

## 0) 모수 parameter, 통계량 statistics 비교 (2page)

08.05

스튜던트 t분포

카이제곱분포

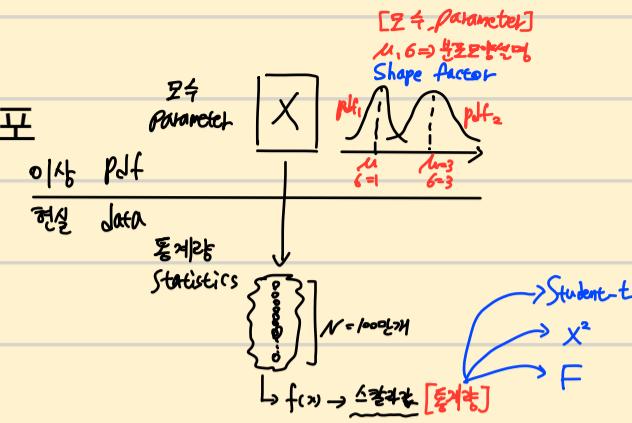
F분포

- 정규분포에서 파생된 분포 : 정규분포에서 생성된 데이터에 수식적용하면 분포모

양이 달라짐

- 적용된 수식에 따라 스튜던트 t, 카이제곱, F분포

- 즉, 다 '통계량 분포' 이다!



### 1) 스튜던트 t분포

(정규분포에서 얻은 표본평균을 표본표준편차로 정규화 한 통계량이 확률변수)

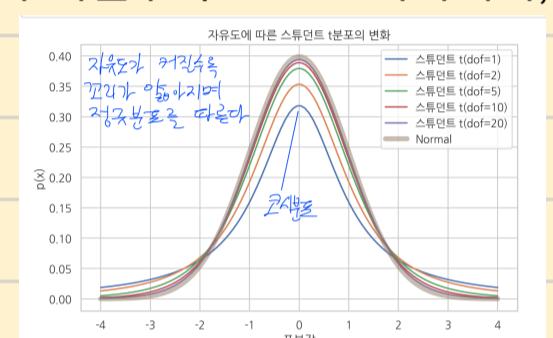
- 분포가 fat-tail (꼬리가 두꺼운) 모형에 적용하기 적합한 분포

- t분포의 pdf (모수 : 기댓값, 표준편차, 자유도)

- 자유도 1 = 코시분포 (양수인 한쪽만 다루면 하프코시분포)

- 자유도에 따른 스튜던트 t 분포의 변화 (자유도가 커질수록 fat-tail 사라지며,

정규분포화 ! ) (4 page)



- t 분포의 모멘트 (8.5.3 - 4)

스튜던트 t분포의 기댓값과 분산은 다음과 같다.

- 기댓값:

$$E[X] = \mu$$

- 분산:

$$\text{Var}[X] = \frac{\nu}{\lambda(\nu-2)}$$

분산의 대한 식은  $\nu > 2$  인 경우만 적용된다.  $\nu = 1, 2$  일 때는 분산이 무한대가 된다.

### 2) t 통계량

- z 통계량을 구하려면 확률분포 ( $X$ , 이상)의 정확한 평균, 표편을 알아야한다. 하지만, 이는 어렵다.

- 따라서, 표본평균을 표준편차로 정규화한 값 = t 통계량 (z통계량의 표본버전)

t 통계량은 자유도가  $N - 1$ 인 스튜던트 t분포를 이룬다.

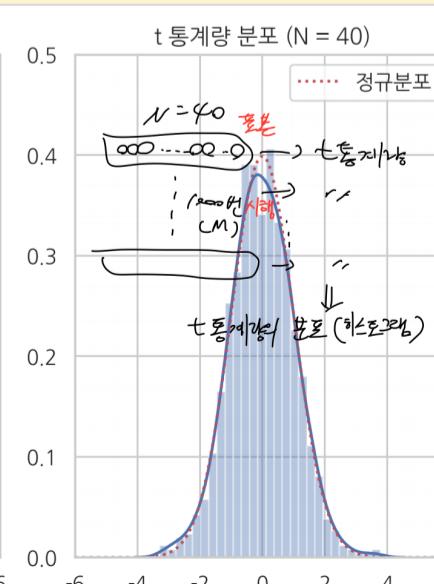
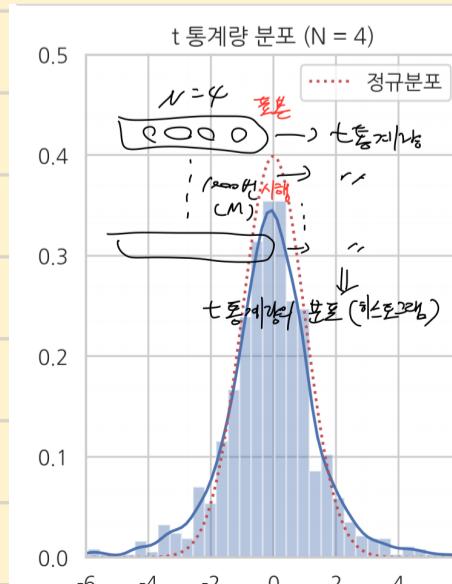
$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}} \sim t(x; 0, 1, N-1) \quad (8.5.5)$$

이 식에서  $\bar{x}$ ,  $s$ 은 각각 표본평균, 표본표준편차다.

$$\bar{x} = \frac{x_1 + \dots + x_N}{N} \quad (8.5.6)$$

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 \quad (8.5.7)$$

- 하지만, 자유도를 30이상으로 높이면, t분포가 정규분포와 거의 동일 ( $t = z$ )



사실상 정규분포와 같다.

t통계량

는 근정계량

$N \uparrow \Rightarrow t_{stat} \approx z_{stat}$

t분포는 정규분포

Why?

$\bar{x}$ 는  $M$ 을 정규화

$t = \frac{\bar{x}}{S}$ 를 정규화

$$E[\bar{x}] = E[x] = \mu$$

$$\frac{1}{N} S^2 = \sigma^2$$

이기 때문에,

$\bar{x}$ 는  $S$ 에 가까워지고,  $S^2$ 은  $\sigma^2$ 에 가까워짐.

08.05

### 3) 카이제곱분포 (정규분포에서 얻은 변수를 제곱+SUM 한 통계량이 확률변수)

스튜던트 t분포

- 8.5.8 ~ 10

카이제곱분포

F분포

그런데 이  $N$  개의 표본들을 단순히 더하는 것이 아니라 제곱을 하여 더하면 양수값만을 가지는 분포가 된다. 이 분포를 카이제곱(chi-squared)분포라고 하며  $\chi^2(x; v)$ 으로 표기한다. 카이제곱분포도 스튜던트 t분포처럼 자유도 모수를 가진다.

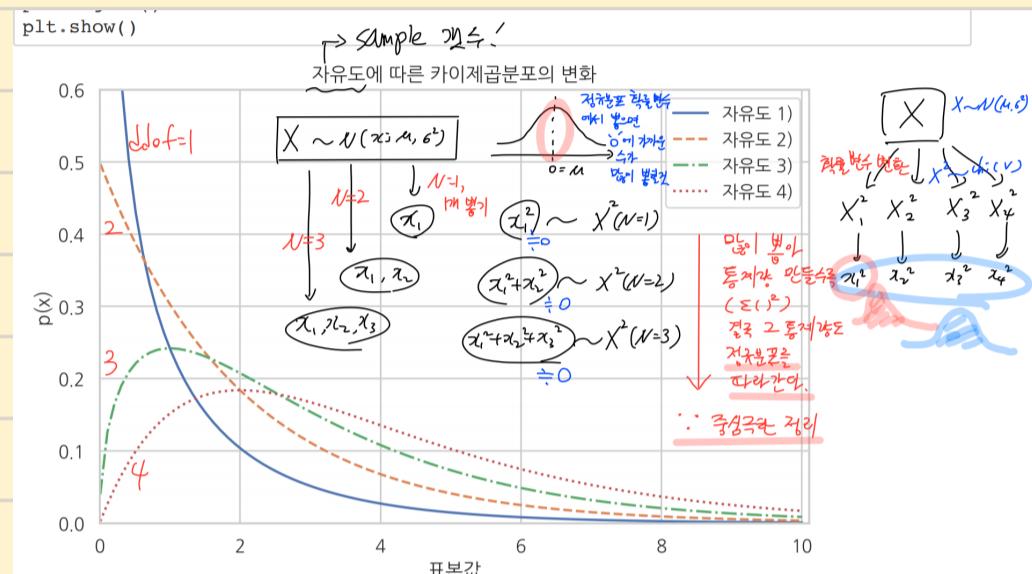
$$x_i \sim \mathcal{N}(x; \mu, \sigma^2) \quad (8.5.8)$$

$$\downarrow \quad (8.5.9)$$

$$\sum_{i=1}^N x_i^2 \sim \chi^2(x; v = N) \quad (8.5.10)$$

#### - 자유도에 따른 카이제곱분포의 변화(8page)

(많이 뽑을 수록 카이제곱 확률변수 (통계량) 의 분포는 정규분포화 된다 !)



