⟨KL-Divergence⟩

정의 : 두 확률 분포의 차이를 계산할 때 사용되는 함수.

어떤 이상적인 분포에 대해, 그 분포를 근사하는 다른 분포를 사용해 샘플링을 할 때 발생할 수 있는 정보 엔트로피 차이를 계산한다.

* 정보 엔트로피 차이는 Cross-entropy를 이용해 구할 수 있다.

먼저 Cross-entropy를 수식으로 표현하면 아래와 같다.

$$H(p,q) = \sum_i p_i \log_2 \frac{1}{q_i} = -\sum_i p_i \log_2 q_i$$

여기서 p_i는 target 값 (확률에 대한 참값), q_i는 우리가 학습한 모델에서 도출된 결과값이다. 즉 우리가 어떤 q_i를 학습하고 있고 이 방향이 올바른 방향으로 진행된다면, Cross-entropy는 작아지는 방향으로 진행될 것이다.

위 Cross-entropy를 이용해 KL-Divergence를 구하면 아래와 같다.

$$\begin{aligned}
H(p,q) &= -\sum_i p_i \log(q_i) \\
&= -\sum_i p_i \log(q_i) \underbrace{- \sum_i p_i \log(p_i)}_{= H(p)} + \sum_i p_i \log(p_i) \\
&= H(p) + \sum_i p_i \log(p_i) - \sum_i p_i \log(q_i) \\
&= H(p) + \sum_i p_i \log\left(\frac{p_i}{q_i}\right)
\end{aligned}$$

이만큼 더해진 것이 바로 P, Q의 정보량 차이, 즉 KL-divergence가 된다.

P의 엔트로피에 이만큼 더해진 것이 Cross-entropy

KL-Divergence는 KL(p||q) 로 쓰고, 위 식에서 유도된 KL(p||q)는 아래와 같다

$$KL\left(p\|q\right)=H\left(p,q\right)-H\left(p\right)$$

또한 KL-divergence의 정확한 식은 아래와 같다.

$$KL\left(p\|q\right)=\begin{cases}\sum_i p_i\log\frac{p_i}{q_i}\quad\text{or}\quad\sum_i p_i\log\frac{q_i}{p_i} & \text{(discrete)}\\ \int p\left(x\right)\log\frac{p\left(x\right)}{q\left(x\right)}dx\quad\text{or}\quad-\int p\left(x\right)\log\frac{q\left(x\right)}{p\left(x\right)}dx & \text{(continuous)}\end{cases}$$

KL(p||q)의 특성중 가장 대표적인 2개는 다음과 같다.

① $KL(p\|q) > 0$

② $KL(p\|q) \neq KL(q\|p)$.

examples)

1 P와 Q 2개의 확률 분포가 다음과 같이 나와있을 때, KL(P||Q) 와 KL(Q||P)를 구하는 과정은 아래와 같다.

x	0	1	2
Distribution $P(x)$	$9/25$	$12/25$	$4/25$
Distribution $Q(x)$	$1/3$	$1/3$	$1/3$

1) $KL(P\|Q) = \sum P(x)\cdot \ln\left(\frac{P(x)}{Q(x)}\right) = 9/25 \cdot \ln \frac{9/25}{1/3} + 12/25 \cdot \ln \frac{12/25}{1/3} + 4/25 \cdot \ln \frac{4/25}{1/3}.$

2) $KL(Q\|P) = \sum Q(x)\cdot \ln\left(\frac{Q(x)}{P(x)}\right) = 1/3 \cdot \ln \frac{1/3}{9/25} + 1/3 \cdot \ln \frac{1/3}{12/25} + 1/3 \cdot \ln \frac{1/3}{4/25}$

2 자주 사용되는 gaussian distribution에 대한 KL-divergence는 아래와 같다.

