Mutual Information, Cross entropy,
KL Divergence

수업 목표

이번 수업의 핵심:

- Mutual information의 정의 및 이해
- Entropy, Joint entropy, Conditional entropy, Mutual information의 관계
- Cross-entropy와 KL divergence의 개념

핵심 개념

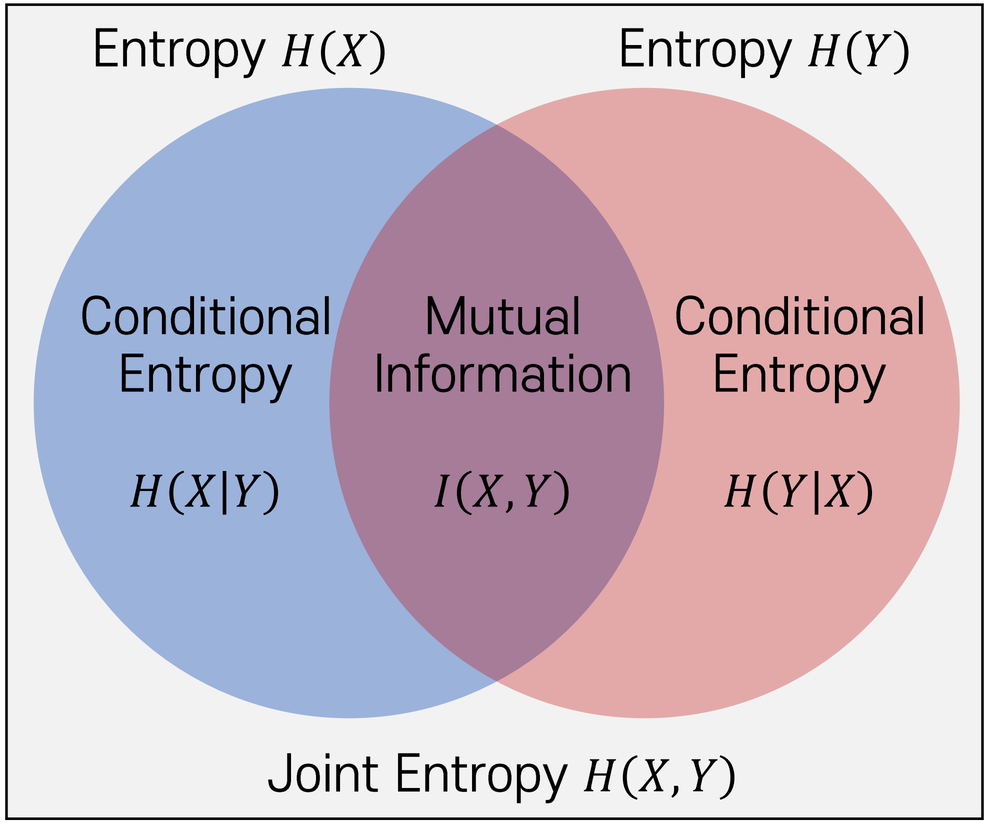
- Mutual information
- Cross entropy, KL divergence

Mutual Information

Mutual Information과 Joint Entropy, Conditional Entropy의 관계

$$I(X,Y) \coloneqq H(X,Y) - H(X|Y) - H(Y|X)$$
$$= H(X) - H(X|Y)$$
$$= H(Y) - H(Y|X)$$
$$= H(X) + H(Y) - H(X,Y)$$

$$= \mathbb{E}_{x,y\sim p(X,Y)}\left[\log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right]$$



Mutual Information

Mutual Information:

1. X와 Y에 대해서 공통으로 얻을 수 있는 정보량
   - 내일 비가 오는지를 X, 모레 비가 오는지를 Y라고 한다면,
     오늘 비가 왔는지에 대한 정보는 X, Y 모두를 알아내는데 도움이 되는 정보임
   - 이런 유형으로 X, Y가 연관되어 있는 정보들을 모두 모으면 Mutual information

- X, Y에 대한 정보를 모두 모은 후에 X에서 Y와 관계된 정보가 빠진 정보량과
  Y에서 X와 관계된 정보가 빠진 정보량을 빼는 것으로 구할 수 있음

$$I(X, Y) \coloneqq H(X, Y) - H(X|Y) - H(Y|X)$$

- 기호 I는 Mutual information를 의미

Mutual Information

Mutual Information:

2. X를 알려줬을 때 Y에 대해서 얻은 정보량 (또는 반대)

- X, Y가 연관되어 있는 정보들을 모두 모으는 또 다른 방법은 Y의 총 정보량에서 X를 알려줬을 때 Y에 남은 정보를 빼는 것
  - 또는 반대로 X의 총 정보량에서 Y를 알려줬을 때 X에 남은 정보를 빼는 것

