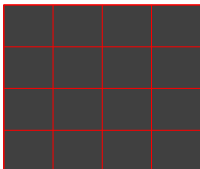


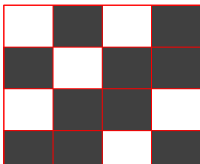
14. 엔트로피(entropy)

물리학에서 엔트로피는 물질의 열적 상태를 나타내는 열역학적 양(量)이다. - 열역학 제2법칙 이를 다르게 표현하면, 물리학에서 엔트로피는 어떠한 물리계의 무질서 정도를 의미한다.

1. 정보이론에서 엔트로피

정보이론에서 엔트로피는 어떠한 데이터를 표현하기 위한 평균 정보량을 의미한다
정보이론에서 엔트로피가 높다는 것은 정보의 양이 많다는 것을 의미한다.

	<ul style="list-style-type: none"> 좌측 그림은 모든 영역이 검은색이므로 데이터의 불확실성이 없다. 그림(데이터)에서 어떤 부분을 선택하더라도 검은색이 추출된다. 데이터 표현은 검은색만을 표현하기 위한 1bit만 있으면 충분하다.
---	--



	<ul style="list-style-type: none"> 좌측 그림은 흰색과 검은색이 무질서하게 분포되어 있다. 데이터를 표현하려면, 여러 비트가 필요하다.(0101 1011 0110 1101) 좌측 그림은 위의 그림에 비해 불확실성이 높다. 좌측 그림은 위의 그림에 비해 엔트로피(정보량)가 높다.
--	--

// 예제 1 : 동전과 주사위를 던질 때 사건

동전	<ul style="list-style-type: none"> 하나의 동전을 던질 때, 나오는 표본공간은 {앞면, 뒷면}이다. 즉, 경우의 수는 2가지이다. 동전 던지기는 주사위 던지기에 비해 사건은 자주 발생된다. 자주 발생하는 사건은 낮은 정보량을 가진다.
주사위	<ul style="list-style-type: none"> 하나의 주사위를 던질 때, 나오는 표본공간은 {1, 2, 3, 4, 5, 6}이다. 즉, 경우의 수는 6가지이다. 주사위 던지기는 동전 던지기에 비해 사건은 덜 자주 발생된다. 덜 자주 발생하는 사건은 더 높은 정보량을 가진다. 정보량이 많을수록 엔트로피가 증가한다.(예측하기가 어렵다)

- 잘 알지만, 동전이 주사위 보다 더 낮은 정보량을 갖는다는 의미이다.(당연)
- 동전 앞면이 나오는 사건은 주사위 눈이 1이 나오는 사건보다 더 자주 발생된다.
- 여기서, 엔트로피는 주사위 던지기가 동전 던지기보다 크다고 할 수 있다.
- 하나의 사건이 확실하게 일어나는 경우의 엔트로피는 0이다.(엔트로피 0은 결정된 정보)

// 예제 2 : 서로 다른 모양의 주사위를 던질 때, 눈이 나올 확률

	<ul style="list-style-type: none"> • 좌측 그림의 주사위를 던질 때, 각 눈의 수가 나올 확률이 1/6로 같다. • 확률변수가 가지는 모든 값의 발생 확률이 같은 주사위이다. • 확률변수가 가지는 확률이 비슷해질수록 엔트로피는 증가한다. • 확률변수가 가지는 모든 값의 발생 확률이 같을 때, 엔트로피는 최대값을 가진다.
	<ul style="list-style-type: none"> • 먼저, 좌측 그림의 주사위는 엄격하게 말하면 주사위가 아니다. • 만약, 좌측 그림의 주사위를 던질 때, 각 눈의 수가 나올 확률은 같지 않다. • 확률변수가 가지는 모든 값의 발생 확률이 같지 않은 주사위이다. • 눈 1이 나올 확률이 눈 3이 나올 확률보다 높다.(예측하기가 쉽다) • 예측하기가 쉽다는 것은 불확실도가 낮은 것이다.(엔트로피가 작다)

- 엔트로피는 불확실도(무질서도)를 나타내는 것이다.
- 엔트로피가 낮을수록 예측하기가 쉽고, 엔트로피가 높을수록 예측하기가 어렵다.
- 블록암호에서 엔트로피는 높을수록 안전하다.(정보의 불확실도가 높으므로 키 유추가 어렵다)

2. 섀넌(shannon) 엔트로피

- 섀넌 엔트로피는 정보 이론 창시자인 수학자 클로드 섀넌이 1948년 창안한 개념이다.
- 엔트로피는 정보기술, 암호학, 데이터 압축 등에서 중심이 되는 개념이다.
- 섀넌 엔트로피는 각 메시지에 포함된 정보의 기댓값(평균)이다.
- 섀넌은 정보 단위를 2진법 또는 비트를 선택했다.(로그 밑이 2가 됨)
- 섀넌 엔트로피에서 사건의 정보량은 음의 로그를 취한 수식이다.
- 확률변수 X 의 값이 x 인 사건의 정보량 : $I(x) = -\log_2 P(x)$

$$I(x) = -\log_2 P(x) = -\log_2 \frac{1}{2} = -\log_2 2^{-1} = 1$$

$$I(x) = -\log_2 P(x) = -\log_2 \frac{1}{4} = -\log_2 2^{-2} = 2$$

$$I(x) = -\log_2 P(x) = -\log_2 \frac{1}{8} = -\log_2 2^{-3} = 3 \leftarrow \text{발생 확률이 낮을수록 } I(x) \text{ 값은 커진다.}$$

// 왜? 섀넌 엔트로피 공식은 음의 로그를 취한 수식일까?.

