

KL-Divergence

크로스 엔트로피 $H(p, q)$ 는 모델에서 예측한 확률(p)과 정답확률(q)을 모두 사용해 측정된 값이다.

학습이 진행될 수록 $H(p, q)$ 는 줄어든다.

KL-Divergence는 두 확률분포의 차이를 계산한다. 어떤 이상적인 분포에 대해, 그 분포를 근사하는 다른 분포를 사용해 샘플링을 한다면 발생할 수 있는 엔트로피 차이이다.

다음 3가지 특성을 가지고 있다.

1. $KL(p||q) = H(p, q) - H(p)$

$$\begin{aligned} H(p, q) &= \sum_i p_i \log_2 \frac{1}{q_i} = - \sum_i p_i \log_2 q_i \\ &= - \sum_i p_i \log_2 q_i + \sum_i p_i \log_2 p_i - \sum_i p_i \log_2 p_i \\ &\quad - \sum_i p_i \log_2 p_i = H(p) \\ &= \sum_i p_i \log_2 p_i - \sum_i p_i \log_2 q_i + H(p) \\ &= \sum_i p_i \log_2 \frac{p_i}{q_i} + H(p) \\ &\quad \sum_i p_i \log_2 \frac{p_i}{q_i} = KL(p||q) \\ &= KL(p||q) + H(p) = H(p, q) \\ KL(p||q) &= H(p, q) - H(p) \end{aligned}$$

$$2. KL(p||q) \geq 0$$

$KL(p||q) \neq KL(q||p)$ 증명

$$KL(p||q) = \sum_i p_i \log_2 \frac{p_i}{q_i}$$

$$KL(q||p) = \sum_i q_i \log_2 \frac{q_i}{p_i}$$

로그함수의 비대칭성

$$\log \frac{a}{b} \neq \log \frac{b}{a}$$

$$\sum_i p_i \log_2 \frac{p_i}{q_i} \neq \sum_i q_i \log_2 \frac{q_i}{p_i}$$

$$KL(p||q) \neq KL(q||p)$$

하지만 $p=q$ 라면,

$$KL(p||q) = 0$$

$$KL(q||p) = 0$$

$$3. KL(p||q) \neq KL(q||p)$$

$$KL(P||Q) \geq 0 \quad \text{증명}$$

$$KL(P||Q) = - \sum_i p_i \log_2 \frac{q_i}{p_i}$$

$$\text{Convex 함수 } f(x) = -\log_2 x$$

$$\text{젠슨 부등식 } E[f(x)] \geq f(E[x])$$

$$E[f(x)] = - \sum_i p_i \log_2 \frac{q_i}{p_i}$$

$$f(E[x]) = -\log_2 \left(\sum_i p_i \frac{q_i}{p_i} \right)$$

$$= -\log_2 \left(\sum_i q_i \right)$$

$$= -\log_2(1) = 0$$

$$E[f(x)] \geq 0$$

$$KL(P||Q) \geq 0$$