정보와 Entropy

수업 목표

이번 수업의 핵심:

- 정보의 개념과 확률적 사건에 대한 정보량 정의
- Entropy의 개념 및 관련한 세 가지 특성의 증명 및 이해
- 다양한 확률 변수의 Entropy 계산

핵심 개념

- 정보, 정보량, bit
- Entropy

정보

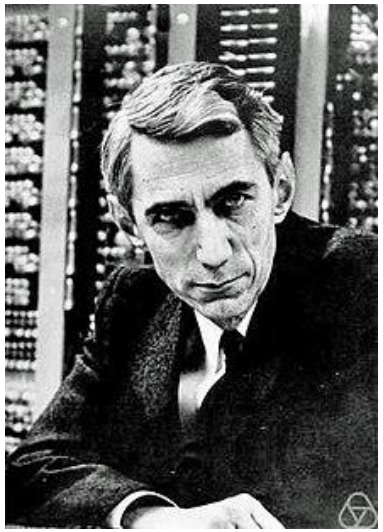
정보: 놀라움의 정도

- 놀라움이 크다 ⇒ 받은 정보가 많다, 놀라움이 작다 ⇒ 받은 정보가 적다

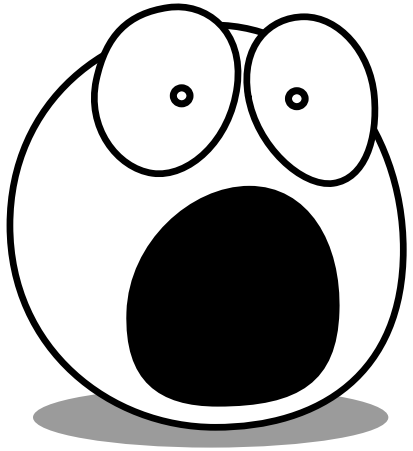
놀라움의 정도: 발생 확률이 적은 일이 일어날수록 놀라움이 큼

- e.g., 한국이 포르투갈을 이겨서 16강에 진출하는 사건

→ 발생 확률이 적어 놀라움이 큼



Claude Shannon
1916~2001



$$\text{정보} \propto \frac{1}{\text{발생 확률}}$$

정보

정보의 단위: bit

- 흔한 일이면 적은 수의 bit로 표현이 가능
- 발생 확률이 낮은 놀라운 일이면, 이를 설명하기 위해 많은 수의 bit가 필요

→ 정보를 다음과 같이 정의해보자

확률변수 X가 x값이라는 정보를 알았을 때 얻은 정보량

$$\text{Information}(X = x) := \log_2 \frac{1}{P(X = x)} = -\log_2 P(X = x)$$

왜 $\log_2$가 쓰였을까?

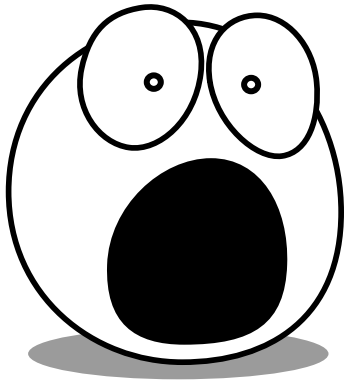
정보

$\log_2$의 배경

- 길이 N의 0, 1로 이루어진 이진 문자열 로또를 Uniform하게 뽑는다고 가정
  - 정보량은 총 N bit
- N bit 문자열 중 내가 선택한 문자열이 나올 확률 $= \frac{1}{2^N}$

놀라움의 정도
2^N **중에 내 것이 나오다니!**

  - 로또 당첨 번호를 누가 알려준다면 얼마 만큼의 bit가 필요한가? N **bit!**
- **즉, 발생 확률이 $\frac{1}{2^N}$인 사건에 대해서 필요한 정보량 혹은 얻은 정보량은 N bit**
  - 정보량의 정의 식에 대입해보자: $N \text{ bits} = -\log_2 \frac{1}{2^N}$



