7. Information and Entropy

Information

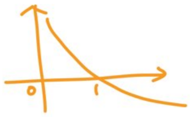
불확실성(Uncertainty)을 나타내는 값으로 random variable X에 대한 불확실성이다.

확률이 낮을수록 불확실성이 높다. 왜냐하면 P(x)는 [0, 1]인데 -log는 0에 가까울수록 무한, 1에 가까울수록 0이기 때문이다.

늘 나타내는 값

$$I(\mathrm{x}) = -\log P(\mathrm{x})$$

random variable X에 대한 불확실성



$0 \le P(x) \le 1$

확률이 낮을수록 (불확실할수록) 높음

Entropy

정보량의 기대값(평균)으로 분포의 평균적인 uncertainty를 나타내는 값이다.

이 Entropy를 통해 분포의 형태를 예측해 볼 수 있는데 entropy가 작을수록 평평하고, 클수록 한 군데에 집중된 sharp한 형태이다.

$$H(P) = -\mathbb{E}_{x \sim P(x)}[\log P(x)]$$

Cross Entropy

분포 P의 관점에서 본 분포 Q의 정보량의 평균이다.

⇒아래 식은 P(X)에서 sampling 한 x를 Q(x) 분포에 넣어주고 평균을 구하는 것이다.

여기서 P를 참인 확률분포, Q를 신경망 확률분포라 생각하면 적용하기 쉽다.

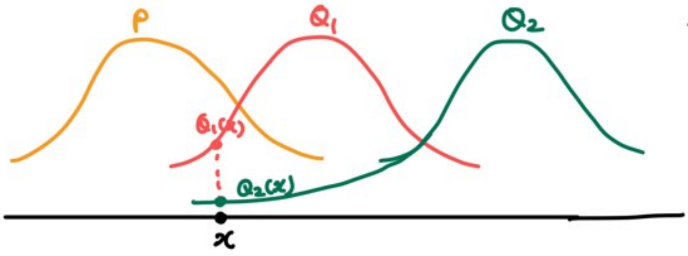
참 ↙ ↙ P_θ : neural network

$$H(P,Q) = -\underbrace{\mathbb{E}_{x\sim P(x)}[\log Q(x)]}_{\text{P(x)에서 sampling한 x를 Q(x)분포에 넣어줌}} = -\int p(x)\log Q(x)\,dx$$

monte carlo

$$\approx -\frac{1}{n}\sum_{i=1}^{n}\log P(x_i)$$

두 분포가 비슷할수록 작은 값을 가진다.



$$\Rightarrow Q_1(x) > Q_2(x)$$

$$\log Q_1(x) > \log Q_2(x)$$

$$-\log Q_1(x) < -\log Q_2(x)$$

P와 더 먼(안 비슷한) Q_2 값이 더 큼.

DNN Optimization using Cross Entropy

$$\mathcal{L}(\theta) = -\mathbb{E}_{x\sim P(x)}\left[\mathbb{E}_{y\sim P(y|x)}[\log P(y|x;\theta)]\right]$$

monte carlo ↓

$$\approx -\frac{1}{N\cdot k}\sum_{i=1}^{N}\sum_{j=1}^{k}\log P\left(y_{i,j}|x_i;\theta\right)$$

k=1 가정 ↓

$$\approx -\frac{1}{N}\sum_{i=1}^{N}\log P(y_i|x_i;\theta)\text{, if } k=1.$$

↓ 정답값이 one-hot encoding 이므로 softmax layer 에서라서

$$= -\frac{1}{N}\sum_{i=1}^{N} y_i^T \cdot \log \hat{y}_i$$

KL-Divergence and Cross Entropy

KL-Divergence와 Cross Entropy를 seta로 미분하면 같다.

$$
\begin{aligned}
\mathrm{KL}(p||p_\theta) &= -\mathbb{E}_{x\sim p(\mathrm{x})}\left[\log\frac{p_\theta(\mathrm{x})}{p(\mathrm{x})}\right] \\
&= -\int p(x)\log\frac{p_\theta(x)}{p(x)}\,dx \\
&= \underbrace{-\int p(x)\log p_\theta(x)\,dx}_{\text{cross entropy}} + \underbrace{\int p(x)\log p(x)\,dx}_{\text{entropy}} \\
&= H(p,p_\theta) - H(p)
\end{aligned}
$$

$$
\nabla_\theta \mathrm{KL}(p||p_\theta) = \nabla_\theta H(p,p_\theta) - \cancelto{0}{\nabla_\theta H}(p)
$$