### 4) F 분포

(카이제곱분포를 따르는 2개의 확률변수를 각각  $N$ (표본 수)로 나눠 비교 한 통계량이 확률변수)

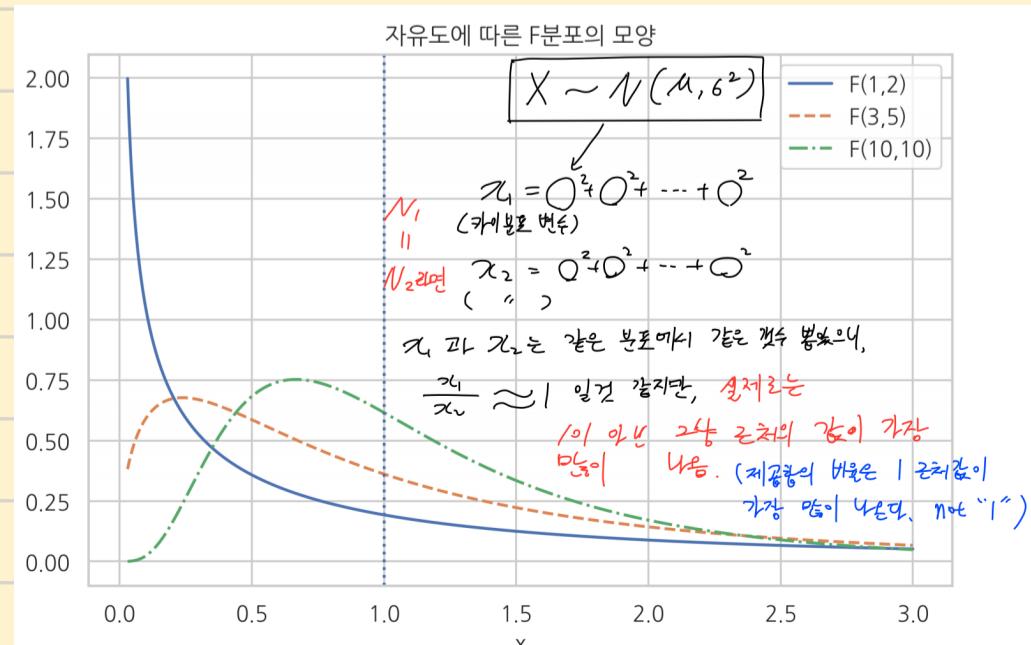
- t분포, 카이제곱분포 = 정규분포 확률변수에서  $N$ 개의 표본을 추출해 생성
- F분포 = 카이제곱분포에서 확률변수 2개 뽑아서 진행. 즉,  $N_1 + N_2$  개 표본 추출해 생성 (8.5.12)

이와 비슷하게 카이제곱분포를 따르는 독립적인 두 개의 확률 변수  $\chi^2_1(x; N_1)$ 과  $\chi^2_2(x; N_2)$ 의 확률 변수 표본을 각각  $x_1, x_2$ 이라고 할 때 이를 각각  $N_1, N_2$ 로 나눈 뒤 비율을 구하면  $F(x; N_1, N_2)$  분포가 된다.  $N_1, N_2$ 는 F분포의 자유도 모수라고 한다.

$$x_1 \sim \chi^2(N_1), x_2 \sim \chi^2(N_2) \rightarrow \frac{x_1}{N_1} \sim F(x; N_1, N_2) \quad (8.5.12)$$

$$x_2 \sim \chi^2(N_2) \quad \text{※ } x_i = \sum_{j=1}^{N_i} x_{ij}^2, x_i \sim \mathcal{N}(N_i, 6^2)$$

- 11 page



08.05

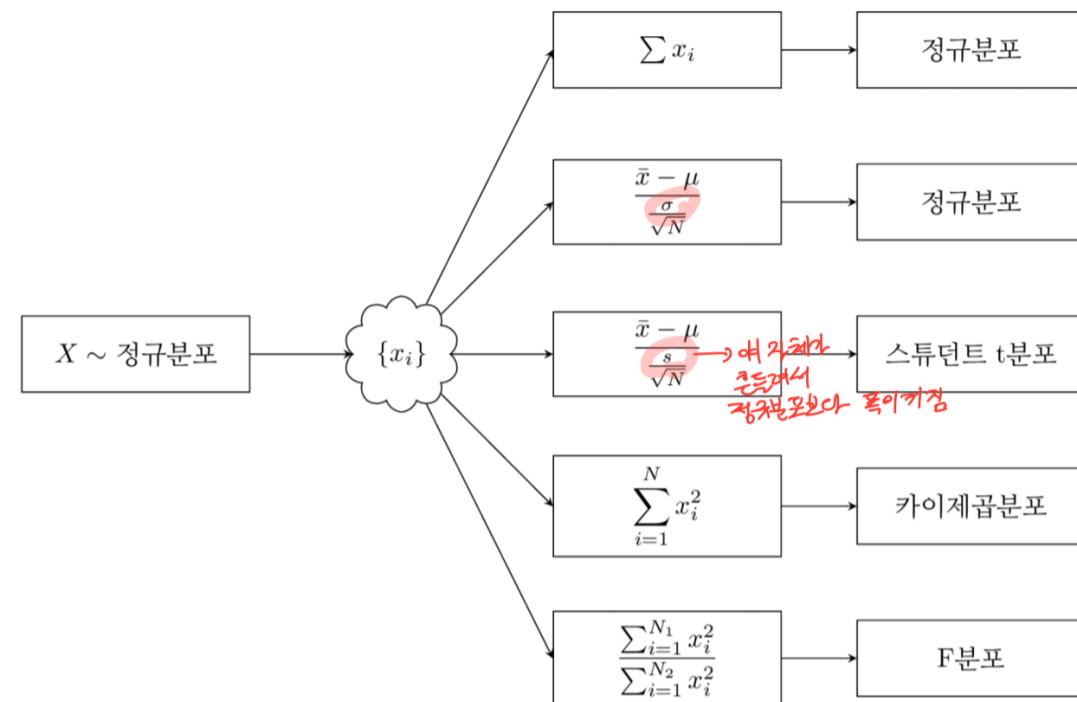
스튜던트 t분포

카이제곱분포

F분포

스튜던트 t분포, 카이제곱분포, F분포는 모두 정규분포의 통계량 분포(statistics distribution)의 일종이다. 선형회귀분석에서 이 통계량 분포들은 각각 다음 값에 대한 확률모형으로 사용된다.

- 스튜던트 t분포: 추정된 가중치에 대한 확률 분포
- 카이제곱분포: 오차 제곱합에 대한 확률 분포
- F분포: 비교 대상이 되는 선형모형의 오차 제곱합에 대한 비율의 확률 분포



08.06

## 다면수 정규분포

- 지금까지는 대부분 출력값이 스칼라인 분포를 봄
- 다변수 정규분포 : 벡터입력, 벡터출력되는 사건을 대상으로 함
- 다변수 정규분포의 pdf : 벡터 입력해서 해당 벡터가 출력될 확률값(스칼라)
- pdf 값이 같은 등고선은 원을 이룬다.

\* 대칭행렬  $\Rightarrow$  실수 고윳값.  
고윳벡터 직교  $\Rightarrow$  항상 대각화 가능

\* 선형독립  $\rightarrow$  직교  
 $\Leftrightarrow$

- 공분산 행렬로부터 상관계수 = 0 이면,

다면수정규분포 Pdf는 분리가능함수가 된다 = 벡터의 다변수들이 서로 독립적인 것  
(Pdf값이 같은 변수들이 독립적으로 있어, 등고선은 원을 이룬다)

다음과 같은 2차원( $D = 2$ ) 다변수정규분포를 생각하자. 2차원이므로 확률변수벡터는

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (8.6.2)$$

이다.

만약 모수가 다음과 같다고 하자.

$$\mu = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (8.6.3)$$

공분산행렬로부터  $x_1$ 과  $x_2$ 가 독립이라는 것을 알 수 있다. 확률밀도함수를 구하면 다음과 같다.

$$|\Sigma| = 1, \Sigma^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (8.6.4)$$

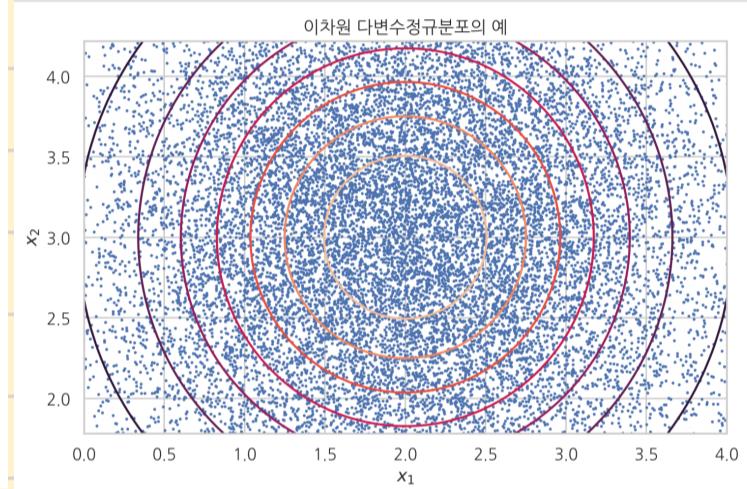
$$(x - \mu)^T \Sigma^{-1} (x - \mu) = [x_1 - 2 \ x_2 - 3] \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 - 2 \\ x_2 - 3 \end{bmatrix} \quad (8.6.5)$$

$$= (x_1 - 2)^2 + (x_2 - 3)^2 \quad \text{분리가능함수} \quad (8.6.6)$$

$$\mathcal{N}(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}((x_1 - 2)^2 + (x_2 - 3)^2)\right) \quad (8.6.7)$$

확률밀도함수값이 같은 등고선은 원이 된다.

$$(x_1 - 2)^2 + (x_2 - 3)^2 = r^2 \quad (8.6.7)$$



예제

만약 모수가 다음과 같다고 하자. 공분산행렬로부터  $x_1$ 과  $x_2$ 가 양의 상관관계가 있다는 것을 알 수 있다.

$$\mu = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \Sigma = \begin{bmatrix} 2 & 3 \\ 3 & 7 \end{bmatrix} \quad (8.6.8)$$

이 때 확률밀도함수는 다음과 같다.

$$|\Sigma| = 5, \Sigma^{-1} = \begin{bmatrix} 1.4 & -0.6 \\ -0.6 & 0.4 \end{bmatrix} \quad (8.6.9)$$

