Ch 05. EM 알고리즘

1. KL 발산
2. EM 알고리즘 도출 ①
3. EM 알고리즘 도출 ②
4. GMM과 EM 알고리즘
5. EM 알고리즘 구현

5.1 KL 발산

우선 표기법을 변경!

$E_{p(x)}[f(x)] = \int f(x)p(x)\,dx$ ⇒ 연속 확률 변수 x가 있고 그 확률밀도가 $p(x)$ 일 때,
함수 $f(x)$의 기댓값을 의미!

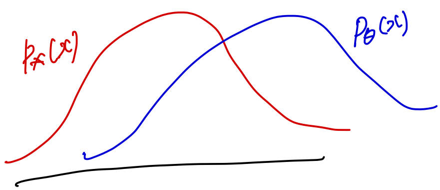
KL 발산 (쿨백-라이블러 발산, Kullback-Leibler divergence) 란?!?!
⇒ 두 확률 분포의 차이!를 측정하는 척도!

$\Rightarrow D_{KL}(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} dx$ 또는! $\sum_x p(x) \log \frac{p(x)}{q(x)}$

KL 발산의 특징은?

1. 두 확률분포가 다를수록 값이 커진다!
2. 두 확률분포가 같으면 0이다!
3. 같지 않으면 0보다 크다!
4. $D_{KL}(p\|q)$와 $D_{KL}(q\|p)$는 다르다!

KL 발산과 최대 가능도 추정의 관계



$p_*(x) \xrightarrow{\text{샘플링}} \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$ ⇒ 이것으로 $p_\theta(x)$ 최대가능도!

즉! $\log \prod_{n=1}^{N} p_\theta(x^{(n)}) = \sum_{n=1}^{N} \log p_\theta(x^{(n)})$

따라서! $\hat{\theta} = \arg\max_\theta \sum_{n=1}^{N} \log p_\theta(x^{(n)})$

핵심 아이디어는!

$p_\theta(x)$를 $p_*(x)$에 최대한 가깝게 만든다
⇕
$p_*(x)$와 $p_\theta(x)$의 KL 발산을 최소화한다

즉! $D_{KL}(p_*\|p_\theta) = \int p_*(x) \log \frac{p_*(x)}{p_\theta(x)} dx$를 계산!

그러나! $p_*(x)$를 정확하게 모른다...
따라서! 몬테카를로 방법 (Monte Carlo method)을 사용

몬테카를로 방법은?

⇒ 확률분포나 기댓값의 근사값을 구하는 기법!

⇒ 무작위 샘플 기반의 결과값의 평균을 구해서 근사!

⇒ 근사 방법은?

1. $p_*(x)$에 따라 샘플 $\{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$을 생성!

2. 각 데이터 $x^{(i)}$에서 $f(x^{(i)})$를 구하고 평균을 구한다

즉! $E_{p_*(x)}[f(x)] = \int p_*(x) f(x)\, dx$

$$\approx \frac{1}{N}\sum_{n=1}^{N} f(x^{(n)}) \quad (x^{(n)} \sim p_*(x))$$

몬테카를로를 KL발산에 적용하면?

$$D_{KL}(p_* \| p_\theta) = \int p_*(x) \log \frac{p_*(x)}{p_\theta(x)}\, dx$$

$$\approx \frac{1}{N}\sum_{n=1}^{N} \log \frac{p_*(x^{(n)})}{p_\theta(x^{(n)})} \quad (x^{(n)} \sim p_*(x))$$

$$= \frac{1}{N}\sum_{n=1}^{N} \left(\log p_*(x^{(n)}) - \log p_\theta(x^{(n)})\right)$$

여기서!, 가장 작은 KL은?

$$\arg\min_\theta D_{KL}(p_* \| p_\theta)$$

$$\approx \arg\min_\theta \left(-\frac{1}{N}\sum_{n=1}^{N} \log p_\theta(x^{(n)})\right)$$

$$= \arg\max_\theta \sum_{n=1}^{N} \log p_\theta(x^{(n)})$$

즉! $\underbrace{\arg\min_\theta D_{KL}(p_* \| p_\theta)}_{\text{최소 KL발산!}} \approx \underbrace{\arg\max_\theta \sum_{n=1}^{N} \log p_\theta(x_n)}_{\text{곧 최대가능도!}}$

