8.2 베르누이분포와 이항분포

가장 단순한분포 & 분류문제에서 널리 사용.

- 두 분포의 개념 학습
- 두 분포가 스팸메일 필터링에 어떻게 쓰이는지 알아보기

베르누이 시행

결과가 두가지 중 하나만 나옴, Bernoulli trial

ex) 동전던지기 - H or T

베르누이 확률변수

베르누이 시행의 결과를 상수 0 또는 1로 바꾼 것 = 베르누이 확률변수 (Bernoulli random variable)

- 이산확률변수 (discrete random variable)
  ↳ 보통 정수 1, 0 으로표현

베르누이 확률분포

$$X \sim Bern(x;\mu)$$

X: 분포X, ~: "따른다", x: 출력값, μ: 분포 parameter (모수)

$$Bern(x;\mu) = \begin{cases} \mu & \text{if } x=1, \\ 1-\mu & \text{if } x=0 \end{cases}$$

$$\boxed{Bern(x;\mu) = \mu^x(1-\mu)^{(1-x)}}$$

베르누이 확률변수 표본값이 1, 0이 아니라 1, -1이라면

$$Bern(x;\mu) = \mu^{(1+x)/2}(1-\mu)^{(1-x)/2}$$

베르누이분포의 모멘트

- 기댓값

$$E[X] = \mu$$

(증명)
$$E[X] = \sum_{x_i \in \Omega} x_i P(x_i)$$
$$= 1 \cdot \mu + 0 \cdot (1-\mu)$$
$$= \mu$$

- 분산

$$Var[X] = \mu(1-\mu)$$

(증명)
$$Var[X] = \sum_{x_i \in \Omega} (x_i - \mu)^2 P(x_i)$$
$$= (1-\mu)^2 \cdot \mu + (0-\mu)^2 \cdot (1-\mu)$$
$$= \mu(1-\mu)$$

* 넘파이 계산

np.var(X, ddof=1)

↳ $\frac{N}{N-1} E[S^2] = \sigma^2$

이항분포 binomial distribution

성공확률이 μ인 베르누이 시행을 N번 반복하는 경우

- N번 중 성공횟수를 X라고 하면, X는 $0 \sim N$ 사이 정수값이 됨.
- 이러한 확률변수를 이항분포(binomial distribution)를 따르는 확률변수라 하며, 다음과 같이 표시.

$$X \sim Bin(x; N, \mu)$$

베르누이 분포와 이항분포는 모두 베르누이 확률변수에서 나온 표본값.

→ 표본데이터 1개 → $Bern(x; \mu)$

표본데이터 N개 ($N>1$) → $Bin(x; N, \mu)$

Y가 베르누이 분포를 따르는 분포일 때,

$$Y \sim Bern(x; \mu)$$

이 확률변수의 N개의 표본을 $y_1, y_2, \cdots, y_N$이라고 하자.

이 값은 모두 0(실패) 아니면 1(성공)이라는 값을 가지기 때문에

N번 중 성공한 횟수는 N개의 표본값의 합이다.

$$x = \sum_{i=1}^{N} y_i$$

Bern 확률변수와 Bin 확률변수를 연결해주는 공식

베르누이분포를 따르는 확률변수 Y의 확률질량함수를 대입하여 정리하면, 이항분포 확률변수 X의 확률질량함수는 다음과 같이된다.

$$Bin(x; N, \mu) = \binom{N}{x} \mu^x (1-\mu)^{N-x}$$

이 식에서 $\binom{N}{x}$ 기호는 조합(combination)이라는 기호로, N개 원소 중에 x개 원소를 순서와 상관없이 선택할 수 있는 경우의 수를 뜻한다. 조합은 다음 공식으로 계산할 수 있다.

$$\binom{N}{x} = \frac{N!}{x!(N-x)!}$$

! 기호는 팩토리얼(factorial)이라 하며 다음과 같이 정의한다.

$$N! = 1 \cdot 2 \cdot 3 \cdot \cdots \cdot (N-1) \cdot N$$

이항분포의 모멘트

이항분포의 기댓값과 분산은 각각 다음과 같다.

- 기댓값

$$E[X] = N\mu$$

(증명)

$$E[X] = E\left[\sum_{i=1}^{N} Y_i\right] = \sum_{i=1}^{N} E[Y_i] = N\mu$$

여기에서 Y_i는 서로 독립인 베르누이 분포이다.

