8.3 카테고리분포와 다항분포

베르누이분포 → 이진분류문제
카테고리분포 → 다중분류문제 (multi-class classification)

카테고리 확률변수

동전이 아니라주사위를 던지면?

카테고리 확률변수 → 1부터 6까지 하나의 정수값 中 1
↳ 범주값, 카테고리, 클래스
↳ 주사위던지기 ⇒ K=6인 카테고리분포

"주사위던지기" (one-hot encoding)

$x=1 \longrightarrow x = (1,0,0,0,0,0)$
$x=2 \longrightarrow x = (0,1,0,0,0,0)$
$\vdots$
$x=6 \longrightarrow x = (0,0,0,0,0,1)$

확률변수 값 표시 $x = (x_1, x_2, x_3, x_4, x_5, x_6)$

$$x_i = \begin{cases} 0 \\ 1 \end{cases}$$ constraint

$$\sum_{k=1}^{K} x_k = 1$$ … 1은 하나만 등장한다는 의미.

카테고리 분포의 모수

$\mu = (\mu_1, \mu_2, \cdots, \mu_K)$

$0 \le \mu_i \le 1$ … 0부터 1 사이 어떤 실수값도 가능.

$\sum_{k=1}^{K} \mu_k = 1$ … ex) $0.2+0.2+0.2+0.2+0.2 = 1$

카테고리 확률분포

카테고리 확률변수의 확률분포인 **카테고리 확률분포**

$$Cat(x_1, x_2, \cdots, x_K ; \mu_1, \cdots, \mu_K)$$

x 를 확률벡터 $x = (x_1, x_2, \cdots, x_K)$,
모수벡터 $\mu = (\mu_1, \mu_2, \cdots, \mu_K)$ 를 사용하여

$$Cat(x ; \mu)$$

로 간단히 표기할 수 있다.

확률질량함수를 다음처럼 표기한다.

$$Cat(x;\mu) = \begin{cases} \mu_1 & \text{if } x = (1,0,0,\cdots,0) \\ \mu_2 & \text{if } x = (0,1,0,\cdots,0) \\ \mu_3 & \text{if } x = (0,0,1,\cdots,0) \\ \vdots & \vdots \\ \mu_K & \text{if } x = (0,0,0,\cdots,1) \end{cases}$$

one-hot-encoding 형태로 간략화 가능.

$$Cat(x;\mu) = \mu_1^{x_1} \mu_2^{x_2} \cdots \mu_K^{x_K} = \prod_{k=1}^{K} \mu_k^{x_k}$$

명제 8.3.1

$K=2$ 인 카테고리분포의 pmf (확률질량함수)가 베르누이분포의 pmf와 같음을 보여라.

$K=2$: $Cat(x;\mu) = \mu_1^{x_1} \mu_2^{x_2}$ — 제약조건 $\mu_1 + \mu_2 = 1$ $\therefore \mu_2 = 1 - \mu_1$