5.2 EM 알고리즘 도출 ①

모델 + 잠재변수 → 확률의 수식화 → 로그 가능도

여러 개의 데이터 $D=\{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$

하나의 데이터 → $\log p_\theta(x) = \log \sum_z p_\theta(x, z)$

매개변수 (θ), 관찰 가능한 확률 변수 (x), 잠재변수 (z)

$$\log p_\theta(D) = \log\left(p_\theta(x^{(1)})\, p_\theta(x^{(2)}) \cdots p_\theta(x^{(N)})\right)$$
$$= \sum_{n=1}^{N} \log p_\theta(x^{(n)})$$
$$= \sum_{n=1}^{N} \log \sum_{z^{(n)}} p_\theta(x^{(n)}, z^{(n)})$$

이때! "로그-합"이라서 어려워!
이를 위해, 임의 확률 $q(z)$를 사용

그러면!? 1단계!

$$\log p_\theta(x) = \log \frac{p_\theta(x,z)}{p_\theta(z|x)}$$
$$= \log \frac{p_\theta(x,z)\, q(z)}{p_\theta(z|x)\, q(z)}$$
$$= \underbrace{\log \frac{p_\theta(x,z)}{q(z)} + \log \frac{q(z)}{p_\theta(z|x)}}$$

아직 로그-합이야!
그래서 이걸 KL발산으로 바꿔!

⇒

그러면 2단계!

$$\log p_\theta(x)$$
$$= \log p_\theta(x) \sum_z q(z)$$
$$= \sum_z q(z) \log p_\theta(x)$$
$$= \sum_z q(z) \left(\log \frac{p_\theta(x,z)}{q(z)} + \log \frac{q(z)}{p_\theta(z|x)}\right)$$
$$= \sum_z q(z) \log \frac{p_\theta(x,z)}{q(z)} + \sum_z q(z) \log \frac{q(z)}{p_\theta(z|x)}$$
$$= \underbrace{\sum_z q(z) \log \frac{p_\theta(x,z)}{q(z)} + D_{KL}(q(z), p_\theta(z|x))}$$

이건 KL발산!

이러면 합-로그!
즉! 극복!

5.3 EM 알고리즘 도출 ②

ELBO (Evidence Lower BOund, 증거하한) 은!? → 가장 낮은 로그 가능도!?
에이.. 아무리 낮아도 이것보다 가능도가 낮을까.. 요런 느낌!?

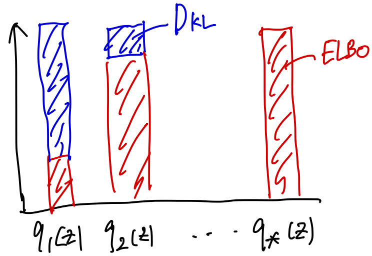
즉!

$$\log p_\theta(x) = \sum_z q(z)\log\frac{p_\theta(x,z)}{q(z)} + \underbrace{D_{KL}(q(z)\,||\,p_\theta(z|x))}_{\geq 0}$$

$$\geq \sum_z q(z)\log\frac{p_\theta(x,z)}{q(z)} \quad \Longrightarrow \text{요것이!?} \quad \text{ELBO!!!?} \quad \xRightarrow[\text{즉!?}]{} \quad ELBO(x;q,\theta) = \sum_z q(z)\log\frac{p_\theta(x,z)}{q(z)}$$

자 그럼

$$\log p_\theta(x) = ELBO(x;q,\theta) + D_{KL}(q(z)\,||\,p_\theta(z|x)) \xRightarrow[\text{즉!}]{\log p_\theta(x)\text{는 일정}} \text{ELBO}\uparrow\ D_{KL}\downarrow \ \text{ or } \ \text{ELBO}\downarrow\ D_{KL}\uparrow$$



즉!? ⟹ DKL 최소화 = ELBO 최대화 = 최대가능도!?