로또 번호	1	0	1	0	1	1	1
내 번호	1	0	1	0	1	1	1

$\frac{1}{2^7}$ **의 확률**

Entropy

Entropy: 확률 변수 X에 대해서 X를 알아냄으로써 얻을 수 있는 정보의 기대값

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

- 이산 확률 변수

$$H(X) = -\sum_i p_X(x_i) \log p_X(x_i)$$

- 연속 확률 변수

$$H(X) = -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx$$

Entropy의 특성

1. Entropy는 항상 음이 아닌 실수의 값을 가진다

$$H(X) \geq 0$$

- 정해져 있지 않은 확률적인 변수 X의 결과를 알려준다면 항상 정보를 얻게 됨
  - 등호는 X가 정해져 있을 때 (즉, Deterministic 할 때)

  예시 1) 동전을 던져서 **앞면이 나온 사건**은 앞면이 나왔다는 정보를 제공

  예시 2) 항상 앞면이 나오는 동전에서 **앞면이 나온 사건**은 아무런 정보를 주지 않음

증명)

$$\forall x_i, \quad p_X(x_i) \in [0,1]$$

$$\Rightarrow -p_X(x_i)\log p_X(x_i) \geq 0$$

$$\Rightarrow -\sum_i p_X(x_i)\log p_X(x_i) \geq 0$$

Entropy의 특성

2. Entropy $H(X)$는 $\mathbb{E}_{x\sim p(x)}[-\log q(x)]$의 하한이다

$$H(X) = \mathbb{E}_{x\sim p(x)}[-\log p(x)] \le \mathbb{E}_{x\sim p(x)}[-\log q(x)]$$

- 이 때 X는 $p(x)$를 따름

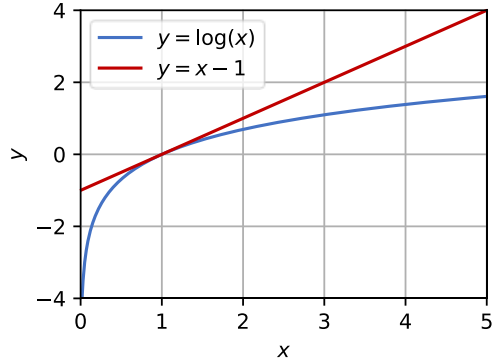
증명) $\mathbb{E}_{x\sim p(x)}[-\log p(x)] - \mathbb{E}_{x\sim p(x)}[-\log q(x)]$

$$= \left(-\sum_i p_X(x_i)\log p_X(x_i)\right) - \left(-\sum_i p_X(x_i)\log q_X(x_i)\right)$$

$$= \sum_i p_X(x_i)\log\frac{q_X(x_i)}{p_X(x_i)} \le \frac{1}{\ln 2}\sum_i p_X(x_i)\left(\frac{q_X(x_i)}{p_X(x_i)} - 1\right)$$

$$= \frac{1}{\ln 2}\sum_i \left(q_X(x_i) - p_X(x_i)\right) = \frac{1}{\ln 2}\left(\sum_i q_X(x_i) - \sum_i p_X(x_i)\right) = 0$$

- 등호는 $\frac{q_X(x_i)}{p_X(x_i)} = 1$일 때, 즉 $p_X(x_i) = q_X(x_i)$



$$\log x = \frac{\ln x}{\ln 2} \le \frac{1}{\ln 2}\cdot(x-1)$$

- 등호는 $x = 1$일 때

Entropy의 특성

2. Entropy $H(X)$는 $\mathbb{E}_{x\sim p(x)}[-\log q(x)]$의 하한이다

- 기계학습에서 분포를 학습할 때 $\mathbb{E}_{x\sim p(x)}[-\log q(x)]$를 최소화 하는 것을 목표로 함

해석) $p(x)$를 $q(x)$로 모델링한다고 해보자

- $\mathbb{E}_{x\sim p(x)}[-\log q(x)]$: $p(x)$에서 뽑히는 x를 가지고 $q(x)$로 놀라고 있는 것
  - 잘못된 확률 분포로 생각하고 있다면 더욱 많이 놀랄 것
  - e.g., $x=0$에 대한 확률을 낮게 생각하고 있었는데 자꾸 $x=0$이 나오는 경우,
    $p(x)$가 낮지 않은데 $-\log q(x)$가 크기 때문에 전체 평균이 커짐

$$H(X) = \mathbb{E}_{x\sim p(x)}[-\log p(x)] \leq \mathbb{E}_{x\sim p(x)}[-\log q(x)]$$

