10.1 엔트로피 ENTROPY

$Y=0$ 또는 $Y=1$인 두가지 값을 취하는 확률변수의 확률분포가 다음과 같이 세 종류가 있다고 하자.

- 확률분포 Y_1 : $P(Y=0)=0.5$, $P(Y=1)=0.5$
- 확률분포 Y_2 : $P(Y=0)=0.8$, $P(Y=1)=0.2$
- 확률분포 Y_3 : $P(Y=0)=1.0$, $P(Y=1)=0.0$

베이지안 관점에서의 Y_1, Y_2, Y_3의 의미

- Y_1 ⇒ y값에 대해 아무것도 모른다 0.5/0.5
- Y_2 ⇒ y값이 0이라고 믿지만 아닐 가능성도 있다는 것을 아는 상태 0.8/0.2
- Y_3 ⇒ y값이 0이라고 100% 확신하는 상태 1.0/0.0

확률분포가 가지는 이러한 차이를 하나의 숫자로 나타낸 것

⇒ 엔트로피

엔트로피의 정의

확률분포가 가지는 정보의 확실도 또는 정보량을 수치로 표현한 것.

- 확률분포에서 특정한 값이 나올 확률 상승 → 엔트로피 감소.
- 확률분포에서 여러가지 값이 비슷한 확률로 나옴 → 엔트로피 증가.
- 확률 또는 확률밀도가 특정값에 몰려있음 (뾰족함) → 엔트로피 감소.
- 확률 또는 확률밀도가 고루고루 퍼져있음 (완만함) → 엔트로피 증가.

★ 물리학에서의 엔트로피

- 물질의 상태가 분산되는 정도 = 엔트로피
  - 고루 분산 ⇒ 엔트로피 ↑
  - 특정 상태로 몰려있음 ⇒ 엔트로피 ↓

수학적으로, 엔트로피는 확률분포함수를 입력으로 받아 숫자를 출력하는 범함수 (functional)로 정의한다. $H[\,]$ 기호로 표시한다.

binary

확률변수 Y가 카테고리분포와 같은 이산확률변수이면, 다음처럼 정의한다.

$$H[Y] = -\sum_{k=1}^{K} p(y_k) \log_2 p(y_k)$$

자연로 아님.

for K = X가 가질 수 있는 클래스의 수

$p(y)$ = 확률질량함수 (pmf)

(★ 확률의 로그값은 항상 음수이므로, 음수기호를 붙여서 양수로 만들어줌.)

확률변수 Y가 정규분포와 같은 연속확률변수이면 다음처럼 정의한다.

$$H[Y] = -\int_{-\infty}^{\infty} p(y) \log_2 p(y)\, dy$$

for $p(y)$ = 확률밀도함수 (pdf)

엔트로피 정의에서 log base = 2로 정의된 이유

→ 정보통신과 관련을 갖는 역사적 배경!

* 엔트로피 계산에서 $p(y)=0$ 인 경우 log값이 정의되지 않으므로, 다음과 같은 극한값을 사용한다.

$$\lim_{p \to 0} p \log_2 p = 0$$

(python 계산)

$0 \cdot \log_2(0)$
↓
$eps * np.log2(eps)$

- 이 값은 미적분 로피탈의 정리 (l'Hôpital's rule)에서 구할 수 있다.

- Y_1, Y_2, Y_3 의 엔트로피 계산

$$H[Y_1] = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$H[Y_2] = -\frac{8}{10}\log_2\frac{8}{10} - \frac{2}{10}\log_2\frac{2}{10} \approx 0.72$$

$$H[Y_3] = \underbrace{-1 \cdot \log_2 1}_{=0.} \underbrace{- 0 \cdot \log_2 0}_{=0.} = 0$$

균고수 퍼짐 ↑ ↓ 특정한 값

III 엔트로피의 ~~확률변수~~ 성질

확률변수가 결정론적이면, 확률분포에서 특정한 하나의 값이 나올 확률이 1이다. 이 때 엔트로피는 0이 되고, 이값은 엔트로피가 가질 수 있는 최솟값이다.

반대로, 엔트로피의 최대값은 이산확률변수의 클래스 갯수에 따라 달라진다. 만약 이산확률분포가 가질 수 있는 값이 2^k개면, 엔트로피의 최대값은 각 값에 대한 확률이 모두 같은 값인 $\frac{1}{2^k}$ 이다.

엔트로피의 값은

$$H = -2^k \times \left(\frac{1}{2^k} \cdot \log_2 \frac{1}{2^k} \right) = k$$

(2^k개만큼을 더함.)

#클래스 2^3개 → $H=3$

#클래스 2^5개 → $H=5$

이다.