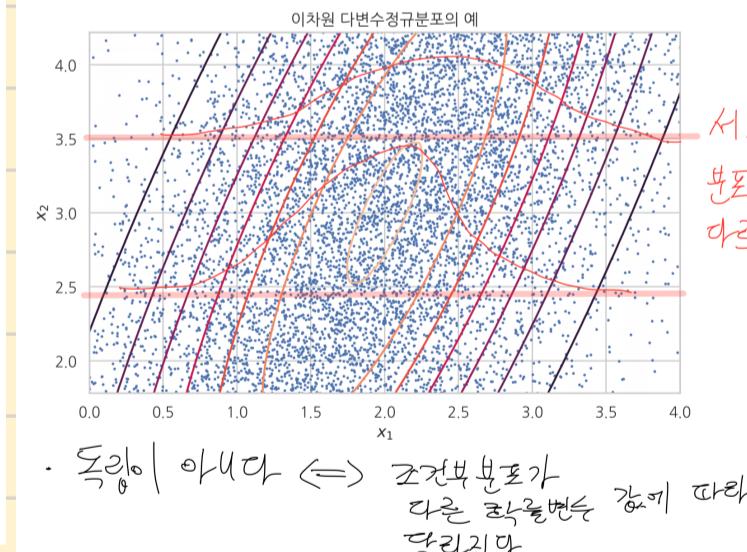
$$(x - \mu)^T \Sigma^{-1} (x - \mu) = [x_1 - 2 \ x_2 - 3] \begin{bmatrix} 1.4 & -0.6 \\ -0.6 & 0.4 \end{bmatrix} \begin{bmatrix} x_1 - 2 \\ x_2 - 3 \end{bmatrix} \quad (8.6.10)$$

$$= \frac{7}{5}(x_1 - 2)^2 - \frac{6}{5}(x_1 - 2)(x_2 - 3) + \frac{2}{5}(x_2 - 3)^2 \quad (8.6.11)$$

$$\mathcal{N}(x_1, x_2) = \frac{1}{2\sqrt{5}\pi} \exp\left(\frac{7}{5}(x_1 - 2)^2 - \frac{6}{5}(x_1 - 2)(x_2 - 3) + \frac{2}{5}(x_2 - 3)^2\right)$$

이 확률밀도함수의 모양은 다음과 같이 회전변환된 타원 모양이 된다.

$$P_{xy} = \frac{6xy}{6x6y} = \frac{6}{6} = \frac{-0.6}{\sqrt{(1.4)^2 + (0.4)^2}} \neq 0, < 0$$



• 독립  $\Leftrightarrow$  조건부 분포가 다른 확률변수 간에 따라 달라진다.

020.4.28. 08.06 다변수정규분포

**8.6 다변수정규분포** 스칼라가 아닌 벡터값 출력!  
D차원 다변수정규분포(MVN: multivariate Gaussian normal distribution)의 확률밀도함수는 평균벡터  $\mu$  와 공분산 행렬  $\Sigma$ 라는 두 개의 모수를 가지며 다음과 같은 수식으로 정의한다.

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right) = \text{스칼라} \quad (8.6.1)$$

이 식에서 각 기호의 의미는 다음과 같다.

- $x \in \mathbb{R}^D$  확률변수벡터
- $\mu \in \mathbb{R}^D$  평균벡터
- $\Sigma \in \mathbb{R}^{D \times D}$  공분산행렬  $\Rightarrow$  양의정부호 + 대칭행렬  $\Rightarrow$  모든 양의 고윳값 + 대각화 가능

다면수정규분포에서 공분산행렬은 양의 정부호인 대칭행렬이어야 한다. 따라서 역행렬이 항상 존재한다. 공분산행렬의 역행렬  $\Sigma^{-1}$ 을 정밀도행렬(precision matrix)이라고 한다.

정밀도  $\Delta = \Sigma^{-1}$  (공분산행렬)  
 $\Delta = \frac{1}{\sigma^2}$  (분산)

지의 분포가  
기여 따라  
다르지 않다.

지의 분포가  
기여 따라  
다르다.

08.06

## 다면수 정규분포

## 1) 다변수정규분포와 고유값 분해

- 공분산 행렬의 대각화  $\rightarrow$  공분산 행렬의 고유값행렬, 고유벡터행렬을 얻는다.
- 다변수정규분포의 pdf == 변수 간 좌표변환 (고유벡터 축으로의 회전 및 평행)

## 이동) 역할

- 다변수정규분포 pdf를 취하면,

다면수 공분산행렬이 갖는 고유값이 pdf 타원의 폭, 고유벡터가 방향이 된다.

- 공분산 행렬의 고유벡터 방향으로 고유값만큼 늘리고, 평균벡터만큼 평행이동

- 고유값이 서로 다르다면, pdf는 타원형! (스케일링이 서로 다르니까)

다면수정규분포의 공분산행렬  $\Sigma$ 은 양의 정부호인 대칭행렬이므로 대각화가능(diagonalizable)이다. 정밀도행렬  $\Sigma^{-1}$ 은 다음과처럼 분해할 수 있다. 이 식에서  $\Lambda$ 는 고윳값행렬,  $V$ 는 고유벡터행렬이다.

$$\Sigma^{-1} = V \Lambda^{-1} V^T \quad (8.6.12)$$

이를 이용하면 확률밀도함수는 다음처럼 좌표 변환할 수 있다.

$$\mathcal{N}(x) \propto \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right) \quad (8.6.13)$$

$\begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix} = \exp\left(-\frac{1}{2}(V^T(x - \mu))^T \Lambda^{-1} (V^T(x - \mu))\right)$

$\begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix} = \exp\left(-\frac{1}{2}(V^{-1}(x - \mu))^T \Lambda^{-1} (V^{-1}(x - \mu))\right)$

이 식에서  $\begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix} = \begin{bmatrix} x'_1 & x'_2 & \dots & x'_n \end{bmatrix}$   $x' = V^{-1}(x - \mu)$   $\therefore$   $x = Vx' + \mu$   $\therefore$   $\mathcal{N}(x) \propto \exp\left(-\frac{1}{2}(x - \mu)^T \Lambda^{-1} (x - \mu)\right)$  (8.6.14)

라고 하자. 이 식은 변환행렬  $V^{-1}$ 의 열벡터인 고유벡터를 새로운 축으로 회전시킨 고유벡터 방향으로 평행이동하는 것을 뜻한다.

최종 확률밀도함수식은 다음과 같다. 이 식에서  $\lambda_i$ 는 고윳값  $\Lambda$ 를 대각성분으로 가지는 대각행렬이므로 새로운 좌표  $x'$ 에서 확률밀도함수는 타원이 된다. 타원의 반지름은 고윳값 크기에 비례된다. 반대로 이기기하면 원래 좌표에서 확률밀도함수는  $\mu$ 를 중심으로 가지고 고윳값에 비례하는 반지름을 가진 타원을 고유벡터 방향으로 회전시킨 모양이다.

$$\mathcal{N}(x) \propto \exp\left(-\frac{1}{2}x'^T \Lambda^{-1} x'\right) \quad (8.6.15)$$

예를 들어 위의 두번째 예제에서 공분산행렬을 고유분해한 다음과 같다.

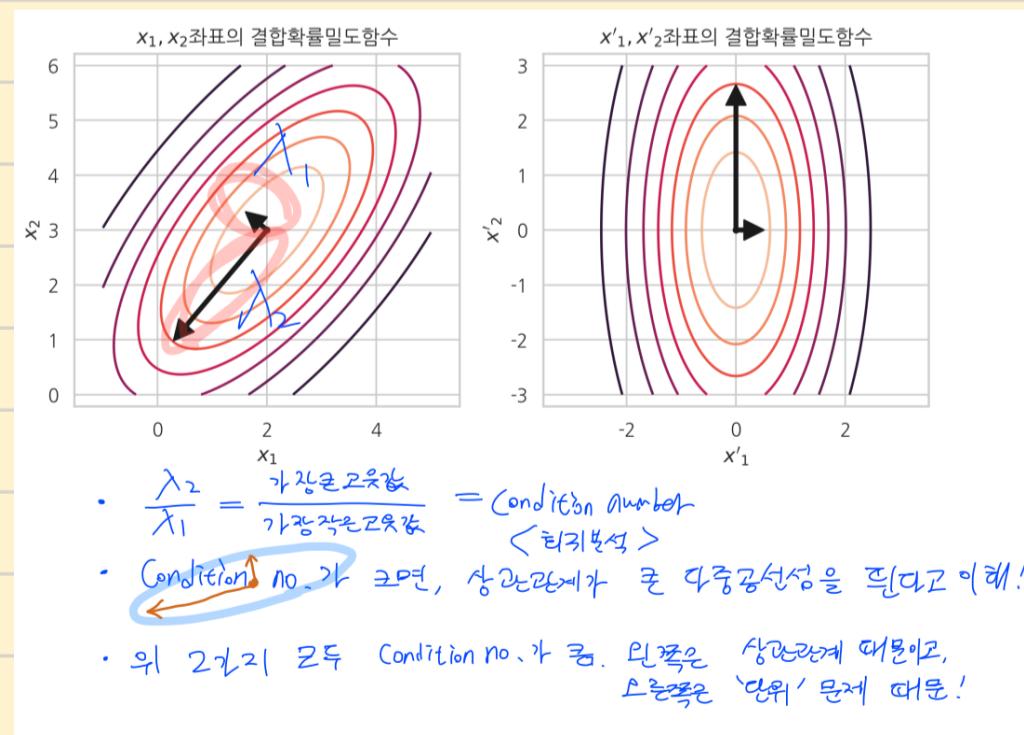
In [3]:

```
mu = [2, 3]
cov = [[4, 3], [3, 5]]
w, v = np.linalg.eig(cov)
고윳값은 lambda_1 = 1.46, lambda_2 = 7.54다.
```

다면수 정규분포  
 $\downarrow$   
 $\mu$  중심  
 $\Rightarrow$  고유값 비례하는  
 $\Rightarrow$  타원은  
 $\Rightarrow$  고윳벡터 방향으로  
 $\Rightarrow$  회전시킨 모양

다면수 정규분포  
 $\downarrow$   
 $\mu$  중심  
 $\Rightarrow$  고윳값 비례하는  
 $\Rightarrow$  타원은  
 $\Rightarrow$  고윳벡터 방향으로  
 $\Rightarrow$  회전시킨 모양

- 다변수정규분포 pdf 타원 모양 + 고유값 ==> 다중공선성 관련 (8page)