등호는 $p(x)=q(x)$ 일 때,
확률 분포를 완벽하게 모델링하면 평균적으로 가장 적게 놀람

Entropy의 특성

3. Entropy는 Uniform distribution일 때 최대가 된다

$$H(X) \leq \log N$$

- N: X가 가질 수 있는 값의 개수

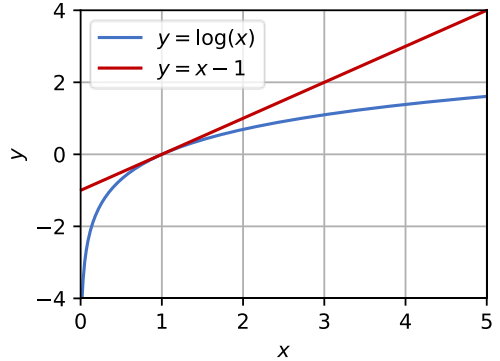
증명) $\mathbb{E}_{x\sim p(x)}[-\log p(x)] - \log N$

$$= \left(-\sum_{i=1}^{N} p_X(x_i)\log p_X(x_i)\right) - \left(\sum_{i=1}^{N} p_X(x_i)\log N\right)$$

$$= \sum_{i=1}^{N} p_X(x_i)\log\frac{1}{Np_X(x_i)} \leq \frac{1}{\ln 2}\sum_{i=1}^{N} p_X(x_i)\left(\frac{1}{Np_X(x_i)} - 1\right)$$

$$= \frac{1}{\ln 2}\sum_{i=1}^{N}\left(\frac{1}{N} - p_X(x_i)\right) = 0$$

- 등호는 $\frac{1}{Np_X(x_i)} = 1$일 때, 즉 Uniform distribution



$$\log x = \frac{\ln x}{\ln 2} \leq \frac{1}{\ln 2}\cdot(x-1)$$

- 등호는 $x = 1$일 때

Entropy의 특성

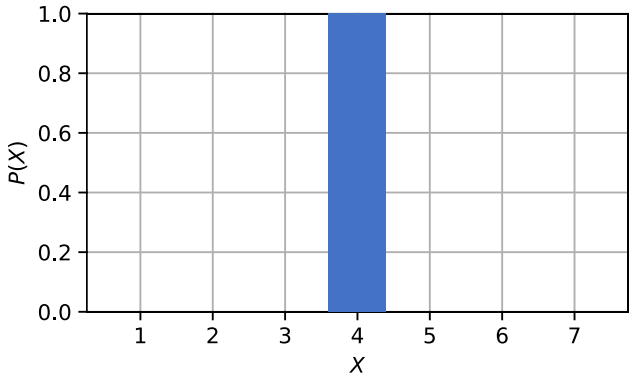
3. Entropy는 Uniform distribution일 때 최대가 된다

$$H(X) \leq \log N$$

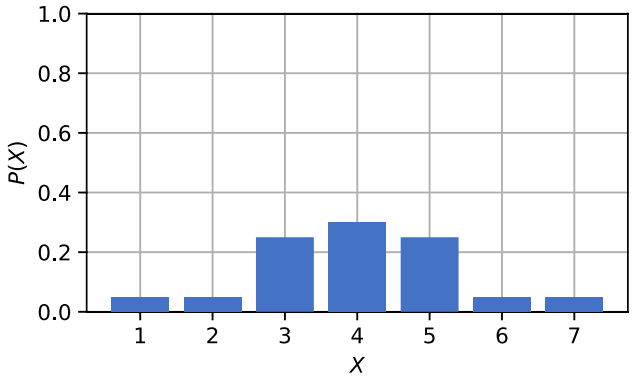
해석) 가장 정보를 많이 필요로 하는 분포는 Uniform distribution이다

