EM 알고리즘

**확률 : 확률분포가 주어졌을 때의 데이터의 특정 확률*

ex) 평균이 23, 분산이 5인 정규분포에서 값이 21~25일 확률

**가능도 : 데이터가 주어졌을 때 특정 확률 분포일 확률*

ex) 값이 24, 27, 28 일 때, 평균이 23, 분산이 5인 정규분포일 확률

▶ EM 알고리즘

- 잠재변수가 존재하는 확률모형의 문제를 풀기 위한 알고리즘 중 하나.

▶ 잠재변수(latent variable)

- 직접 관찰/측정 되는 것이 아닌 연구자가 설정한 확률변수
- 문제를 더 쉽게 하기 위해서 사용

▶ 잠재변수의 활용

- $P(X|\theta)$를 구하기 위해 잠재변수를 활용

- $P(X|\theta) = \sum_z P(X, Z|\theta)$

-> $\mathrm{argmax}_\theta P(X|\theta) = \mathrm{argmax}_\theta \log P(X|\theta) = \mathrm{argmax}_\theta \log \sum_z P(X, Z|\theta)$

- $q(Z)$를 도입 *$q(Z)$ *: 잠재변수의 확률밀도함수(확률질량함수)*

- $\ln \sum_z P(X, Z|\theta) = \ln \sum_z q(Z) \frac{P(X, Z|\theta)}{q(Z)}$

- $\sum_z q(Z) \frac{P(X, Z|\theta)}{q(Z)}$ 뜯어보기

- $q(Z)$ -> z에 대한 확률밀도함수, $\frac{P(X, Z|\theta)}{q(Z)} = z'$ -> z에 의해 생성되는 확률변수의 함수

- $\therefore \sum_z q(Z) \frac{P(X, Z|\theta)}{q(Z)} = E_{z'}$

- Jensen's Inequality(옌슨 부등식) 도입

**Jensen's Inequality(여기선 x가 z'가 된 것이라고 보면 됨*

- convex함수일 때 : $E[f(x)] \geqq f(E[x])$
- concave함수일 때 : $E[f(x)] \leqq f(E[x])$

-> $\ln \sum_z q(Z) \frac{P(X, Z|\theta)}{q(Z)} \geqq \sum_z q(Z) \ln\left(\frac{P(X, Z|\theta)}{q(Z)}\right) = L(q, \theta)$ **log는 concave 함수이기 때문*

-> $L(q, \theta)$가 log likelihood의 Lower bound이고, 이 lower bound를 최대화

- log likelihood와 $L(q,\theta)$의 차이

$$\ln P(X|\theta) - L(q,\theta) = \ln P(X|\theta) - \sum_z q(Z)\ln(\frac{P(X,Z|\theta)}{q(Z)})$$
$$= \ln P(X|\theta) - \sum_z q(Z)\ln(\frac{P(Z|X,\theta)P(X|\theta)}{q(Z)})$$
$$= \ln P(X|\theta) - \sum_z q(Z)(\ln(\frac{P(Z|X,\theta)}{q(Z)}) + \ln P(X|\theta))$$
$$= \ln P(X|\theta) - \sum_z q(Z)\ln(\frac{P(Z|X,\theta)}{q(Z)}) - \ln P(X|\theta)\Sigma_z q(Z)$$
$$= -\sum_z q(Z)\ln(\frac{P(Z|X,\theta)}{q(Z)})$$

- $-\sum_z q(Z)\ln(\frac{P(Z|X,\theta)}{q(Z)})$ 살펴보기

- $-\sum_z q(Z)\ln(\frac{P(Z|X,\theta)}{q(Z)}) = \sum_z q(Z)\ln(\frac{q(Z)}{P(Z|X,\theta)}) = KL(q(Z)||P(Z|X,\theta))$

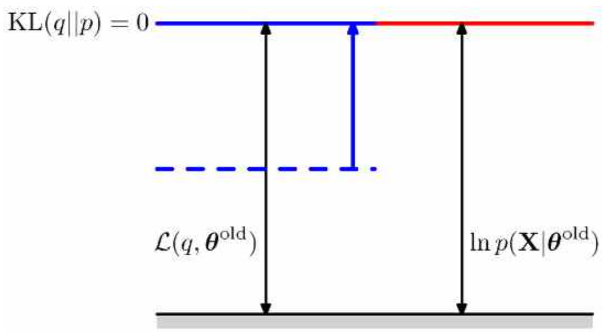
*KL-divergence -> $\sum_i p_i \ln\frac{p_i}{q_i}$

▶ E-step

- 정해진 θ(M-step을 통해 or 초기 임의값)값을 이용하여 가장 적절한 $q(Z)$를 찾는 과정
- $\ln(P(X|\theta))$를 고정시켰다면, lower bound와 $\ln(P(X|\theta))$를 같게하는 $q(Z)$가 최적값일 것

-> $KL(q(Z)||P(Z|X,\theta)) = 0$을 만들겠다.

=> $q(Z) = P(Z|X,\theta))$를 통해 $q(Z)$를 구함



▶ M-step

- 정해진 $q(Z)$(E-step을 통해)를 통해 가장 적절한 θ를 구하는 과정
- $KL(q(Z)||P(Z|X,\theta))$가 고정되어 있기 때문에, $L(q,\theta)$를 최대화시키면 log likelihood 값 또한 증가될 것.

- E-step에서 구한 $q(Z)$를 $L(q,\theta)$에 대입

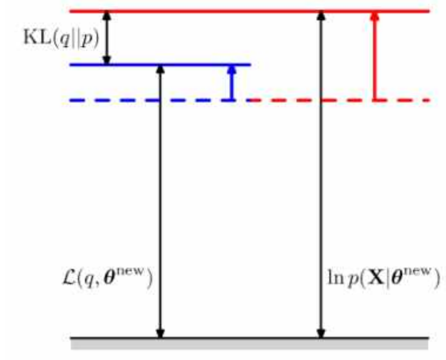
- $L(q,\theta) = \sum_z (P(Z|X,\theta^{old})\ln \frac{P(X,Z|\theta)}{P(Z|X,\theta^{old})})$
  $= \sum_z (P(Z|X,\theta^{old})(\ln P(X,Z|\theta) - \ln P(Z|X,\theta^{old})))$
  $= \sum_z (P(Z|X,\theta^{old})\ln P(X,Z|\theta)) - \Sigma_z (P(Z|X,\theta^{old})\ln P(Z|X,\theta^{old}))$

• $\sum_z (P(Z|X,\theta^{old})\ln P(X,Z|\theta)) - \sum_z (P(Z|X,\theta^{old})\ln P(Z|X,\theta^{old}))$ 뜯어보기

- 앞의 항 : θ에 대한 방정식 -> $E_{\ln(P(X,Z|\theta))}$

- 뒤의 항 : θ에 대해 관계없는 식 -> 상수

=> 앞의 항을 최대화하는 파라미터를 구함



▶ **총정리**

- EM 알고리즘이란 잠재변수가 있는 확률모형을 풀기 위해 반복적인 과정을 거침

<순서>

1. 임의의 파라미터 θ^{old}로부터 E-step을 통해 $L(q)$를 풀이 -> q를 구함
2. 구해진 q를 이용하여 M-step을 통해 $L(\theta)$를 풀이 -> 최적의 θ를 구함
3. M-step에서 구해진 θ를 이용하여 E-step으로 돌아감
4. 수렴 조건을 만족할 때까지 반복