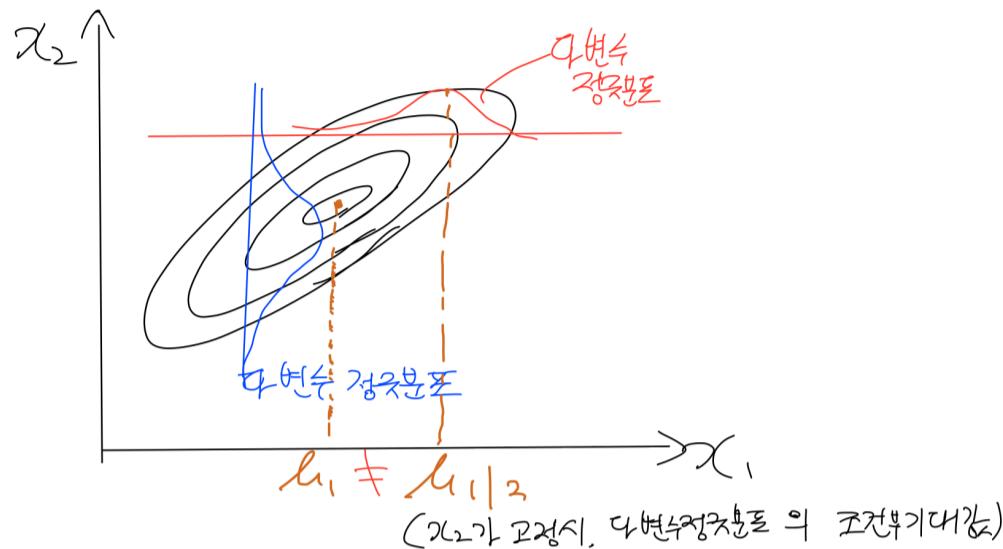
08.06

## 다면수 정규분포

### 2) 다변수 정규분포의 조건부 확률분포 (자른 단면)

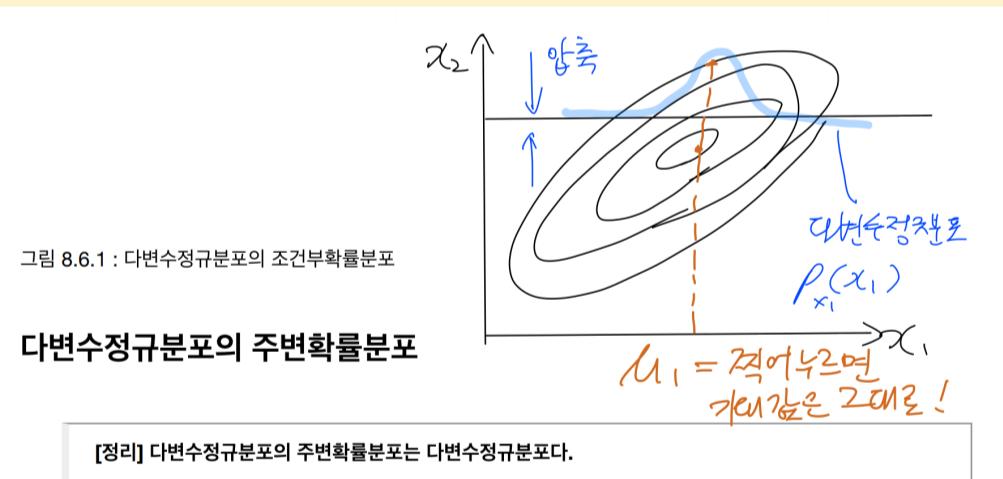
- 조건부 확률분포 (어떤 변수의 값이 정해져 있을 때) 도 역시 다변수 정규분포 !
- 즉, 다변수 정규분포 확률밀도함수 (타원)를 자른 단면도 다변수 정규분포 !

#### 다면수정규분포의 조건부확률분포



### 3) 다변수 정규분포의 주변확률분포 (찌부시킨 것)

- 다변수 정규분포의 주변확률분포도 다변수정규분포



08.07

베타분포

조정이 가능

&gt; 이런 모양의 분포를 만들고 싶다 -&gt; 모수를 조정해 인공적으로 만들 수 있는 분포

감마분포

&gt; 분포의 표현 보다는, 추정한 모수의 신뢰의 정도를 표현하는 데 사용됨(베이지안 관

디리클레분포

점)

## 1) 베타분포 (0 &lt; &lt; 1인 모수 추정에 활용)

- ex) 베르누이분포 뮤 모수 추정

- 0 &lt; 표본공간 &lt; 1 == 추정 대상 모수

- 하이퍼모수 a, b

- 확률변수 확률값 제한 조건 == 추정 대상 모수 제한조건(0 &lt; &lt; 1, sum=1)

- 1page, 2page

**베타분포(Beta distribution)**은  $a$ 와  $b$ 라는 두 모수를 가지며 표본 공간은 0과 1사이의 실수다. 즉 0과 1 사이의 표본값만 가질 수 있다.

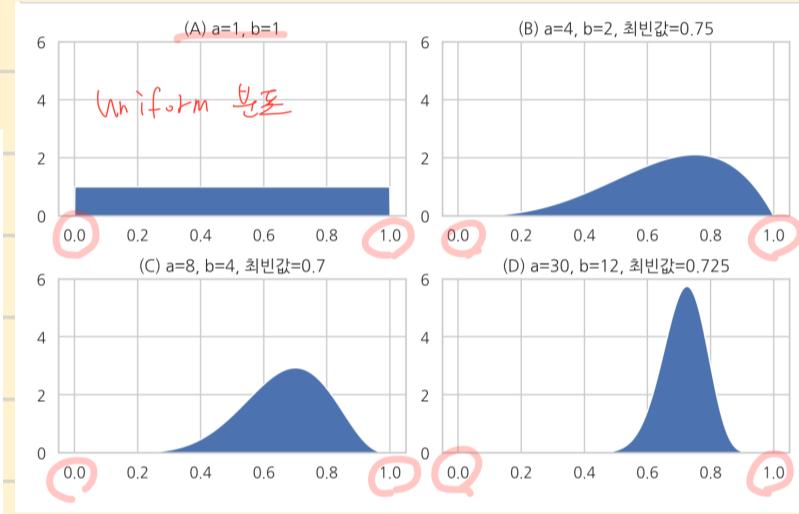
$$\text{Beta}(x; a, b), 0 \leq x \leq 1 \quad (8.7.1)$$

베타분포의 확률밀도함수는 다음과 같다.

$$\text{Beta}(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} \quad (8.7.2)$$

이 식에서  $\Gamma(a)$ 는 감마함수(Gamma function)라는 특수함수로 다음처럼 정의된다.

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx \quad (8.7.3)$$

여러 모수  $a, b$ 값에 대해 베타분포의 모양을 그려보면 다음과 같다.- 베타분포의 모멘트는 하이퍼모수  $a, b$ 의 식 (3page)(분산은  $a$ 와  $b$ 가 커질수록 작아지는 구조 -> 확률분포의 폭이 작아짐)

• 기댓값

$$E[x] = \frac{a}{a+b} \quad \text{Moment} \Leftarrow \text{모두 } a, b \text{의 식!} \quad (8.7.4)$$

• 최빈값 : 확률분포가 가장 커지는 위치

$$\text{mode} = \frac{a-1}{a+b-2} \quad (\text{설계} \rightarrow \text{최대하는 mode}, \text{var(x) 값 정하고, 그에 맞춰 } a, b \text{를 구한다.}) \quad (8.7.5)$$

• 분산 : 확률분포의 폭

$$\text{Var}[x] = \frac{ab}{(a+b)^2(a+b+1)} \quad \begin{matrix} 2차식 \\ 3차식 \end{matrix} \quad (8.7.6)$$

최빈값 수식을 보면  $a = b$ 일 때  $x = 0.5$ 에서 가장 확률밀도가 커지는 것을 알 수 있다. 또한 분산 수식에서 분모가 3차식, 분자가 2차식이기 때문에  $a, b$ 의 값이 커질수록 분산 즉, 확률분포의 폭이 작아진다.

## 2) 베타분포와 베이지안 추정

- '베이지안' 추정 : 추정(모수)을 확정값이 아닌, 가능성으로 확률분포로 나타낸다.

3page)

## 베타분포와 베이지안 추정

(베르누이 분포, 모두 시도정)



$$\hat{\mu} = \frac{x_i}{N} \neq \mu$$

 $\hat{\mu} = [a=0, b=0 \text{인 Beta분포}]$ 

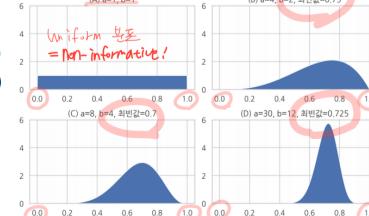
모수추정 결과를 그림으로 표현!  
So, 확률분포의 폭은 가능성을 A로 바꿔놓는다.

베터분포는 0부터 1까지의 값을 가질 수 있는 베르누이분포의 모수  $\mu$ 의 값을 베이지안 추정한 결과를 표현한 것이다. 베이지안 추정은 모수가 가질 수 있는 모든 값에 대해 가능성을 확률분포로 나타낸 것을 말한다. 실제로 베르누이분포의 모수를 베이지안 추정하는 것은 나중에 다루게 된다. 여기에서는 결과만 보였다.

위 그림이 베이지안 추정 결과라면 각각은 베르누이분포의 모수  $\mu$ 에 대해 다음과 같이 추정한 것과 같다.

- (A): 베르누이분포의 모수  $\mu$ 를 추정할 수 없다. (정보가 없음)  $a=b=1$  non informative
- (B): 베르누이분포의 모수  $\mu$ 값이 0.75일 가능성이 가장 크다. (정확도 낮음)
- (C): 베르누이분포의 모수  $\mu$ 값이 0.70일 가능성이 가장 크다. (정확도 중간)
- (D): 베르누이분포의 모수  $\mu$ 이 0.725일 가능성이 가장 크다. (정확도 높음)

연습 문제 8.7.1



3) 감마분포 ( $0 < x < \infty$ 인 모수 추정에 활용)

베타분포

- ex) 정규분포의 분산 모수 추정

감마분포

-  $0 < x < \infty$  == 추정 대상 모수

디리클레분포

- 하이퍼모수  $a, b$ 

- 4page

감마분포(Gamma distribution)도 베타분포처럼 모수의 베이지안 추정에 사용된다. 다만 베타분포가 0부터 1 사이값을 가지는 모수를 베이지안 방법으로 추정하는 데 사용되는 것과 달리 감마분포는 0부터 무한대의 값을 가지는 양수 값을 추정하는데 사용된다.

*ex)  $\mathcal{N}(\mu, \sigma^2)$ 의  $\sigma^2$ 을 추정하고 싶다  $\Rightarrow \beta$  분포가 아닌  $\Gamma$  분포!*

감마분포의 확률 밀도 함수는  $a$ 와  $b$ 라는 두 모수를 가지며 수학적으로 다음과 같이 정의된다.

$$\text{Gam}(x; a, b) = \frac{1}{\Gamma(a)} b^a x^{a-1} e^{-bx} \quad (8.7.7)$$

감마분포의 확률 밀도 함수는 모수  $a, b$ 의 값에 따라 다음과 같은 형상을 가진다.