III 엔트로피의 추정

이론적인 확률밀도함수가 없고, 실제 데이터가 주어진 경우,

① 데이터에서 확률질량함수를 측정한후

② 추정결과를 가반으로 엔트로피계산

ex) 데이터가 모두 80개 있고, 그중 $Y=0$인 데이터가 k개, $Y=1$인 데이터가 k개 있는 경우 엔트로피는 1이다.

$$P(y=0) = \frac{40}{80} = \frac{1}{2}$$

$$P(y=1) = \frac{40}{80} = \frac{1}{2}$$

$$H[Y] = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = \frac{1}{2} + \frac{1}{2} = 1.$$

사이파이의 stats 서브패키지 — entropy 함수 제공
(base 인수값 = 2)

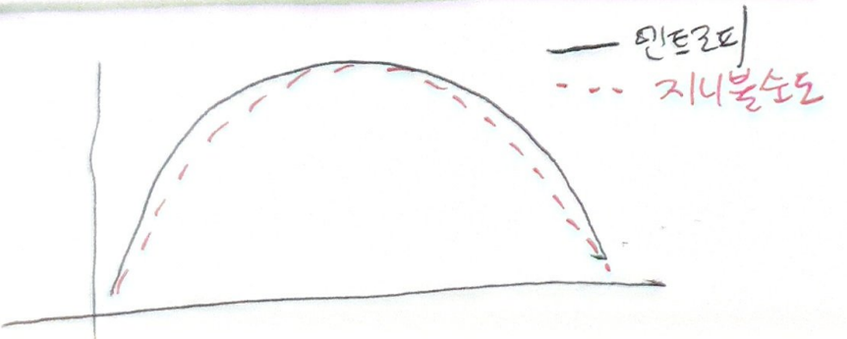
```
sp.stats.entropy([0.5, 0.5],
                 base=2)
```

지니불순도

엔트로피와 유사한 개념으로 지니불순도 (Gini impurity) 라는 것이 있다.

- 엔트로피와 유사하게, 확률분포가 치우쳐 있는가를 재는 척도
- ✓ log를 사용하지 않으므로 계산량이 더 적어서 엔트로피 대용으로 사용됨
  * 경제학에서의 '지니계수'와는 다르다.

$$G[Y] = \sum_{k=1}^{K} p(y_k)(1 - p(y_k))$$



엔트로피 최대화 ~(상식적으로 알고있을 것!)

기댓값 0, 분산 σ^2이 주어졌을 때 엔트로피 $H[p(x)]$를 가장 크게 만드는 확률밀도함수 $p(x)$는 정규분포가 된다. 이는 다음처럼 증명한다.

우선 확률밀도함수가 지켜야 할 제한조건은 다음과 같다.

(1) 확률밀도함수 총면적 = 1

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

(2) 기댓값 = 0

$$\int_{-\infty}^{\infty} x\,p(x)dx = 0$$

(3) 분산 = σ^2

$$\int_{-\infty}^{\infty} x^2 p(x)dx = \sigma^2$$

최대화할 목적함수(objective functional) → entropy.

$$H[p(x)] = -\int_{-\infty}^{\infty} p(x)\log_2 p(x)dx$$

라그랑주 승수법으로 제한조건을 추가하면 다음과 같아진다.

$+\lambda_1, \lambda_2, \lambda_3$

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

$$+ \lambda_1\left(\int_{-\infty}^{\infty} p(x)\,dx - 1\right)$$

$$+ \lambda_2\left(\int_{-\infty}^{\infty} x\,p(x)\,dx - 0\right)$$

$$+ \lambda_3\left(\int_{-\infty}^{\infty} x^2 p(x)\,dx - \delta^2\right)$$

$$= \int_{-\infty}^{\infty} \left(-p(x)\log p(x) + \lambda_1 p(x) + \lambda_2 x p(x) + \lambda_3 x^2 p(x) - \lambda_1 - \lambda_3\delta^2\right) dx$$

변분법에서 도함수를 다음과 같이 계산하여,

$$\frac{\delta H}{\delta p(x)} = -\log p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3 x^2 = 0$$

따라서 확률밀도함수의 형태는,

$$p(x) = \exp\left(-1 + \lambda_1 + \lambda_2 x + \lambda_3 x^2\right)$$

정규분포 특히 1차원의 확률밀도함수의 면적, 기댓값, 분산을 계산하고
주어진 제한조건을 만족하도록 연립방정식을 풀면, 라그랑주 승수를 다음처럼 구할 수 있다.
(풀이과정은 생략)

$$\lambda_1 = 1 - \frac{1}{2}\log 2\pi\delta^2$$

$$\lambda_2 = 0$$

$$\lambda_3 = -\frac{1}{2\delta^2}$$

위 값을 대입하면, 정규분포

$$\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

"정규분포 ⟶ Entropy ↑↑"

따라서, 정규분포는

- 기댓값과 표준편차를 알고있는 확률분포 중에서 가장 엔트로피가 크고 따라서 가장 정보가 적은 확률분포

✡ 정규분포는 베이즈 추정에 있어서 "사실상의 무정보" 사전 확률분포로 자주 사용한다.