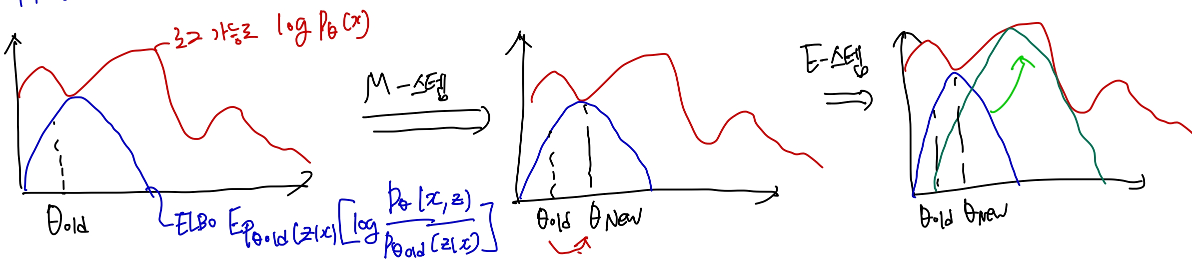
그럼!? 우선 q(z) 최적화!? ⟹ $\theta = \theta_{old}$로 고정 $\xRightarrow{D_{KL}=0}$ $p_{\theta_{old}}(z|x) = q(z)$

↳ $ELBO(x; q=p_{\theta_{old}}(z|x), \theta) = \sum_z p_{\theta_{old}}(z|x)\log\frac{p_\theta(x,z)}{p_{\theta_{old}}(z|x)}$

이것이 E-스텝!?!? (Expectation value-step) ⟹ $= E_{p_{\theta_{old}}(z|x)}\left[\log\frac{p_\theta(x,z)}{p_{\theta_{old}}(z|x)}\right]$

그럼!? 다음은 θ 최적화 ⟹ θ에 대한 미분!? ⟹ 최적 θ 발견!? ⟹ 이것이 M-스텝 (Maximization-step)

이후 E-스텝, M-스텝 반복.. 즉!?



2절 데이터가 여러 개일 때?

관측 데이터 $= \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$, 임의 확률 분포 $= \{q^{(1)}, q^{(2)}, \dots, q^{(N)}\}$ 일 때?

2절! 로그 가능도와 ELBO의 관계는?

$$\sum_{n=1}^{N} \log \sum_{z^{(n)}} p_\theta(x^{(n)}, z^{(n)}) \geq \sum_{n=1}^{N} \mathrm{ELBO}(x^{(n)}; q^{(n)}, \theta)$$
$$= \sum_{n=1}^{N} \sum_{z^{(n)}} q^{(n)}(z^{(n)}) \log \frac{p_\theta(x^{(n)}, z^{(n)})}{q^{(n)}(z^{(n)})}$$

EM 알고리즘은!

1. E-스텝 : $\{q^{(1)}, q^{(2)}, \dots, q^{(N)}\}$ 갱신 (θ 고정)

   → 각 n에 대해 $q^{(n)}(z) = p_\theta(z \mid x^{(n)})$으로 설정.

2. M-스텝 : θ 갱신 ($\{q^{(1)}, q^{(2)}, \dots, q^{(N)}\}$ 고정)

   → $\sum_{n=1}^{N} \mathrm{ELBO}(x^{(n)}; q^{(n)}, \theta)$가 최대가 되는 θ 갱신

3. 종료 : 이전 로그 가능도와 이후 로그 가능도 비교

즉! GMM에 대한 EM은?

E-스텝은 각 n, k에 대해

$$q^{(n)}(k) = \underbrace{\frac{\phi_k \mathcal{N}(x^{(n)}; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \phi_j \mathcal{N}(x^{(n)}; \mu_j, \Sigma_j)}}_{\text{전체 GMM 중 k 파라미터 분포가 잡힐 확률?!?!}}$$

종료 판정은

$$\underbrace{\frac{1}{N} \overbrace{\sum_{n=1}^{N} \log \underbrace{\sum_{j=1}^{K} \phi_j \mathcal{N}(x^{(n)}; \mu_j, \Sigma_j)}_{\text{GMM 수식}}}^{\text{GMM 로그 가능도}}}_{\text{GMM 평균 로그 가능도}}$$

M-스텝은 각 k에 대해

$$\phi_k = \frac{1}{N} \sum_{n=1}^{N} q^{(n)}(k)$$

$$\mu_k = \frac{\sum_{n=1}^{N} q^{(n)}(k)\, x^{(n)}}{\sum_{n=1}^{N} q^{(n)}(k)}$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} \left(q^{(n)}(k) (x^{(n)} - \mu_k)(x^{(n)} - \mu_k)^T \right)}{\sum_{n=1}^{N} q^{(n)}(k)}$$

⟹ 모든 데이터에 대한 계산!
k 번째 파라미터 기반으로?
전체 데이터에 대입?!?!
이후, 가장 적절한 파라미터 선택?!?!