사이파이의 stats 서브패키지에서 제공하는 `gamma` 클래스는 모수  $b = 1$ 로 고정되어  $a$  값만 설정할 수 있다.  $b$ 를 바꾸려면  $x$ 값 스케일과 계수를 수동으로 설정하여야 한다.

4) 디리클레 분포 ( $0 < x_i < 1$ 인 모수 추정. 단,  $k$ 개 클래스의 모수 추정)

- ex) 카테고리분포 뮤벡터 모수 추정

-  $0 < x_i < 1$  == 추정 대상 모수

- 하이퍼모수 (알파1 ~ 알파k)

- 확률변수 확률값 제한 조건 == 추정 대상 모수 제한조건( $0 < x_i < 1, \sum x_i = 1$ )

- 5page

둘다 0과 1사이 모수추정  $\text{Var}[X] = \frac{1}{b^2}$  (8.7.10)

디리클레분포  $\text{Dir}(K, \alpha_1, \alpha_2, \dots, \alpha_K)$   $\leftarrow$  디리클레 분포! (K클래스) ( $0 < \alpha_i < 1, \sum \alpha_i = 1$ )

디리클레분포(Dirichlet distribution)는 베타분포의 확장판이라고 할 수 있다. 베타분포는 0과 1사이의 값을 가지는 단일(univariate) 확률변수의 베이지안 모형에 사용되고 디리클레분포는 0과 1사이의 사이의 값을 가지는 다변수(multivariate) 확률변수의 베이지안 모형에 사용된다.

예를 들어  $K = 3$ 인 디리클레분포를 따르는 확률변수는 다음과 같은 값들을 표본으로 가질 수 있다.

$$(0.2, 0.3, 0.5) \rightarrow \text{모두 } \alpha_i \text{ 가 될 수 있다.} \quad (8.7.11)$$

$$(0.5, 0.5, 0) \rightarrow \text{모두 } \alpha_i \text{ 가 될 수 있다.} \quad (8.7.12)$$

$$(1, 0, 0) \rightarrow \text{모두 } \alpha_i \text{ 가 될 수 있다.} \quad (8.7.13)$$

디리클레분포의 확률밀도함수는 다음과 같다.

$$\text{Dir}(x; \alpha) = \text{Dir}(x_1, x_2, \dots, x_K, \alpha_1, \alpha_2, \dots, \alpha_K)$$

$$= \frac{1}{B(\alpha_1, \alpha_2, \dots, \alpha_K)} \prod_{i=1}^K x_i^{\alpha_i - 1} \quad (8.7.14)$$

이 식에서  $x = (x_1, x_2, \dots, x_K)$ 은 디리클레분포의 표본값 벡터이고  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_K)$ 은 모수 벡터다.

$B(\alpha_1, \alpha_2, \dots, \alpha_K)$ 는 베타함수라는 특수함수로 다음처럼 정의한다.

$$B(\alpha_1, \alpha_2, \dots, \alpha_K) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)} \quad (8.7.15)$$

디리클레분포의 확률값  $x$ 는 다음 제한조건을 따른다.

$$= \text{Dir 확률변수의 출현값} \quad 0 \leq x_i \leq 1, \quad \sum_{i=1}^K x_i = 1 \quad \Rightarrow \text{(\alpha_i의 모수가 } (M_1, \dots, M_K) \text{ 드어야하기 때문!)}$$

08.07

베타분포

감마분포

디리클레분포

## 5) 베타분포와 디리클레분포의 관계

-  $k=2$  인 디리클레분포(cat 모수추정) == 베타분포(베르누이 모수추정)

## 6) 디리클레분포의 응용

-  $x+y+z = 1$  인, 양의 난수  $x, y, z$  생성. 모든 경우가 균등하게!

<==> 카테고리 확률변수  $x, y, z$ 이고, 대신, 그것의 모수 뮤가 모두 균등하게!

<==> 디리클레 분포를 통해 균등한 뮤를 추정하자!

<==> 클래스 3인, 하이퍼모수가 모두 1인 디리클레분포로 모수 추정!

### 디리클레분포의 응용

다음과 같은 문제를 풀어보자.

$x, y, z$ 가 양의 난수일 때 항상  $x + y + z = 1$ 이 되게 하려면 어떻게 해야될까요? 모든 경우가 균등하게 나와야 합니다.

이 문제는  $K = 3$ 이고  $\alpha_1 = \alpha_2 = \alpha_3$ 인 디리클레분포의 특수한 경우이다.

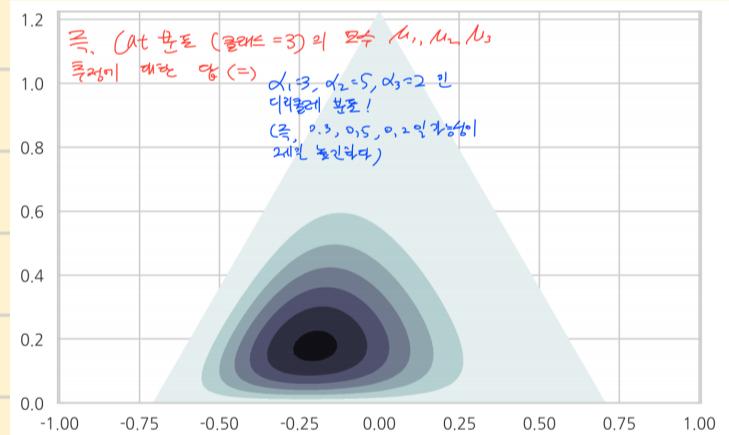
( $\hat{\mu}_1 = \hat{\mu}_2 = \hat{\mu}_3$ )

Non-informative는 디리클레분포!

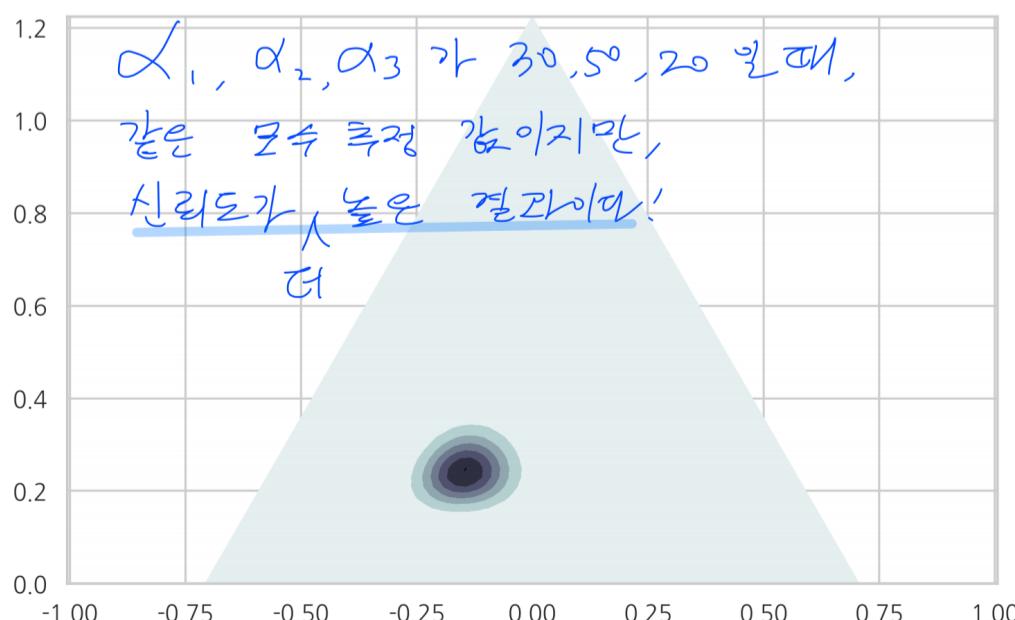
## 7) 디리클레분포와 베이지안 추정

- 디리클레분포를 통해, 카테고리 분포의 모수를 추정할 수 있다.(베이지안 추정)

(13page 하단, 16page, 17page)



모수  $\alpha$ 가  $(1, 1, 1)$ 이 아닌 경우에는 다음과 같이 특정 위치에 분포가 집중되도록 할 수 있다. 이 특성을 이용하면 카테고리분포의 모수 벡터  $\mu$ 를 추정한 결과를 나타낼 수 있다.



09.01

확률분포의 추정  
(Shape Factor  
추정)

09.01 확률분포의 추정

= 모수(모멘트)를 추정하는 것

### 1) 확률분포의 결정

- 1 : 확률변수가 우리가 배운 기본분포 중 어느 확률분포를 따르는지 탐색
- 2 : 데이터로부터 확률분포의 모수 값을 구한다. (= 모수추정)

### 1 : 1page

확률분포를 알아내는 일은 다음처럼 두 작업으로 나뉜다. ✕

1. 확률변수가 우리가 배운 베르누이분포, 이항분포, 정규분포 등의 기본 분포 중 어떤 확률분포를 따르는지 알아낸다.
2. 데이터로부터 해당 확률분포의 모수의 값을 구한다. *모수추정!*

첫 번째 작업 즉, 확률변수가 어떤 확률분포를 따르는가는 데이터가 생성되는 원리를 알거나 데이터의 특성을 알면 추측할 수 있다. 히스토그램을 그려서 확률분포의 모양을 살펴보고 힌트를 얻을 수도 있다.

- 데이터는 0 또는 1 뿐이다. → 베르누이분포
- 데이터는 카테고리 값이어야 한다. → 카테고리분포
- 데이터는 0과 1 사이의 실수 값이어야 한다. → 베타분포
- 데이터는 항상 0 또는 양수이어야 한다. → 로그정규분포, 감마분포, F분포, 카이제곱분포, 지수분포, 하프코시분포 등
- 데이터가 크기 제한이 없는 실수다. → 정규분포 또는 스튜던트 t분포, 코시분포, 라플라스분포 등

이 규칙에는 예외가 있을 수 있다. 예를 들어 항상 양수인 데이터인 경우에도 정규분포로 모형화가 가능하다면 정규분포를 사용할 수 있다 정규분포와 스튜던트 t분포와 같이 둘 중 어느 것인지 구분하기 힘든 경우에는 뒤에서 설명할 정규성 검정이나 KS검정을 사용한다.