- 수식이 음의 로그 형태이어야 정보량이 많을수록 수식의 값은 커지게 된다.
- 엔트로피는 정보량을 나타낸다.
- 엔트로피가 클수록 많은 정보량의 의미를 표현하려면 수식이 음의 로그 형태이어야 한다.

// 동전과 주사위를 각각 1개 던질 때 정보량

동전	<ul style="list-style-type: none"> • 동전을 던져 앞면이 나오는 확률은 1/2 • 정보량 $I(x) = -\log_2 \frac{1}{2} = 1$ (주사위에 비해 낮은 정보량)
주사위	<ul style="list-style-type: none"> • 주사위를 던져 눈이 1이 나오는 확률은 1/6 • 정보량 $I(x) = -\log_2 \frac{1}{6} = 2.5849$ (동전에 비해 높은 정보량)

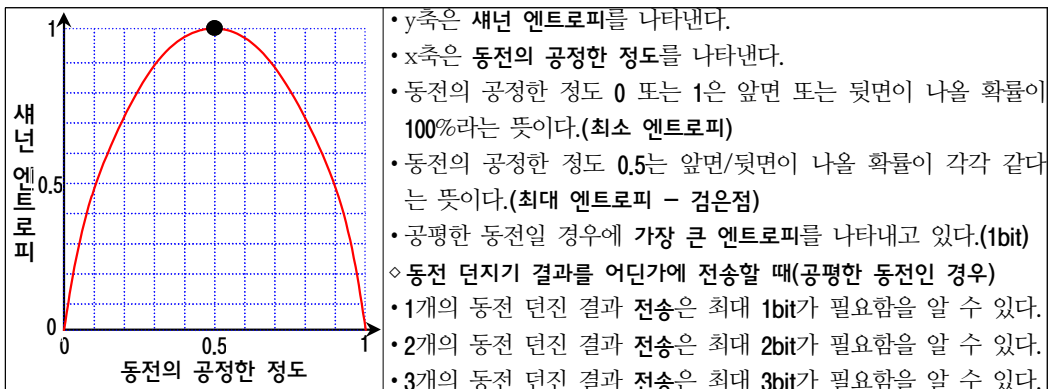
- 1개 동전을 던지면, 2가지 결과가 발생하고, 새넨 엔트로피는 1bit가 된다.
- 2개 동전을 던지면, 4가지 결과가 발생하고, 새넨 엔트로피는 2bit가 된다.
- m 개 동전을 던지면, 2^m 가지 결과가 발생하고, 새넨 엔트로피는 m bit가 된다.
- 비트의 개수와 새넨이 동일한 경우는 모든 결과의 발생 확률이 동일한 경우로 한정된다.
- 즉, 엔트로피는 정보를 최적으로 인코딩(부호화)하기 위해 필요한 bit 수이다.
- 예 : 월(000), 화(001), 수(010), 목(011), 금(100), 토(101), 일(110) → 3bit가 필요

// 정리 : 1개의 동전을 던질 때 표본공간의 확률이 다음과 같은 경우(균일분포)

표본공간	경우의 수	확률	설 명
앞면	1	$\frac{1}{2}$	앞면이 나올 확률이 $\frac{1}{2}$ 이라는 뜻
뒷면	1	$\frac{1}{2}$	뒷면이 나올 확률이 $\frac{1}{2}$ 이라는 뜻

- 새넨 엔트로피 $I(x) = -\log_2 P(x) = -\log_2 \frac{1}{2} = -\log_2 2^{-1} = 1$
- 위의 식에서 밑이 2인 경우 정보량의 단위를 새넨(shannon) 또는 비트(bit)라고 한다.

// 동전을 한번 던졌을 때 새넨 엔트로피의 변화



- 하나의 동전 던지기의 결과는 1bit에 해당하는 정보를 가지게 된다.
- 1bit가 가질 수 있는 엔트로피의 최댓값은 1이다.

