Probability & Statistics

정보 이론

강사 양석환

정보 이론 개요

정보이론 개요

- **정보 이론 (Information Theory)이란?**
  - **디지털 정보의 정량화, 저장, 그리고 의사소통을 연구하는 과학적 연구**
  - **1920년대 해리 나이퀴스트(Harry Nyquist)와 랄프 하틀리(Ralph Hartley)에 의해 처음 정립됨**
  - **1940년대에 클로드 섀넌(Claude Shannon)이 더욱 구체화됨**
    - 1948년 섀넌은 Bell Systems Technical Journal에 "A Mathematical Theory of Communication"이라는 논문을 제출하면서, 정보라는 것이 어떻게 하면 절대적인 정확도로, 그리고 고유하게 정량화 될 수 있는지 보여줌
    - 논문에 따르면 본질적으로 모든 종류의 의사소통(Communication), 예컨대 신호, 텍스트, 라디오 신호, 그리고 사진까지 모두 비트(bits)로 변환될 수 있음
    - 이 논문은 디지털 시대의 청사진을 제공했다고 평가됨

정보이론 개요

- 정보의 불확정성
  - 일반적으로 정보는 불확정성의 해상도(the resolution of uncertainty)라고 표현됨
  - 불확정성이 큰 것일수록 유의미한 정보가 담겨있고, 불확정성이 작으면 유의미한 정보가 적다고 이해할 수 있음
  - 1948년 클로드 섀넌에 의해 이러한 정보의 추상적인 개념이 형식화됨
    - 정보는 가능한 메세지의 집합으로 생각되며,
    - 이때 정보 이론의 목표는 이 메세지를 노이즈가 있는 채널을 통해 전달한 뒤 메시지를 수신하는 측에서 오차확률을 최소화하는 방법을 연구하는 것

정보이론 개요

- 섀넌의 '노이즈 채널 코딩 정리'
  - 노이즈가 있는 어떠한 채널이든 간에 각 채널에는 고유한 채널용량(Channel capacity)이 존재해서 채널용량의 한계를 벗어나지 않는 비율(rate, 단위는 일반적으로 bits/s)에 한해서는 정보를 오차 없이 얼마든지 보낼 수 있다
  - 이러한 채널 용량을 정의한 것이 섀넌의 업적 중에서도 가장 중요한 것 중 하나
  - 채널용량이 존재한다고 예언함으로써 정보이론에서 무엇을 연구해야 하며 무엇이 달성될 수 있는지 예언함
    → 이로 인하여 여러 기관에서 정보 이론 연구를 지원하기 시작했고 이때 투자했던 것들이 현대의 많은 기술들을 낳게 됨

정보 이론의 응용

- 데이터 압축
  - 일상에서 흔히 사용하는 ZIP파일은 정보이론의 소스코딩/데이터 압축의 대표적인 응용 중 하나
  - DSL은 정보이론의 채널 코딩/오류 탐지 및 수정의 응용 중 하나
  - 이러한 응용들은 보이저호 탐사, CD의 발명, 스마트폰과 현대 인터넷의 발전 등에도 큰 기여를 함
- 언어학
  - 정보이론을 활용하여 소음이 많은 공간에서도 대화가 가능한 이유 등을 설명
  - 음향적 손실이 많은데도 대화가 가능한 것은 언어를 통한 의사소통이 정보량이 많은 음절/어절을 포착하는 과정이기 때문으로 설명될 수 있음
  - 화자와 청자 또한 이러한 사실에 민감하기 때문에 주변에 소음이 많을 때, 엔트로피가 높은 단위를 말할 때 과잉 조음하는 등의 전략을 취함

엔트로피

엔트로피 개요

- 정보 이론은
  - 확률론, 통계학, 컴퓨터과학, 통계역학 그리고 전자공학의 교차점에 위치함
  - 정보 이론의 하위분야로는 소스코딩, 알고리즘 복잡도 이론, 알고리즘 정보 이론 등이 있음
  - 정보 이론에서 핵심 측정량은 엔트로피
- 엔트로피

열역학에서의 엔트로피와는 개념이 조금 다름

  - 엔트로피는 무작위 변수 혹은 무작위 과정의 결과에 포함되어 있는 불확정성의 양을 정량화 함
  - 예를 들어, 양쪽 면이 나올 확률이 같은 동전은, 6개의 면이 나올 확률이 같은 주사위보다 적은 정보량을 제공하며, 따라서 적은 엔트로피를 가짐
  - 엔트로피 이외에도 정보이론에서 중요시하는 측정량으로는 상호정보, 채널용량, 상대 엔트로피 등이 있음

엔트로피

- 하나의 확률 변수 X가 x값을 갖기 위한 정보량의 정의

$$-log_2 P(X = x)$$

로그의 밑이 2 → 이때 정보량의 단위: 샤넌 또는 bit
자연로그 사용 시 → 정보량의 단위: 내트(nat)

  - 예시: 동전 던지기에서 앞면이 나올 정보량과 주사위를 던져서 1이 나올 정보량을 비교해 보면
    - 동전 던지기에서 앞면이 나올 확률 = $\frac{1}{2}$ → 정보량 = $-log_2 \frac{1}{2} = 1$
    - 주사위를 던져서 1이 나올 확률 = $\frac{1}{6}$ → 정보량 = $-log_2 \frac{1}{6} = 2.5849$
  - 예시의 결과: 주사위의 정보량이 높음 → 불확실성이 높음

- 엔트로피는 확률 변수 X의 정보량에 대한 기댓값으로 다음을 정의함

$$H(X) = -\sum_i p_i log\, p_i = E(-log\, p(X))$$

엔트로피

- **결합 엔트로피**
  - 2개의 이산 확률변수의 결합 엔트로피는 (X, Y) 짝에 대한 엔트로피임

$$H(X,Y) = -\sum_{x,y} p(x,y) log\, p(x,y) = E_{X,Y}(-log\, p(X,Y))$$

- **조건부 엔트로피**
  - 확률변수 Y에 대한 다른 확률변수 X의 불확실성, 즉 조건부 엔트로피는 다음과 같이 정의됨

$$H(X|Y) = E_Y(H(X|Y=y) = -\sum_y p(y) \sum_x p(x|\, y) log\, p(x|y) = -\sum_{x,y} p(x,y) log\, p(x|y)$$

엔트로피

- 상호 정보(Mutual Information)
  - 다른 확률변수를 관찰함으로써 관심있는 확률변수에 대하여 얻을 수 있는 정보량에 대한 측도

$$I(X;Y) = \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} = E_{X,Y}(log \frac{p(x,y)}{p(x)p(y)})$$

  - 만약 2개의 확률변수가 독립이라면, 상호 정보는 0 이 됨
  - 일반적으로 $p(x,y) \geq p(x)p(y)$ 이므로 상호정보는 항상 0보다 크거나 같게 됨

$$log \frac{p(x,y)}{p(x)p(y)} = log\, p(x,y) - log\, p(x) - log\, p(y)$$

  - 따라서

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

엔트로피

- **교차 엔트로피**
  - **확률변수 X에 대하여 X의 실제 분포를 $p(X)$,임의의 분포를 $q(X)$라고 할 때, 교차 엔트로피의 정의**

$$H(p,q) = -E_p log\ q = -\sum_x p(x) \log q(x)$$