### 2) 모수 추정 방법론

- 1) 모멘트 방법
- 2) 최대가능도 추정법
- 3) 베이지안 추정법

09.01

확률분포의 추정  
(Shape Factor  
추정)

## 1) 모멘트 방법

: 데이터값(표본)의 모멘트 == 이상(확률변수)의 모멘트 가정(2page 상단)

ex) 베르누이분포의 모수 추정 (모멘트 방법)

베르누이분포 모수 뮤 = \*베르누이분포 기대값 = 표본의 기대값(표본평균)\*

| $X$ 의 pdf | data      |
|-----------|-----------|
| $E(X)$    | $\bar{x}$ |
| 분산        | 표본분산      |
| skewness  | 표본분산      |
| kurtosis  | 표본분산      |

ex) 정규분포의 모수 추정 (모멘트 방법)

정규분포 모수 뮤 = \*정규분포 기대값 = 표본의 기대값(표본평균)\*

정규분포 모수 분산 = \*정규분포 분산 = 표본분산\*

ex) 베타분포의 모수 추정 (모멘트 방법)

베타분포 모수 a, b = \*베타분포의 기대값, 분산 공식

= 표본기대값, 표본분산\*

\*\*연립방정식

\*Kernel Density Estimation (5page)

= Non-parametric 추정 방법

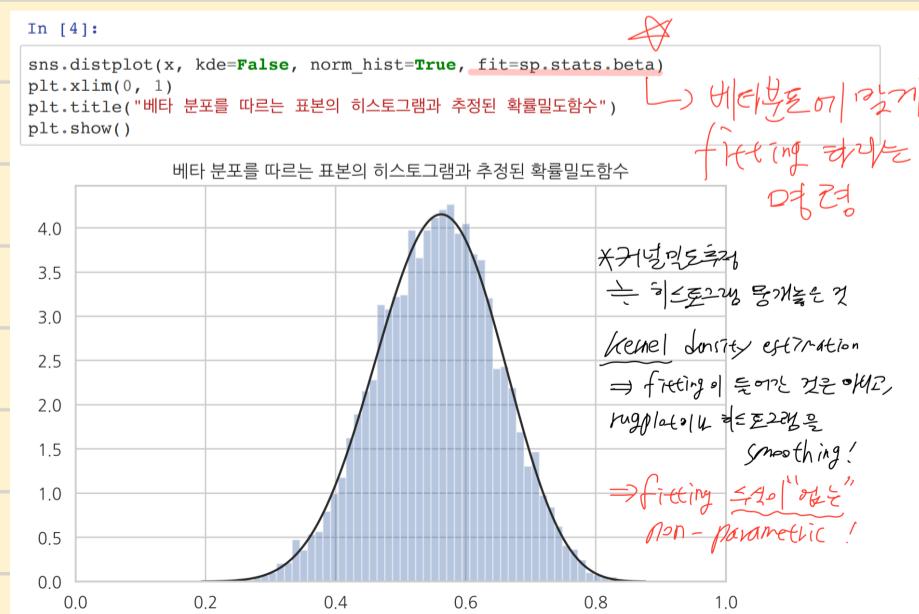
1) parametric 추정 방법 : fitting 수식을 통해 모수 추정 -> 분포를 추정하는 것

\*seaborn에 특정 분포에 맞게 모수를 추정해 분포를 만들도록 하는 명령

= .distplot(fit= )

2) non-parametric 추정 : fitting 수식 없이 주어진 히스토그램을 smoothing.

하게 만듬



09.02

최대가능도 추정법

MLE

Maximum

Likelihood

Estimation

- 모멘트 방법 : 가정. "표본모멘트 == 확률분포의 이론적 모멘트"

- MLE : 가정이 아닌, "가장 가능성 높게" 확률분포의 모수를 추정하고자 한다.

Q. "가장 가능성 높다"는 어떻게 정의하는가?

### 1) 가능도함수 (likelihood function)

- 가능도함수는 reverse engineering!

- 기존의 관점(확정된 모수로 확률분포를 그려 확률값(표본값) 계산), 변수 =  $x$ (다면수가 되기도 함)

- 가능도 관점(확정된 확률값(데이터)로 모수를 추정  $\rightarrow$  확률분포 추정), 변수 = 모수(다면수가 되기도 함)

- 1 - 3page

이제부터는 여러가지 확률분포  $X$ 에 대한 확률밀도함수 또는 확률질량함수를 다음과 같이 대표하여 쓰기로 한다.

$$\text{Pf} \leftarrow p(x; \theta) \rightarrow \text{모수 } (\mu, \sigma^2, \alpha, \dots) = \text{상수}$$

증명, 확률변수 확률값 = 변수

확률밀도함수에서는 모수  $\theta$ 가 이미 알고 있는 상수계수고  $x$ 가 변수다. 하지만 모수 추정 문제에서는  $x$  즉, 이미 실현된 표본값은 알고 있지만 모수  $\theta$ 를 모르고 있다. 이때는 반대로  $x$ 를 이미 알고 있는 상수계수로 놓고  $\theta$ 를 변수로 생각한다. 물론 함수의 값 자체는 변함없이 주어진  $x$ 가 나올 수 있는 확률밀도다. 이렇게 확률밀도함수에서 모수를 변수로 보는 경우에 이 함수를 가능도함수 (likelihood function)라고 한다. 같은 함수를 확률밀도함수로 보면  $p(x; \theta)$ 로 표기하지만 가능도함수로 보면  $L(\theta; x)$  기호로 표기한다.

모수추정문제 : "3.2번은 데이터 ( $x_1, \dots, x_n$ )가 나왔는데, 이를 나누기 한 분포의  $\theta$ (모수)는 무엇?"

$L(\theta; x) = p(x; \theta)$

모수추정시 기준 방식 = 알고 있는 모수로 shape를 설정하고, 거기서 학습변수 확률값들을 (reverse engineering)

모수 = 상수계수! ( $\mu_0, \sigma^2$ )

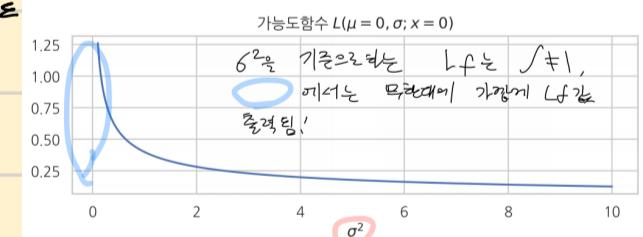
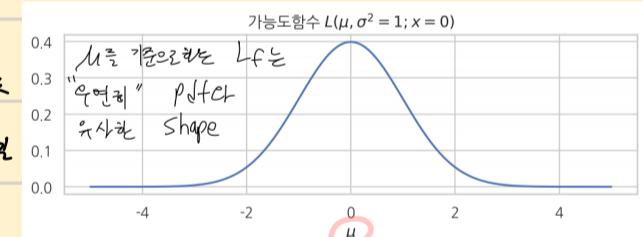
#### [ Pf 와 Lf 비교 ]

공통점 :  $p(x; \mu, \sigma^2)$  출력값 =  $L(\mu, \sigma^2; x)$  출력값 = 확률밀도값

차이점 :  $p(x; \mu, \sigma^2)$ 의 경우,  $x$ 가 변수,  $x$ 를 기준으로 Pf가 결정되는 때는  $\int p(x) dx = 1$

$L(\mu, \sigma^2; x)$ 의 경우,  $\mu, \sigma^2$ 가 변수,  $\mu$ 나  $\sigma^2$ 를 기준으로 가능도함수를 적용하면, '1'이 된다는 '보장'이 없음

즉,  $L(\mu, \sigma^2; x)$ 는 확률밀도함수가 아니다. 정의 충족X ( $\int \neq 1$ )



- 변수가 달라지면, 그래프가 달라진다.

ex) 베르누이 분포 pdf, Lf (4page)

#### 예제

베르누이분포의 확률질량함수는 다음과 같은 함수다. 이때 입력  $x$ 는 0과 1이라는 두 가지 값만 받을 수 있다.

$$p(x; \mu_0) = \mu_0^x (1 - \mu_0)^{1-x}$$

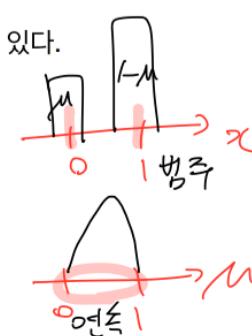
변수 상수

하지만 가능도함수는 다음과 0부터 1까지의 연속적인 실수값을 입력으로 받는 함수가 된다.

$$L(\mu; x_0) = \mu^{x_0} (1 - \mu)^{1-x_0}$$

변수 상수

수식은 같지만 함수의 변수가 다르다는 점에 주의하라.



## - MLE 정리 (5page)

09.02

최대가능도 추정법

MLE

Maximum

Likelihood

Estimation

- 확률밀도함수  $f(x; \theta)$ 
  - $\theta$  값을 이미 알고 있음
  - $\theta$ 는 상수,  $x$ 는 변수
  - $\theta$ 가 이미 정해져 있는 상황에서의  $x$  값의 상대적 확률
  - 적분하면 전체 면적은 항상 1
- 가능도함수  $L(\theta) = p(x | \theta)$ 
  - $x$ 가 이미 발생. 값을 이미 알고 있음
  - $x$ 는 상수,  $\theta$ 는 변수
  - $x$ 가 이미 정해져 있는 상황에서의  $\theta$  값의 상대적 확률
  - 적분하면 전체 면적이 1이 아닐 수 있다.

## 2) 최대가능도 추정법 (MLE : Maximum Likelihood Estimation)

: 최적화. 변수인 모수를 이것저것 trial해봐서 Lf 출력값(확률값)이 가장 높게 나오는 것을 찾는

(5page 중, 하단, 6page)

### 최대가능도 추정법

최대가능도 추정법(Maximum Likelihood Estimation, MLE)은 주어진 표본에 대해 가능도를 가장 크게 하는 모수  $\theta$ 를 찾는 방법이다. 이 방법으로 찾은 모수는 기호로  $\hat{\theta}_{MLE}$ 와 같이 표시한다.

