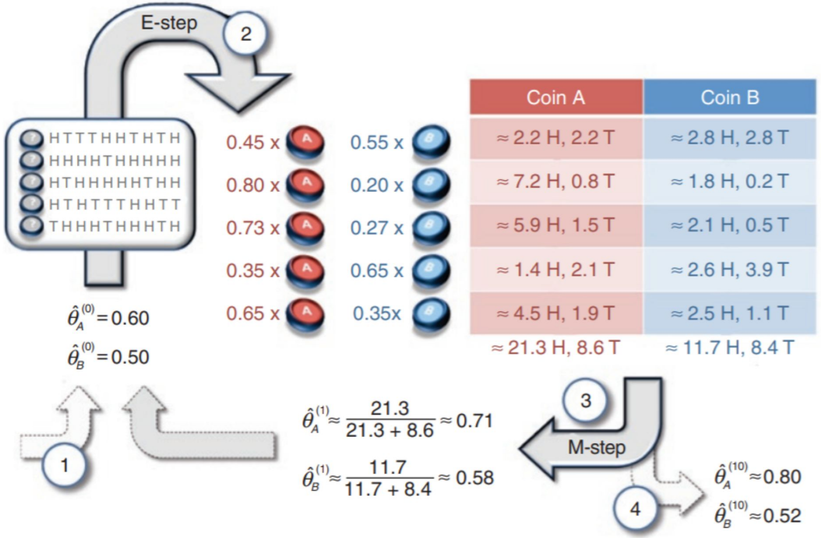
EM(Expectation-Maximization) Algorithm

- 예시 (State Unknown)
  - M-Step
    1. E-Step에서 계산한 responsibility를 사용해 확률 계산 (아래 설명)
  - 이후 E -> M -> E -> M -> .. Step을 반복하여 local maximum으로 답을 찾을 수 있음

- 1차수 sequence : HTTTHHTHTH (H: 5번 T: 5번)
- 동전 A인 경우 1차수 H개수: 0.45*5 = 2.2
- 동전 A인 경우 1차수 T개수: 0.45*5 = 2.2
- 동전 B인 경우 1차수 H개수: 0.55*5 = 2.8
- 동전 B인 경우 1차수 T개수: 0.55*5 = 2.8
- …

- 동전 A인 경우 전체 H : 21.3
- 동전 A인 경우 전체 T : 8.6
- 동전 B인 경우 전체 H : 11.7
- 동전 B인 경우 전체 T : 8.4

추정값 update
- $\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3+8.6} \approx 0.71$
- $\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7+8.4} \approx 0.58$



EM(Expectation-Maximization) Algorithm

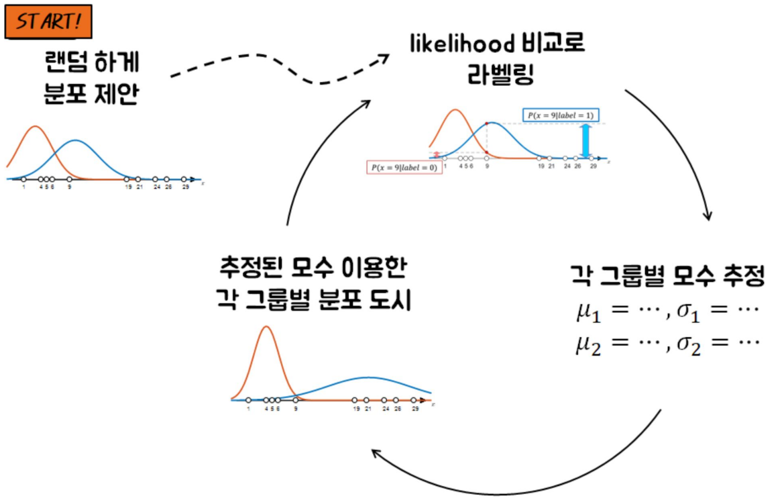
- 예시에서는 Hidden variable을 2개로 가정하였으나, 다른 개수일수도 있음
- variable에 대한 가정이 잘 맞아야 결과도 좋게 나올 수 있음
- EM 알고리즘은 정답 label 없이 전혀 알지 못하는 hidden variable에 대한 값을 풀 수 있다는 특징이 있으며, 가정을 잘 세웠다면 해당 알고리즘을 적절히 활용하여 원하는 결과를 얻을 수 있음
- 수렴 속도가 느리다는 단점이 있음 (iterative하게 계산)

EM(Expectation-Maximization) Algorithm

- 응용 분야 : GMM (Gaussian Mixture Model)
  - label 이 주어지지 않은 데이터셋에 대해 N개의 정규분포를 이룰 것이라 가정하고 clustering을 수행 (label 별 분포 추정)
  - N개의 정규분포? : Gaussian Mixture!