- 즉, Mutual information을 다음과 같이 구할 수도 있음:

$$I(X, Y) = H(Y) - H(Y|X)$$

$$= H(X) - H(X|Y)$$

Cross-Entropy

Cross-Entropy:

$$\mathrm{CE}(P,Q) = -\mathbb{E}_{x\sim p(x)}[\log q(x)]$$

- P, Q는 각각 $p(x), q(x)$를 따르는 확률 변수

- Entropy의 두번째 특성에서 언급한 $-\mathbb{E}_{x\sim p(x)}[\log q(x)]$를 활용
  - $p(x)$에서 뽑히는 x를 가지고 $q(x)$로 놀라고 있는 것
  - $-\mathbb{E}_{x\sim p(x)}[\log q(x)]$의 하한은 $H(P)$이며 등호는 $p(x) = q(x)$일 때
  - $q(x)$를 모델링할 때, $\mathrm{CE}(P,Q)$를 최소화하면 $q(x)$를 $p(x)$와 최대한 같도록 모델링할 수 있음
  - 이를 활용하여 많은 기계학습 방법론이 Cross-entropy loss를 사용

Kullback-Leibler Divergence

KL Divergence: 한 분포에서 다른 분포까지 떨어진 확률 분포 간의 거리

- 대칭적이지 않기 때문에 엄밀하게 거리는 아님! 대신 다른 좋은 성질들이 존재함

$$D_{KL}(P \parallel Q) := \underline{\mathbb{E}_{x\sim p(x)}\left[\log \frac{p(x)}{q(x)}\right]}$$

Non-symmetric! i.e., $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$

- Cross-entropy와의 관계

$$D_{KL}(P \parallel Q) = \mathrm{CE}(P, Q) - H(P)$$

즉, $q(x)$를 모델링 할 때 Cross-entropy를 최소화하는 것은
KL divergence를 최소화하는 것과 동치

$$\underset{q(x)}{\operatorname{argmin}}\, \mathrm{CE}(P, Q) = \underset{q(x)}{\operatorname{argmin}}\big(D_{KL}(P \parallel Q) + H(P)\big) = \underset{q(x)}{\operatorname{argmin}}\, D_{KL}(P \parallel Q)$$

KL Divergence의 성질

1. $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$

- KL divergence는 교환법칙이 성립하지 않음

$$D_{KL}(P \parallel Q) = \mathbb{E}_{x\sim p(x)}\left[\log \frac{p(x)}{q(x)}\right]$$

$$D_{KL}(Q \parallel P) = \mathbb{E}_{x\sim q(x)}\left[\log \frac{q(x)}{p(x)}\right]$$

- $x\sim p(x)$인지 $x\sim q(x)$의 개념이 $q(x)$를 모델링 하는데 있어서 중요

KL Divergence의 성질

2. 모든 P, Q에 대하여 $D_{KL}(P \parallel Q) \geq 0$이며, 등호성립 조건은 $P = Q$

- KL divergence의 가장 유용한 성질
  - 항상 음이 아닌 실수이기 때문에 KL divergence는 두 분포 간의 거리의 개념으로 쓰임
  - 등호성립조건이 $P = Q$이기 때문에, $D_{KL}(P \parallel Q) > 0$이라면 $P \neq Q$라는 성질도 사용할 수 있음

증명)

1. Entropy의 특성 2: $H(P)$는 $-\mathbb{E}_{x\sim p(x)}[\log q(x)]$ (Cross-entropy)의 하한이다
2. KL divergence와 Cross-entropy와의 관계: $D_{KL}(P \parallel Q) = \mathrm{CE}(P, Q) - H(P)$

$$\therefore D_{KL}(P \parallel Q) = \mathrm{CE}(P, Q) - H(P) = -\mathbb{E}_{x\sim p(x)}[\log q(x)] - H(P) \geq 0$$

KL Divergence의 성질

3. $p(x) > 0$이고 $q(x) = 0$인 지점이 존재하면 $D_{KL}(P \parallel Q) = \infty$

- $q(x) = 0$인 지점에서 x가 뽑히면 무한히 놀라움
- 즉, $D_{KL}(P \parallel Q)$을 줄일 때 $p(x) > 0$이고 $q(x) = 0$인 지점부터 없애려고 함

증명) $p(x_i) > 0$이고 $q(x_i) = 0$

- $D_{KL}(P \parallel Q) = \mathbb{E}_{x \sim p(x)}\left[\log \frac{p(x)}{q(x)}\right] \geq p_X(x_i) \log \frac{p_X(x_i)}{\cancel{q_X(x_i)}} \overset{\infty}{=} \underbrace{p_X(x_i)}_{\neq 0} \log \infty = \infty$

요약

- 여러 확률 변수에 대해 공통적으로 얻을 수 있는 Mutual Information의 계산 방법
- Cross-entropy와 KL divergence의 개념과 성질
- Cross-entropy와 KL divergence의 관계 및 이를 목적함수로 사용하는 최적화 문제의 사례