$$Cat(x;\mu) = \mu_1^{x_1} (1-\mu_1)^{x_2}$$

$x_1 + x_2 = 1$ 이므로 $x_2 = 1 - x_1$

✓ $$Cat(x;\mu) = \mu_1^{x_1} (1-\mu_1)^{(1-x_1)} = Bern(x;\mu)$$

카테고리분포의 파라미터

- 기댓값

$$E[x_k] = \mu_k$$

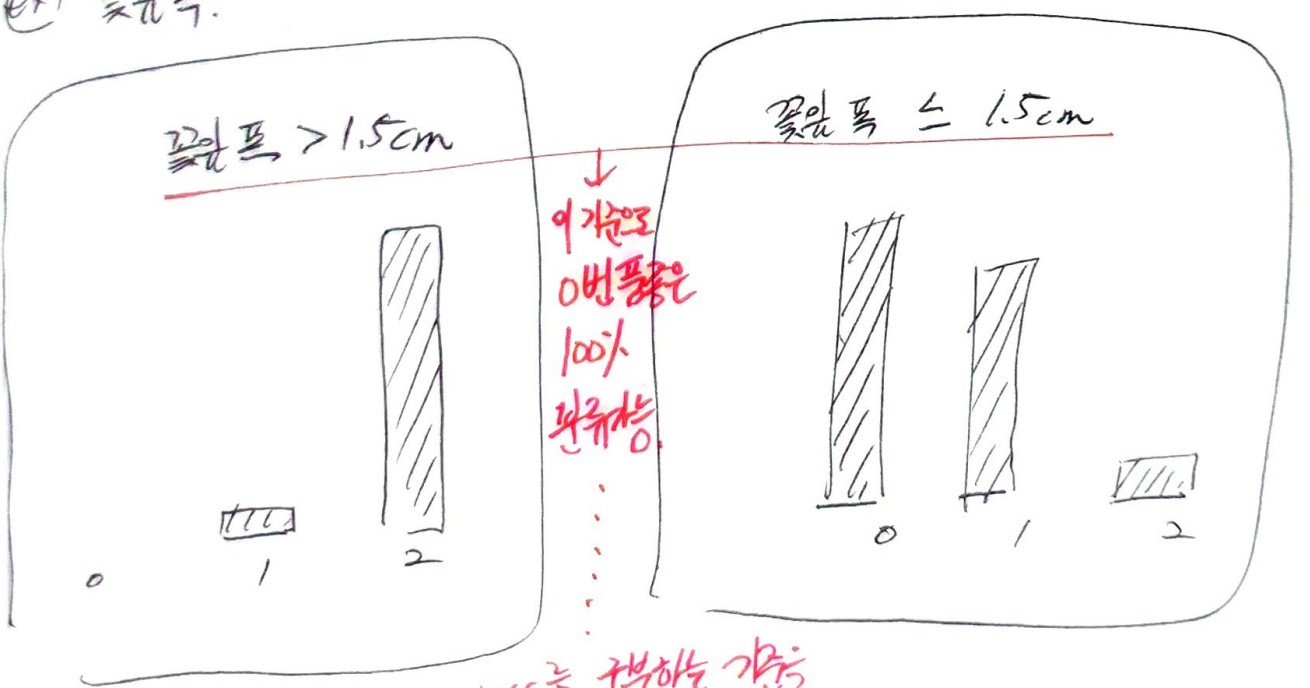
- 분산

$$Var[x_k] = \mu_k(1-\mu_k)$$

다중 분류문제 "multi-class classification"

예측범주가 2개 이상인 지도학습 문제 — "카테고리분포를 사용함."

(ex) 꽃잎폭.



class를 구분하는 기준을 탐색하고 적용

decision tree

⋮

random forest

다항분포 Multinomial distribution

베르누이 확률변수 데이터가 복수이면, 데이터의 합은 이항분포를 이룬다
카테고리 확률변수 데이터가 복수이면, 데이터의 합은 다항분포를 이룬다.

동전던지기 $\times N$ 앞면나온 횟수 ~ 이항분포
주사위던지기 $\times N$ 각 면이 나온 횟수집합의 분포 ~ 다항분포

[다항분포의 확률질량함수]

$$Mu(x; N, \mu)$$

$$Mu(x; N, \mu) = \binom{N}{x} \prod_{k=1}^{K} \mu_k^{x_k} = \binom{N}{x_1, \cdots, x_K} \prod_{k=1}^{K} \mu_k^{x_k}$$

이 식에서 조합계수는 다음과 같이 정의된다.

$$\binom{N}{x_1, \cdots x_K} = \frac{N!}{x_1! \cdots x_K!}$$

연습문제 8.8.3

$K=2$인 다항분포의 확률질량함수가 이항분포의 확률질량함수와 같음을 보여라.

$K=2$일때 다항분포의 확률질량함수

$$Mu(x; N, \mu) = \binom{N}{x_1, \cdots x_K} \prod_{k=1}^{K} \mu_k^{x_k} = \binom{N}{x_1, x_2} \prod_{k=1}^{2} \mu_1^{x_1} \mu_2^{x_2}$$

표본데이터 총합이 N개이므로 $x_1 + x_2 = N$ ... ①
모수의 총합이 1이므로 $\mu_1 + \mu_2 = 1$ ... ②

①, ②을 식에 대입하면

$$Mu(x; N, \mu) = \binom{N}{x_1, x_2} \prod_{k=1}^{2} \mu_1^{x_1} (1-\mu_1)^{N-x_1} = Bin(x; N, \mu)$$

다항분포의 모멘트

다항분포의 기댓값과 분산

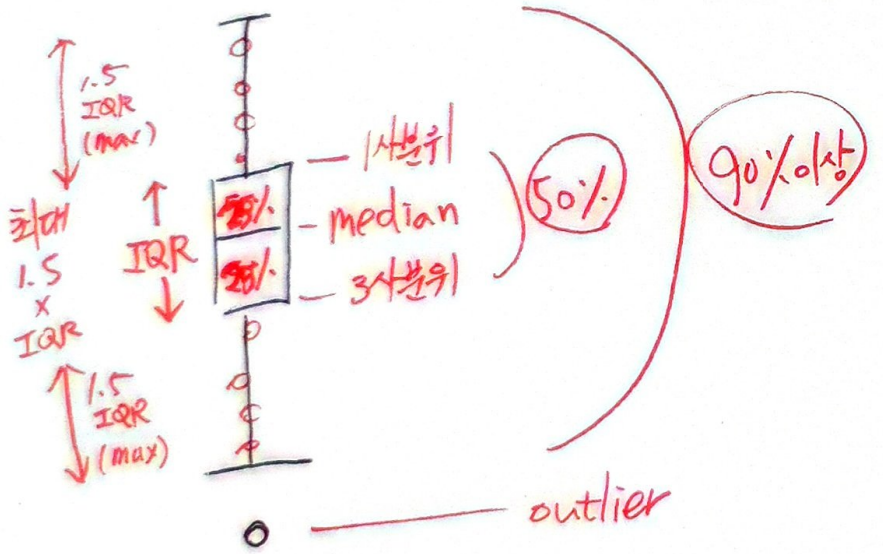
- 기댓값

$$E[x_k] = N\mu_k$$

- 분산

$$Var[x_k] = N\mu_k(1-\mu_k)$$

* Boxplot



** boxplot은 분포의 형태가 대체로 짐작될 것을 가정함.
(대칭 중앙에 많이 몰려있을 것이다)