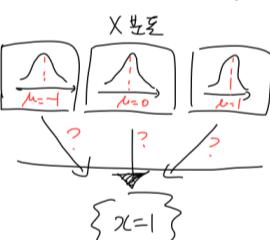
$$\hat{\theta}_{MLE} = \arg \max_{\theta} L(\theta; x)$$

$L(\theta; x)$  가능도  
 $= P(x=x_0 | \theta=1)$  비교해서  
 $P(x=x_0 | \theta=0)$  가장 큰 값을 주는  $\theta$  찾기!

### 예제

정규분포를 가지는 확률변수의 분산  $\sigma^2 = 1$ 은 알고 있으나 평균  $\mu$ 를 모르고 있어 이를 추정해야 하는 문제를 생각해보자. 확률변수의 표본은 하나  $x_1 = 1$ 을 가지고 있다고 하자. 이 경우 어떤  $\mu$  값이 가장 가능성(가능도)이 커 보이는가? 다음 그림에는  $\mu = -1, \mu = 0, \mu = 1$ , 세 가지 후보를 제시한다. 이 세 가지  $\mu$  값에 대해 1이 나올 확률밀도의 값이 바로 가능도다.

① 데이터는 확률변수라는 특성에서 뛰어난다.  
 ② but,  $\theta=1 (\mu=1)$  인 분포에서 뛰어난 예상인지,

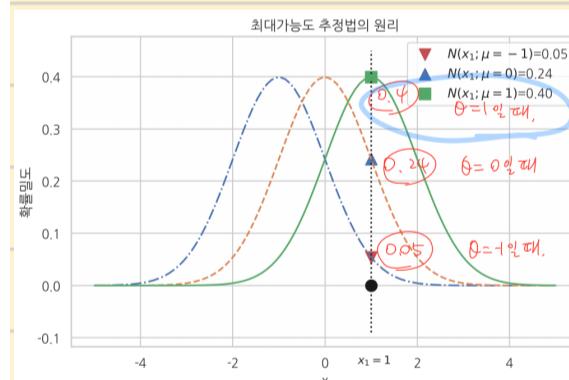


③ Likelihood 정근은,

$$P(x=1 | \theta=1) P(x=0 | \theta=0) P(x=-1 | \theta=-1) \Rightarrow 각각 계산해서$$

어느 모수( $\theta$ )로 분포를 만들었을 때  $x=1$ 의 가능성(가능도)이 가장 높을까?  
 $\therefore \max_{\theta} P(x=1 | \theta)$

$= \max_{\theta} \max_{\text{변수}} L(\theta; x_0)$   
 (설명: 확률밀도)



## 3) 복수의 표본데이터가 있는 경우, 가능도함수

: 반복시행 = 독립

: 같은 확률변수에서 표본을 1개가 아닌, 여러개 뽑으면, 표본들은 서로 독립

!(분리가능함수화) (7page 상단+하단 예제)

### 복수의 표본 데이터가 있는 경우의 가능도함수

일반적으로는 추정을 위해 확보하고 있는 확률변수 표본의 수가 하나가 아니라 복수 개  $\{x_1, x_2, \dots, x_N\}$ 이므로 가능도함수도 복수 표본값에 대한 결합확률밀도  $p_{X_1 X_2 \dots X_N}(x_1, x_2, \dots, x_N; \theta)$ 가 된다. 표본 데이터  $x_1, x_2, \dots, x_N$ 은 같은 확률분포에서 나온 독립적인 값들인므로 결합 확률밀도함수는 다음처럼 곱으로 표현된다.

$$L(\theta; x_1, \dots, x_N) = p(x_1, \dots, x_N; \theta) = \prod_{i=1}^N p(x_i; \theta)$$

연속

정규분포로부터 다음 세 개의 표본 데이터를 얻었다.

{1, 0, -3}

이 경우의 가능도함수는 다음과 같다.

$$\begin{aligned}
 L(\theta; x_1, x_2, x_3) \\
 = N(x_1, x_2, x_3; \theta) \\
 = N(x_1; \theta) \cdot N(x_2; \theta) \cdot N(x_3; \theta)
 \end{aligned}$$

09.02

- 정규분포, 베르누이분포의 모수추정 MLE 법 예제 직접 풀어보기

최대가능도 추정법

MLE

Maximum

Likelihood

Estimation

#### 4) 로그가능도함수

- 가능도함수를 로그를 취해 사용하는 경우가 대다수 (9apge 중단)

이유 1 : 로그 변환해도 최대값의 위치는 변하지 않음

이유 2 : 반복시행 -> 독립 -> pdf, 가능도함수가 곱셈으로 분리가능 -> 로그 취할 시 덧셈으로 계산 편의성

일반적으로 최대가능도 추정법을 사용하여 가능도가 최대가 되는  $\theta$ 를 계산해려면 수치적 최적화(numerical optimization)를 해야 한다.

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} L(\theta; \{x_i\})$$

그런데 보통은 가능도를 직접 사용하는 것이 아니라 로그 변환한 로그가능도함수  $LL = \log L$ 을 사용하는 경우가 많다.

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} \log L(\theta; \{x_i\})$$

#### 5) 베르누이분포의 최대가능도 모수 추정 (11p)

최대가능도 추정법에 의한 베르누이분포의 모수는 1이 나온 횟수와 전체 시행횟수의 비율이다.

#### 6) 카테고리분포의 최대가능도 모수 추정 (13p)

최대가능도 추정법에 의한 카테고리분포의 모수는 각 범주값이 나온 횟수와 전체 시행횟수의 비율이다.

#### 7) 정규분포의 최대가능도 모수 추정 (14p)

최대가능도 추정법에 의한 정규분포의 기댓값은 표본평균과 같고 분산은 (편향)표본분산과 같다.

#### 8) 다변수정규분포의 최대가능도 모수 추정 (16p)

최대가능도 추정법에 의한 다변수정규분포의 기댓값은 표본평균벡터와 같고 분산은 표본공분산행렬과 같다.

#### 9) 베르누이, 카테고리, 정규, 다변수 정규 모두 모멘트 추정법 결과와 같다.

- 모멘트, MLE와 달리 추정값(모수값)이 가질 수 있는 모든 가능성의 분포를 추정

09.03

베이즈 추정법

결과로 제시

- 베이즈 추정법 사용 이유 : 추정 결과가 상수라면, 추정의 신뢰도와 신뢰구간을 구할 수 없기 때문.

## 1) 베이즈 추정법의 기본원리 (1p 하단)

- 베이즈 추정 : 베이즈 정리를 활용한 모수 추정

- 추정 결과 : 확률분포로 제시 (ex] 추정결과 = 베타(1,1))

- 사전분포(prior)를 모를 때, non-informative 분포로 대체

non-infor분포 :  $a=b=1$  베타분포, 알파 $\alpha=1$  디리클레, 0을 기대값으로 하는 정규분포

### 베이즈 추정법의 기본 원리

모수 (=변수)

수학적으로 베이즈 추정법은 주어진 데이터  $\{x_1, \dots, x_N\}$ 를 기반으로 모수  $\mu$ 의 조건부 확률분포  $p(\mu|x_1, \dots, x_N)$ 을 계산하는 작업이다. 조건부 확률분포를 구하므로 베이즈 정리를 사용한다.

$$p(\mu | x_1, \dots, x_N) = \frac{p(x_1, \dots, x_N | \mu) \cdot p(\mu)}{p(x_1, \dots, x_N)} \propto p(x_1, \dots, x_N | \mu) \cdot p(\mu) \quad (9.3.1)$$

- $p(\mu)$ 는 모수의 사전(Prior)분포다. 사전 분포는 베이지안 추정 작업을 하기 전에 이미 알고 있던 모수  $\mu$ 의 분포를 뜻한다. 모수의 분포에 대해 아무런 지식이 없는 경우에는 균일(uniform) 분포 Beta(1, 1)나 0을 중심으로 가지는 정규분포  $\mathcal{N}(0, \sigma_0^2)$  등의 무정보분포(non-informative distribution)를 사용할 수 있다. 무정보 분포에 대해서는 다음 장에서 공부한다.
- $p(\mu | x_1, \dots, x_N)$ 은 모수의 사후(Posterior)분포다. 수학적으로는 데이터  $x_1, \dots, x_N$  가 주어진 상태에서의  $\mu$ 에 대한 조건부 확률 분포다. 우리가 베이즈 추정법 작업을 통해 구하고자 하는 것이 바로 이 사후 분포다.
- $p(x_1, \dots, x_N | \mu)$ 은 가능도(likelihood)분포다. 모수  $\mu$ 가 특정한 값으로 주어졌을 때 주어진 데이터  $\{x_1, \dots, x_N\}$  가 나올 수 있는 확률값을 나타낸다.

$L(\mu; x_1, \dots, x_N) \Rightarrow \mu$ 는 조정지역이며 어느 shape인 때의  
변수  $\mu$ 의 학률값

\*  $p(x_1, \dots, x_N | \mu) \cdot p(\mu)$   
 $= L(\mu; x_1, \dots, x_N) \cdot p(\mu) \Leftrightarrow$   $\mu$ 에 따른 표본값( $x$ )의 학률값  $\times$   $\mu$  분포  
 $\Leftrightarrow$  분포가  $(\mu)$  shape 일 때,  $\mu$ 가 표본값( $x$ )의 학률  
 $= x$  가 표본값일 때, 분포가  $(\mu)$  shape 일 학률.  
(예상)  
(예상)

localhost:8888/noteconvert/html/Desktop/0 데이터 사이언스 쿠루 수학/교재 → 트론/09.03 베이즈 추정법.ipynb?download=false

1/11

- 베이즈 추정법에서 모수의 추정분포는 2가지 방법으로 표현됨

### 1) 모수적 방법(parametric) = 하이퍼모수 계산

- 하이퍼모수를 통해 확률분포로 추정 결과 제시( ex] 모수 추정 결과 = 베타(1,1) )

- 베이즈 추정법 == 하이퍼모수값 계산 작업

사전분포(베타분포 등, 모른다면 non-infor로) ==> 사후분포(사전분포에 계산된 새 하이퍼모수로 분포 제시)

### 2) 비모수적 방법(non-parametric) = 실제 표본집합만 생성

- 모수의 분포와 동일한 분포를 갖는 실제 표본집합을 생성 -> 히스토그램, 최빈값 등으로 분포를 표현

09.03

## 베이즈 추정법

### 2) 베이즈적, 모수 추정법 1 : 베르누이 분포의 모수 추정

- 사전분포 : 베타분포(non-informative)

- 사후분포 : 베타분포(갱신된 하이퍼모수)

\*켤레 사전분포!