// 예제 1 : 2개의 동전을 던질 때 표본공간의 확률이 다음과 같은 경우(균일분포)

표본공간	경우의 수	확률	설 명
앞면/앞면	1	$\frac{1}{4}$	앞면/앞면이 나올 확률이 $\frac{1}{4}$ 이라는 뜻
앞면/뒷면	1	$\frac{1}{4}$	앞면/뒷면이 나올 확률이 $\frac{1}{4}$ 이라는 뜻
뒷면/앞면	1	$\frac{1}{4}$	뒷면/앞면이 나올 확률이 $\frac{1}{4}$ 이라는 뜻
뒷면/뒷면	1	$\frac{1}{4}$	뒷면/뒷면이 나올 확률이 $\frac{1}{4}$ 이라는 뜻

• 새년 엔트로피 $I(x) = -\log_2(P(x)) = -\log_2 \frac{1}{4} = 2 \leftarrow$ 균일분포인 경우

// 예제 2 : 만약, 2개의 동전을 던질 때 표본공간의 확률이 다음과 같은 경우(비균일분포)

표본공간	경우의 수	확률	설 명
앞면/앞면	1	$\frac{1}{2}$	앞면/앞면이 나올 확률이 $\frac{1}{2}$ 로 가장 높다는 뜻
앞면/뒷면	1	$\frac{1}{8}$	앞면/뒷면이 나올 확률이 $\frac{1}{8}$ 로 가장 낮다는 뜻
뒷면/앞면	1	$\frac{1}{8}$	뒷면/앞면이 나올 확률이 $\frac{1}{8}$ 로 가장 낮다는 뜻
뒷면/뒷면	1	$\frac{1}{4}$	뒷면/뒷면이 나올 확률이 $\frac{1}{4}$ 이라는 뜻

$$\begin{aligned}
 \bullet I(x) &= \left(\frac{1}{2} \times -\log_2 \frac{1}{2}\right) + \left(\frac{1}{8} \times -\log_2 \frac{1}{8}\right) + \left(\frac{1}{8} \times -\log_2 \frac{1}{8}\right) + \left(\frac{1}{4} \times -\log_2 \frac{1}{4}\right) \\
 &= \left(\frac{1}{2} \times 1\right) + \left(\frac{1}{8} \times 3\right) + \left(\frac{1}{8} \times 3\right) + \left(\frac{1}{4} \times 2\right) \\
 &= 0.5 + 0.375 + 0.375 + 0.5 = 1.75
 \end{aligned}$$

// 새년 엔트로피를 수식으로 정리하면 다음과 같다.

.....

$$\text{새년 엔트로피 } I(x) = -\sum(\text{사건발생확률}) \cdot \log_2(\text{사건발생확률}) = -\sum_i P_i \cdot \log_2(P_i)$$

.....

- 엔트로피는 가능한 모든 사건이 같은 확률로 일어날 때 최댓값을 갖는다.
- 각 표본공간의 확률이 모두 동등한 상황에서 조금만 벗어나도 엔트로피는 감소한다.
- 엔트로피는 균일분포일수록 높고, 비균일분포일수록 낮아진다.
- 엔트로피는 정보량을 의미한다. 엔트로피 값이 클수록 정보량이 많다는 것을 의미한다.

기출문제 분석

1. 엔트로피에 대한 설명으로 옳은 것만을 모두 고른 것은? [2020년 국가 7급]

- ㄱ. 한 비트가 가질 수 있는 엔트로피의 최댓값은 1이다.
- ㄴ. 블록 암호문의 엔트로피는 낮을수록 안전하다.
- ㄷ. 엔트로피는 정보량 또는 정보의 불확실도를 측정하는 수학적 개념이다.
- ㄹ. 어떤 확률변수가 가질 수 있는 모든 값의 발생 확률이 같을 경우, 엔트로피는 최솟값을 갖는다.

- ① ㄱ, ㄴ ② ㄱ, ㄷ
- ③ ㄴ, ㄹ ④ ㄷ, ㄹ

☞ 엔트로피(entropy)

- ㄴ. 블록 암호문의 엔트로피는 낮을수록 안전하다.(×)
→ ㄴ. 블록 암호문의 엔트로피는 높을수록 안전하다.
- ㄹ. 어떤 확률변수가 가질 수 있는 모든 값의 발생 확률이 같을 경우, 엔트로피는 최솟값을 갖는다.(×)
→ 어떤 확률변수가 가질 수 있는 모든 값의 발생 확률이 같을 경우, 엔트로피는 최댓값을 갖는다.

// 동전던지기

- 동전 던지기를 시행했을 때 결과값의 엔트로피는 공정한 동전일 때 가장 높게 나온다.
- 공정한 동전은 앞, 뒷면이 나올 확률이 각각 1/2로 같은 경우이다.
- 이런 경우가 불확실성이 가장 극대화되고 결과를 예상하기 가장 어렵다는 것을 의미한다.
- 동전 던지기의 결과는 1bit에 해당하는 정보를 가지게 된다.
- 1bit가 가질 수 있는 엔트로피의 최댓값은 1이다.

◆ 정보이론에서 엔트로피

- 엔트로피는 정보량 또는 정보의 불확실도(무질서도)를 측정하는 수학적 개념이다.
- 정보이론에서 발생될 확률이 낮을수록, 어떤 정보일지는 불확실하게 된다.
- 이때, 우리는 '정보가 많다', '엔트로피가 높다'라고 표현한다.
- 암호문의 엔트로피는 높을수록 안전하다.
- 엔트로피의 단위는 그 정의에 사용된 로그의 밑이 무엇인지에 따라 구분된다.
- 엔트로피 단위 : 섀넌(shannon), 내트(nat), 하틀리(hartley) 등이 있다.