EM(Expectation-Maximization) Algorithm

- 응용 분야 : GMM (Gaussian Mixture Model)



EM(Expectation-Maximization) Algorithm

- 응용 분야 : GMM (Gaussian Mixture Model)

Repeat until convergence: {

(E-step) For each i, j, set

$$w_j^{(i)} := p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma) \tag{7}$$

(M-step) Update the parameters:

$$\phi_j := \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)} \tag{8}$$

$$\mu_j := \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}} \tag{9}$$

$$\Sigma_j := \frac{\sum_{i=1}^{m} w_j^{(i)} \left(x^{(i)} - \mu_j\right) \left(x^{(i)} - \mu_j\right)^T}{\sum_{i=1}^{m} w_j^{(i)}} \tag{10}$$

}

여기서 i는 데이터의 순번, j는 라벨을 의미하며, $x^{(i)}$는 i번째 데이터셋, $z^{(i)}$는 i번째 데이터셋의 라벨을 의미한다.

또 $w_j^{(i)}$는 i번째 데이터가 j번 그룹에 속할 확률을 의미한다.

ϕ는 그룹들 간의 데이터 비율을 의미한다. 가령, 0번 그룹과 1번 데이터는 6:4로 존재한다고 하면 $\phi = [0.6, 0.4]$이다.

μ_j는 j번 그룹의 평균값을 의미하고, Σ_j는 표준편차 혹은 공분산행렬을 의미한다.

EM(Expectation-Maximization) Algorithm

- 응용 분야 : GMM (Gaussian Mixture Model)
  - E-step
    - likelihood
      - 분포의 높이에 대한 값
    - prior
      - 모든 데이터에 대해서 각 그룹에 있을 평균 확률 (비율)
    - evidence
      - 모든 그룹에 대해서 likelihood * prior를 더한 것
      - 확률로 만들기 위한 normalize factor
  - M-step

M-step에서는 E-step에서 계산한 $w_j^{(i)}$ 값들을 이용해 모수를 추정하는 과정이다.

이 모수들은 최대우도법을 이용하여 모두 쉽게 계산할 수 있다. ($x^{(i)}$는 모두 주어진 데이터라는 점을 꼭 생각하자.)

그것은 우리에게 $x^{(i)}$라는 데이터가 주어졌고, ϕ, μ, Σ 라는 모수를 줌으로써 각 label에 대한 확률 분포를 가정했을 때

그 분포의 높이에 따라 (i)번째 데이터가 j번 그룹에 속할 확률을 계산하겠다는 뜻이다.

예를 들어 $w_j^{(i)}$라는 값은 그룹이 총 3개 였다고 하고, 0, 1, 2번 그룹에 속할 확률이 각각 0.8, 0.15, 0.05라고 한다면

$$w_0^{(i)} = p(z^{(i)} = 0|x^{(i)}) = 0.8 \tag{11}$$

$$w_1^{(i)} = p(z^{(i)} = 1|x^{(i)}) = 0.15 \tag{12}$$

$$w_2^{(i)} = p(z^{(i)} = 2|x^{(i)}) = 0.05 \tag{13}$$

라고 쓸 수 있는 것이다. (i번째 데이터에 대해서 각 그룹에 속할 확률을 모두 더하면 1이 되어야 한다는 점도 빠뜨리지 말자.)

여기서 조금 더 구체적인 확률 계산은 베이즈 정리에 따라 수행할 수 있다.

식은 복잡해 보이지만, 해당 label에 포함될 Prior x likelihood 값을 각 label에 포함될 Prior x likelihood 값을 모두 더한 값으로 나눈 것이 (i)번째 데이터가 j번 그룹에 속할 확률이 된다.

식을 써보자면 다음과 같다.

$$p(z^{(i)} = j|x^{(i)}; \phi, \mu, \Sigma) = \frac{p(x^{(i)}|z^{(i)} = j; \mu, \Sigma)p(z^{(i)} = j; \phi)}{p(x^{(i)}; \phi, \mu, \Sigma)} \tag{14}$$

$$= \frac{p(x^{(i)}|z^{(i)} = j; \mu, \Sigma)p(z^{(i)} = j; \phi)}{\sum_{k=1}^{l} p(x^{(i)}|z^{(i)} = k; \mu, \Sigma)p(z^{(i)} = k; \phi)} \tag{15}$$

끝!