

Homework 3

Yiling Hu 001537980

2 KL Divergence

1). Let $f(x) = \log\left(\frac{p(x)}{q(x)}\right)$, then we have:

$$\begin{aligned} D_{KL}(P, q) &= \sum_{x \in X} p(x) f(x) \\ &= E_{p(x)}[f(x)] \\ &= E_{p(x)}\left[\log\left(\frac{p(x)}{q(x)}\right)\right] \\ &= E_{p(x)}\left[-\log\left(\frac{q(x)}{p(x)}\right)\right] \end{aligned}$$

Now we let $z = \frac{q(x)}{p(x)}$, then we have:

$$f(z) = -\log(z)$$

Since $f(z) = -\log(z)$ is a convex function, we can apply Jensen's inequality, and then:

$$E[f(z)] \geq f(E[z])$$

$$E[f(z)] = E_{p(x)}\left[-\log\left(\frac{q(x)}{p(x)}\right)\right] = E_{p(x)}\left[\log\left(\frac{p(x)}{q(x)}\right)\right] = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

$$E[z] = E_{p(x)}\left[\frac{q(x)}{p(x)}\right] = \sum_x p(x) \cdot \frac{q(x)}{p(x)} = \sum_x q(x)$$

Since $q(x)$ is a probability distribution, $E[z] = 1$

$$\text{Then, } f(E[z]) = f(1) = -\log(1) = 0$$

Thus, we have $E[f(z)] \geq 0$, and we have proved that $E_{p(x)}[f(z)] = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right) = D_{KL}(p, q)$

Therefore $D_{KL}(p, q) \geq 0$.

$$2) \quad D_{KL}(p, q) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

① Let $p(x) = q(x)$, then we have

$$D_{KL}(p, q) = \sum_x p(x) \log\left(\frac{p(x)}{p(x)}\right) = \sum_x p(x) \cdot \log(1) = 0$$

② Let $D_{KL}(p, q) = 0$, as we proved in question 1).

$$\text{We have } D_{KL}(p, q) = E[f(z)] = 0 = f(E[z])$$

As we defined in question 1), $f(z) = -\log(z)$,

which is a convex function, then if $E[f(z)] = f(E[z])$,

We must have z as a constant:

$$\text{Let } \frac{q(x)}{p(x)} = c, \quad \sum_x q(x) = c \cdot \sum_x p(x)$$

Since $p(x)$ and $q(x)$ are distributions, $\sum_x p(x) = \sum_x q(x) = 1$

Thus, we have $1 = c \cdot 1$ and $c = 1$

Therefore $p(x) = q(x)$.

$$3). \quad D_{KL}(P_{data}, P_{\theta})$$

$$= \sum_x P_{data}(x) \cdot \log\left(\frac{P_{data}(x)}{P_{\theta}(x)}\right)$$

$$= \sum_x P_{data}(x) \cdot \log(P_{data}(x)) - \sum_x P_{data}(x) \cdot \log(P_{\theta}(x))$$

Let $f(x) = \log p_\theta(x)$, we have:

$$\sum_x P_{\text{data}}(x) \cdot f(x) = E_{P_{\text{data}}(x)}[f(x)]$$

Since $P_{\text{data}}(x)$ is a uniform distribution, we have

$$E_{P_{\text{data}}(x)}[f(x)] = \sum_{i=1}^n P_{\text{data}}(x) f(x)$$

$$= \sum_{i=1}^n \frac{1}{n} \cdot f(x)$$

$$= \frac{1}{n} \cdot \sum_{i=1}^n \log p_\theta(x^{(i)})$$

$$\begin{aligned} \text{Since } D_{KL}(P_{\text{data}}, P_\theta) &= \sum_x P_{\text{data}}(x) \cdot \log(P_{\text{data}}(x)) - \sum_x P_{\text{data}}(x) \cdot \log(P_\theta(x)) \\ &= C \end{aligned}$$

Since $\sum_x P_{\text{data}}(x) \cdot \log(P_{\text{data}}(x))$ has no θ , it can be seen as a constant, and thus we have:

$$D_{KL}(P_{\text{data}}, P_\theta) = C - \frac{1}{n} \cdot \sum_{i=1}^n \log(p_\theta(x^{(i)}))$$

$\arg \min(C - \frac{1}{n} \cdot \sum_{i=1}^n \log(p_\theta(x^{(i)})))$ can be seen as

$$\arg \min(-\sum_{i=1}^n \log(p_\theta(x^{(i)}))) = \arg \max(\sum_{i=1}^n \log(p_\theta(x^{(i)})))$$