- 모든 정의역의 어떤 값도 동일한 확률로 나오기에 무엇이 나올지 알 수 없음
- 가장 적은 정보가 필요한 분포는 가능한 값이 오로지 하나인 Distribution
  - $H(X) \geq 0$ 참고

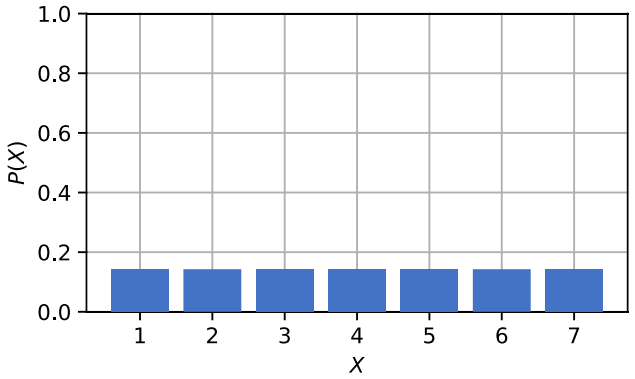
이건 다 알지! 무조건 4!



3에서 5 정도에서 나오겠네



전혀 모르겠어...

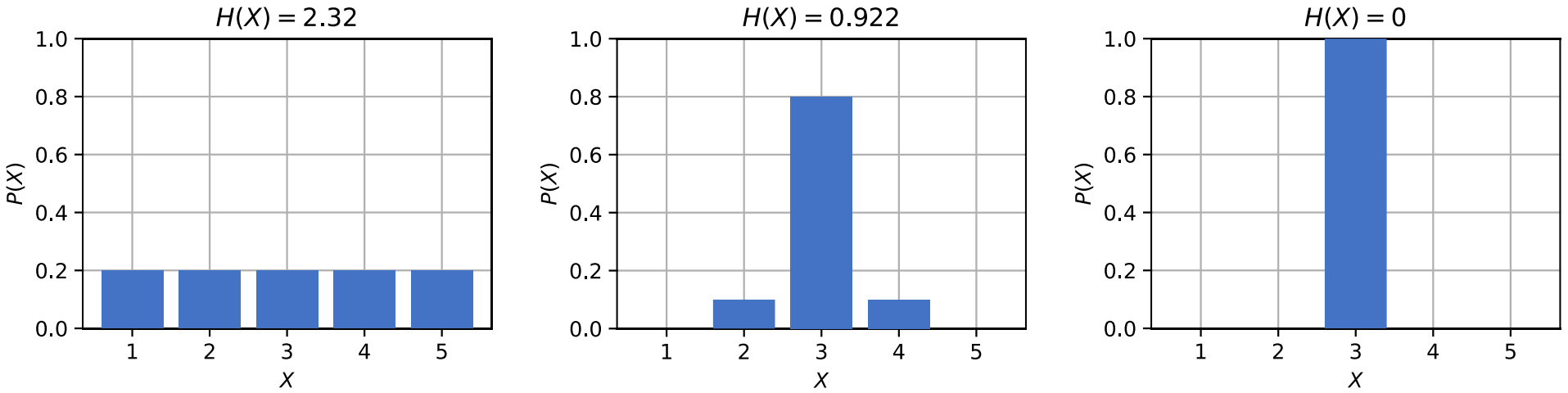


다양한 확률 변수의 Entropy

이산 확률 분포의 Entropy

- [0.2, 0.2, 0.2, 0.2, 0.2]: $\log 5 = 2.32$
- [0.0, 0.1, 0.8, 0.1, 0.0]: $0.2 \log 10 + 0.8 \log 1.25 = 0.922$
- [0.0, 0.0, 1.0, 0.0, 0.0]: $1 \log 1 = 0$