- 분산

$$Var[X] = N\mu(1-\mu)$$

(증명)

$$Var[X] = Var\left[\sum_{i=1}^{N} Y_i\right] = \sum_{i=1}^{N} Var[Y_i] = N\mu(1-\mu)$$

베르누이분포와 이항분포의 모수추정

데이터에서 모수의 값을 찾아내는 것을 모수추정 (parameter estimation) 이라 한다.

$$\hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{N_1}{N}$$

(N = 전체 데이터수, N_1 = 1이 나온 횟수)

베르누이 분포의 활용

베르누이분포는 다음과 같은경우에 사용된다.

1. 분류예측 문제의 출력 데이터가 두 값으로 구분되는 카테고리인 경우에 분류결과 즉, 두 값중 어느 값이 가능성이 높은지를 표현하는데 사용된다 (베르누이 관점)
2. 입력 데이터가 0 또는 1 혹은 참/거짓, 두 개의 값으로 구분되는 카테고리인 경우, 두 종류의 값이 나타나는 비율을 표현하는데 사용 (빈도주의적 관점)

<예제>

spam mail filter : "메일 10통, 스팸메일 6통"

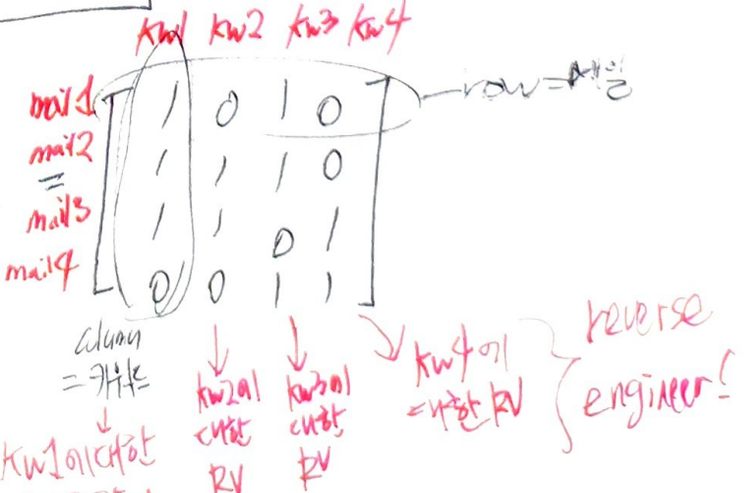
$$p(y) = Bern(y; \mu = 0.6)$$

(확률변수 $Y = 1$ for spam mail).

* 특정한 단어를 갖고있는가? ··· Bag of Words, 4개 단어를 확인.

$$x = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad X_{spam} \vdots$$

특징행렬!



이 때, 스팸 메일의 특성은 4개의 베르누이 확률변수의 튜플 (X_1, X_2, X_3, X_4)로 나타낼 수 있다.

- X_1 : 메일이 첫번째 키워드를 포함하면 1, 아니면 0이 되는 확률변수

$$p(\underbrace{X_1=1}_{kw1} \mid \underbrace{Y=1}_{spam}) = Bern(x_1; \mu_{spam,1})$$

- X_2 : 메일이 두번째 키워드를 포함하면 1, 아니면 0이 되는 확률변수

$$p(\underbrace{X_2=1}_{kw2} \mid \underbrace{Y=1}_{spam}) = Bern(x_2; \mu_{spam,2})$$

- X_3 : 메일이 세번째 키워드를 포함하면 1, 아니면 0이 되는 확률변수

$$p(\underbrace{X_3=1}_{kw3} \mid \underbrace{Y=1}_{spam}) = Bern(x_3; \mu_{spam,3})$$

- X_4 : 메일이 네번째 키워드를 포함하면 1, 아니면 0이 되는 확률변수

$$p(\underbrace{X_4=1}_{kw4} \mid \underbrace{Y=1}_{spam}) = Bern(x_4; \mu_{spam,4})$$

학습의 각 열로부터 각 베르누이 확률분포의 모수의 추정값을 구하면 다음과 같다.

$$\hat{\mu}_{spam,1} = \frac{3}{4}, \quad \hat{\mu}_{spam,2} = \frac{2}{4}, \quad \hat{\mu}_{spam,3} = \frac{3}{4}, \quad \hat{\mu}_{spam,4} = \frac{2}{4}$$

* 키워드 4개를 사용하는 스팸 메일 필터를 만드는데, 스팸 메일과 정상 메일들의 특성을 모두 모형화하려면 베르누이 확률변수가 몇 개 필요한가?

① 입력 for spam → 4개
② 입력 for non-spam → 4개
③ 출력 (bayesian) → 1개

9개!