✓ 정규분포 ~ 자연계 현상 ~ 자연계는 엔트로피가 높아지려고하는경향

가변길이 인코딩

엔트로피를 최대로 할때 통신문에서 데이터가 지니고있는 정보량을 계산하기위해 고안되었다.

예를 들어, 4개의 문자 A,B,C,D로 이루어진 다음 문서가 있다고하자.

'BDABABACBABAACBBAAACBBCAB...' (200자)

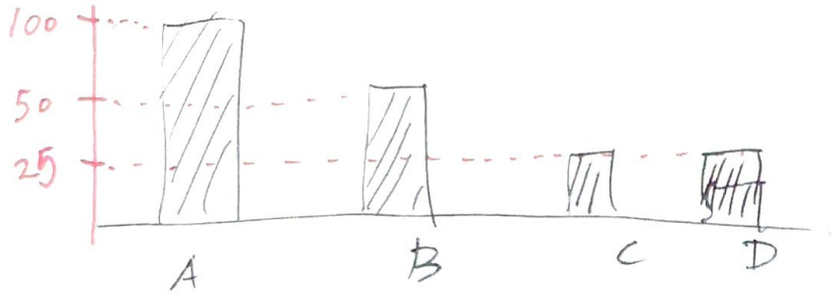
이문서를 0과 1로 이루어진 이진코드로 변환할때, 다음과같이 인코딩한다.

- A = "00"
- B = "01"
- C = "10"
- D = "11"

⟶ (400자)

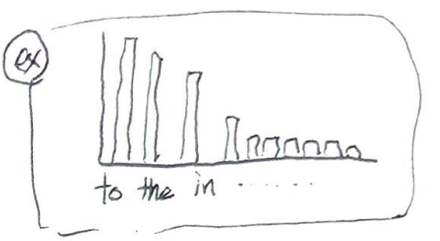
? 글자수를 더 줄일 수 없을까?

→ 글자 수의 분포 확인.



$P(Y=A)=\frac{1}{2}$, $P(Y=B)=\frac{1}{4}$, $P(Y=C)=\frac{1}{8}$, $P(Y=D)=\frac{1}{8}$

(* 지프의 법칙 (Zipf's Law)에 따르면,
이러한 분포는 형성의 글자 빈도수에서도 흔히 나타난다.



확률분포가 위와 같을 때, 다음처럼 인코딩한 후의 이진수 수를 줄일 수 있다.

more frequent

- A = "0"
- B = "10"
- C = "110"
- D = "111"

} 글자마다 인코딩한 이진수 숫자가 다르다
= 가변길이 인코딩 (variable length encoding)

↳ 가장 많이 출현하는 'A' : 1글자
'C', 'D'는 3글자, but 빈도수적에서 영향↓

문서의 분포 = 위에서 지정한 분포일 때,
인코딩된 이진수의 숫자는 350개가 된다.

$(200\times\frac{1}{2})\cdot 1 + (200\times\frac{1}{4})\cdot 2 + (200\times\frac{1}{8})\cdot 3 + (200\times\frac{1}{8})\cdot 3 = 350$

따라서, 알파벳 한 글자를 인코딩하는데 필요한 평균 비트수는 $350/200 = 1.75$
이 값은 확률변수의 엔트로피 값과 같다 ✓

$H = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{2}{8}\log_2\frac{1}{8} = 1.75$

• 연습문제 10.1.4

A, B, C, D, E, F, G, H의 8글자로 이루어진 문서가 있고,
각각의 글자가 나올 확률이 다음과 같다고 가정하자.

$$\left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right\}$$

① 이 문서를 위한 가변길이 인코딩 방식을 서술하고,
② 한 글자를 인코딩하는 데 필요한 평균 비트수를 계산하라.

$2^3 = 8$

NON-가변길이 인코딩	가변길이 인코딩
000	0 — $\frac{1}{2}$
001	10 — $\frac{1}{4}$
010	110 — $\frac{1}{8}$
011	1110 — $\frac{1}{16}$
100	11100 } $\frac{1}{64}$
101	11101 } $\frac{1}{64}$
110	11110 } $\frac{1}{64}$
111	11111 } $\frac{1}{64}$

(avg. 3bit)

↳ 한 글자를 인코딩하는 데 필요한 평균 비트수 = 엔트로피

$$H = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{1}{64}\log_2\frac{1}{64} - \frac{1}{64}\log_2\frac{1}{64} - \frac{1}{64}\log_2\frac{1}{64} - \frac{1}{64}\log_2\frac{1}{64}$$

$$= 2 \text{ bit}$$