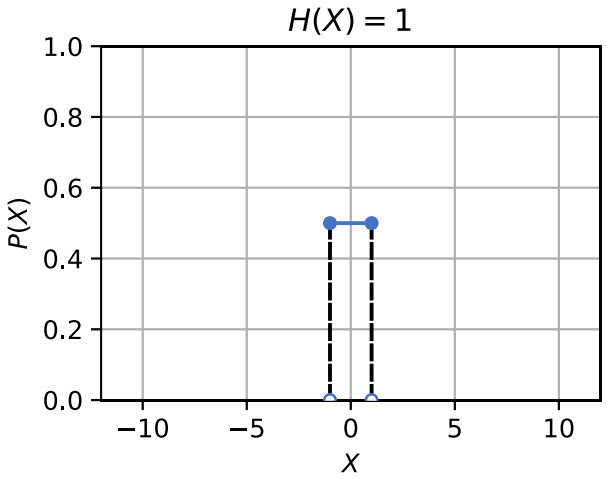
Deterministic 할수록
Entropy 감소

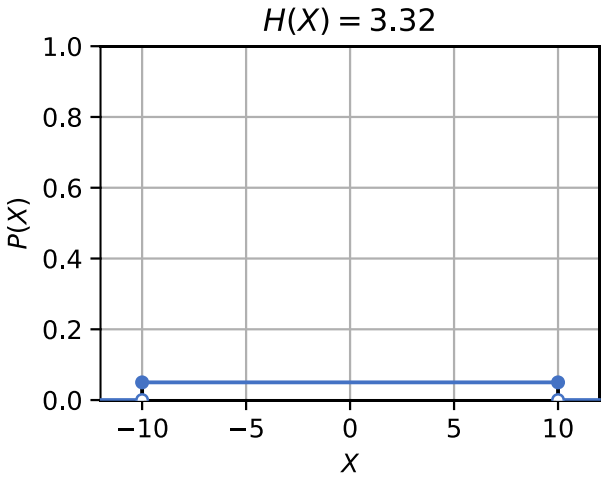


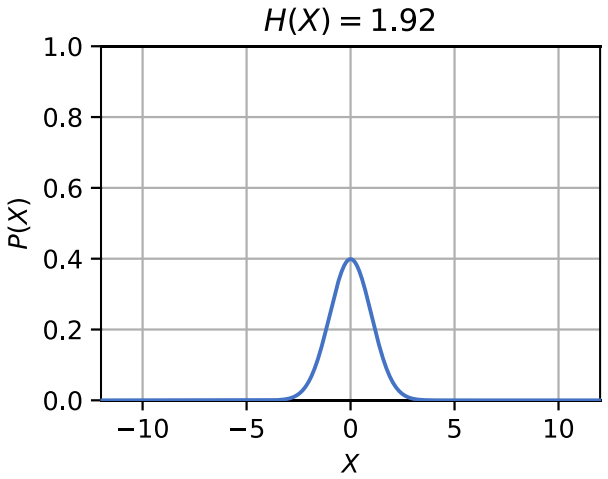
다양한 확률 변수의 Entropy

연속 확률 분포의 Entropy

- Uniform $U(-1,1)$: $\int_{-1}^{1} \frac{1}{2} \log 2 \, dx = 1$
- Uniform $U(-10,10)$: $\int_{-10}^{10} \frac{1}{10} \log 10 \, dx = 3.32$
- Gaussian $\mathcal{N}(0,1)$: $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \log \sqrt{2\pi} e^{\frac{1}{2}x^2} \, dx = 1.92$







요약

- 확률적 사건에 대한 정보량의 개념
- 확률 변수에 대해서 얻을 수 있는 정보량의 기대값 계산 및 Entropy의 정의
- Entropy의 여러가지 특성과 성질
- 이산/연속 확률 변수의 Entropy 계산 및 Deterministic과 Entropy의 관계