$$KL(p\|q) = \int p(x)\cdot \ln \frac{p(x)}{q(x)} dx.$$

$$p(x) = N(\mu_p, \Sigma_p) = (2\pi)^{-k/2}\cdot|\Sigma_p|^{-1/2}\cdot e^{\left(-\frac{1}{2}(x-\mu_p)^T\Sigma_p^{-1}(x-\mu_p)\right)}$$

$$q(x) = N(\mu_q, \Sigma_q) = (2\pi)^{-k/2}\cdot|\Sigma_q|^{-1/2}\cdot e^{\left(-\frac{1}{2}(x-\mu_q)^T\Sigma_q^{-1}(x-\mu_q)\right)}$$

p와 q 모두 k-dimension이라 가정하면

$$KL(p\|q) = E_p[\log(p) - \log(q)] = E_p\left[\frac{1}{2}\log\frac{|\Sigma_q|}{|\Sigma_p|} - \frac{1}{2}(x-\mu_p)^T\Sigma_p^{-1}(x-\mu_p) + \frac{1}{2}(x-\mu_q)^T\Sigma_q^{-1}(x-\mu_q)\right]$$

$$= \frac{1}{2}\log\frac{|\Sigma_q|}{|\Sigma_p|} - \frac{1}{2}E\left[(x-\mu_p)^T\Sigma_p^{-1}(x-\mu_p)\right] + \frac{1}{2}E\left[(x-\mu_q)^T\Sigma_q^{-1}(x-\mu_q)\right]$$

먼저, $(x-\mu_p)^T\Sigma_p^{-1}(x-\mu_p)$ 는 trace 연산을 통해 구할 수 있다. 즉

$(x-\mu_p)^T\Sigma_p^{-1}(x-\mu_p) = tr\{(x-\mu_p)(x-\mu_p)^T\Sigma_p^{-1}\}$ 이다.

$$\therefore E\left[(x-\mu_p)^T\Sigma_p^{-1}(x-\mu_p)\right] = E\left[tr\{(x-\mu_p)(x-\mu_p)^T\Sigma_p^{-1}\}\right]$$

$$= tr\{E[(x-\mu_p)(x-\mu_p)^T]\cdot\Sigma_p^{-1}\}$$

이 때 $E[(x-\mu_p)(x-\mu_p)^T] = \Sigma_p$ 이므로

$tr\{E[(x-\mu_p)(x-\mu_p)^T]\cdot\Sigma_p^{-1}\} = tr\{\Sigma_p\Sigma_p^{-1}\} = tr\{I_k\} = k$ 가 된다.

$E[(x-\mu_q)^T\Sigma_q^{-1}(x-\mu_q)] = (\mu_p-\mu_q)^T\Sigma_q^{-1}(\mu_p-\mu_q) + Tr\{\Sigma_q^{-1}\Sigma_p\}$ 이다.[1]

위 식을 정리하면 $KL(p\|q) = \frac{1}{2}\left(\log\frac{|\Sigma_q|}{|\Sigma_p|} - k + (\mu_p-\mu_q)^T\Sigma_q^{-1}(\mu_p-\mu_q) + tr\{\Sigma_q^{-1}\Sigma_p\}\right)$

가 나오고,

만약 $q \sim N(0,I)$ 라면

$KL(p\|q) = \frac{1}{2}[\mu_p^T\mu_p + tr\{\Sigma_p\} - k - \log|\Sigma_p|]$ 가 된다.

[1] Matrix Cookbook section 8.2, eq 380.

8.2.2 Mean and variance of square forms

Mean and variance of square forms: Assume $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$

$$E(\mathbf{x}\mathbf{x}^T) = \mathbf{\Sigma} + \mathbf{m}\mathbf{m}^T \quad (377)$$

$$E[\mathbf{x}^T\mathbf{A}\mathbf{x}] = \mathrm{Tr}(\mathbf{A}\mathbf{\Sigma}) + \mathbf{m}^T\mathbf{A}\mathbf{m} \quad (378)$$

$$\mathrm{Var}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \mathrm{Tr}[\mathbf{A}\mathbf{\Sigma}(\mathbf{A}+\mathbf{A}^T)\mathbf{\Sigma}] + ... + \mathbf{m}^T(\mathbf{A}+\mathbf{A}^T)\mathbf{\Sigma}(\mathbf{A}+\mathbf{A}^T)\mathbf{m} \quad (379)$$

$$E[(\mathbf{x}-\mathbf{m}')^T\mathbf{A}(\mathbf{x}-\mathbf{m}')] = (\mathbf{m}-\mathbf{m}')^T\mathbf{A}(\mathbf{m}-\mathbf{m}') + \mathrm{Tr}(\mathbf{A}\mathbf{\Sigma}) \quad (380)$$