- 모멘트, MLE : 확정된 값

- 베이즈 추정 : 새로운 하이퍼모수를 계산해 그를 바탕으로 하는 베타분포 제시

\*사전분포를 모른다는 가정 하, 무정부분포(베타(1,1))로 prior

- 베르누이 분포의 모수를 베이즈 추정한다 == 새로운 하이퍼모수 계산( $a', b'$ )

하이퍼모수( $a, b \rightarrow a', b'$ )가 갱신되고, 사전분포와 사후분포는 동일한 형태의 pdf가 된다.

\*이처럼, 사후분포와 사전분포가 모수값만 다르고 같은 pdf로 표현될 수 있도록 해주는 경우를

: 켤레 사전확률분포(conjugate prior)라고 한다.

ex) 동전에 대한 정보가 없을 때, 뮤값은 얼마나?

1. 베이즈 추정법에 의해, 사전확률분포 : 베타(1,1),  $a=b=1$

: non-informative, 0부터 1사이 값은 확실한 데, 뭐가 나올진 모르겠다.

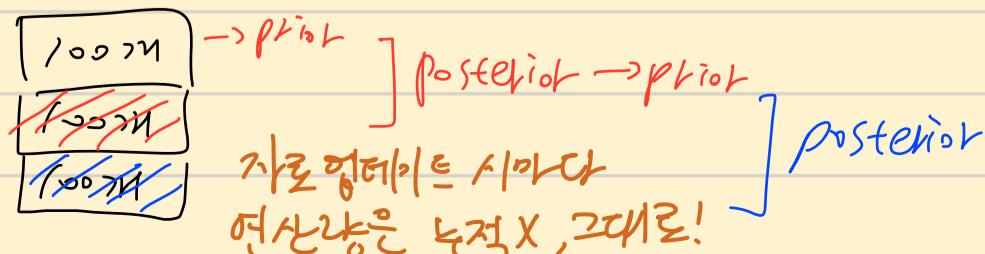
2. 10번 던져서 6번이 나옴 -> 사후적 추정 확률(posterior)분포 제시 가능

: 사후확률분포 : 새로운 하이퍼모수( $a', b'$ )를 갖는 베타분포

\*베이즈 추정법의 장점 : 순차적 연산 가능(데이터 업데이트 시마다, 연산량은 그대로 가능) 3page

베이즈 추정법의 장점은 순차적(sequential) 계산이 가능하다는 점이다. 예를 들어 매 50개의 데이터를 수집하는 경우를 생각하자. 베이즈 추정법을 사용하면 첫날 50개의 데이터로 모수를 추정한 뒤 다음날에는 추가적인 데이터 50개를 사용하여 모수값을 더 정확하게 수정할 수 있다. 이 과정에서 계산량은 증가하지 않는다. 그다음 날도 마찬가지다.

하지만 최대가능도 추정법을 사용하면 첫날에는 데이터 50개를 이용하여 모수를 추정하지만 둘째 날에는 100개의 데이터를 사용하여 모수를 추정해야다. 그다음 날에는 150개의 데이터를 사용하여 계산을 해야 한다. 데이터가 더 수집되면 점점 추정에 사용되는 데이터의 수가 증가하고 그에 따라 계산량도 증가한다.



09.03

## 베이즈 추정법

## 3) 베이즈적, 모수 추정법 2 : 카테고리 분포의 모수 추정

- 사전분포 : 디리클리(non-informative하게)
  - 사후분포 : 디리클리(갱신된 하이퍼모수)
- \*켤레 사전분포!

## 4) 베이즈적, 모수 추정법 3 : 정규분포의 기댓값 모수 추정

- 사전분포 : 정규분포(추정 대상 모수의 분포와 동일한 분포 사용)
  - 사후분포 : 정규분포(갱신된 하이퍼모수)
- \*켤레 사전분포!

## - 추정 결과

이번에는 정규분포의 기댓값 모수를 베이지안 방법으로 추정한다. 분산 모수  $\sigma^2$ 은 알고 있다고 가정한다.

기댓값은  $-\infty$ 부터  $\infty$ 까지의 모든 수가 가능하기 때문에 모수의 사전 분포로는 정규분포를 사용한다.

$$p(\mu) = N(\mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) \quad (9.3.12)$$

데이터는 모두 독립적인 정규분포의 곱이므로 가능도 함수는 다음과 같다.

$$p(x_1, \dots, x_N | \mu) = \prod_{i=1}^N N(x_i | \mu) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \quad (9.3.13)$$

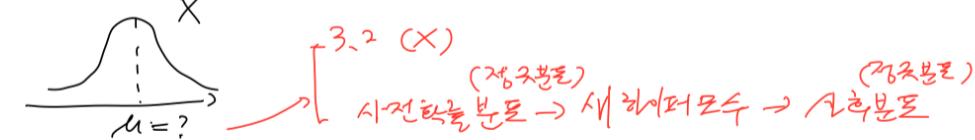
$$\begin{aligned} p(\mu | x_1, \dots, x_N) &\propto p(x_1, \dots, x_N | \mu)p(\mu) \\ &\propto \exp\left(-\frac{(\mu - \mu'_0)^2}{2\sigma'^2_0}\right) \end{aligned} \quad (9.3.14)$$

베이즈 정리를 이용하여 사후 분포를 구하면 다음과 같이 갱신된 하이퍼모수를 가지는 정규분포가 된다.

$$\textcircled{b} \quad \mu'_0 = \frac{\textcircled{a} \sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 + \frac{\textcircled{c} N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \frac{\sum x_i}{N} \quad \text{켤레 사전분포!}$$

$$\textcircled{d} \quad \frac{1}{\sigma'^2_0} = \frac{1}{\sigma^2_0} + \frac{N}{\sigma^2}$$

예제



① prior에서 가져온  $\mu_0$   
 ② posterior로 측정한  $\mu'_0$   
 ③ 새로 들어온 데이터 =  $\frac{\sum x_i}{N}$   
 ④ 가중치,  $\textcircled{d} + \textcircled{e} = 1$

∴ posterior는 prior와 새로 들어온 데이터를  
가중평균!

(예전 것과 새것을 잘 석자)  
 ( $N$ 이↑이면, 새것에 대한 가중치↑)  
 새 데이터 크기

⑤ precision = 고정 precision  
 (새로운) = 더 정밀!  
 + 증수

\* 분수↓ = precision↑  
 (더 정밀)

\*  $N \uparrow \Rightarrow$  precision 증가폭↑  
 (데이터↑수↑) [ 새 데이터의 가중치↑]

09.04

검정과 유의확률

검정 : 데이터 뒤에 숨어있는 확률변수의 분포에 대한 가설이 맞는지, 틀리는지 정량적으로 증명

### 1) 가설과 검정

모수 검정(parameter testing) : 확률분포의 모수값이 특정한 값을 가진다는 가설을 검정

### 2) 귀무가설 (null hypothesis, H0)

- 검정 작업에 있어, 가장 첫, 기준점이 되는 가설
- 귀무가설을 기각할 수 없으면, 귀무가설이 맞다.
- 1차 가설
- 항상 등식이어야 한다. (부등식 가설은 안됨)

> ex) 모수 검정

$$H_0 : \theta = \theta_0$$

### 3) 대립가설 (Ha)

- 귀무가설에 대립되는 가설

### 4) 귀무가설과 대립가설

- 귀무가설은 등식, 대립가설은 등식이 아니다.
- 내가 증명하고 싶은것은 귀무가설, 대립가설 둘 다 가능
- 귀무가설, 대립가설이 서로 여집합일 필요는 없다.

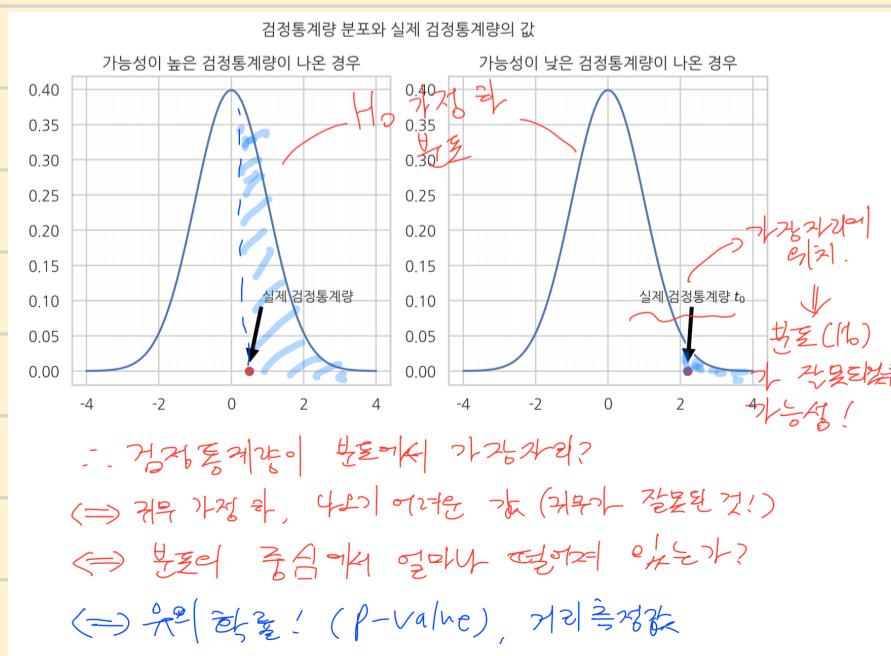
## 5) 검정통계량

## 검정과 유의확률

- 검정 증거 (시험 점수 같은 역할)
- 통계량 : 많은 데이터를 압축시켜 하나의 스칼라
  - > 검정에 사용되면 그것이 바로 '검정통계량'
- 검정통계량도 확률변수
- 검정통계량
  1. 베르누이분포 확률변수 -> 검정통계량 : 성공한 횟수  $n$  (이항분포 확률변수)
  2. 정규분포 확률변수(분산 알 때) -> 검정통계량 :  $z$ 통계량
  3. 정규분포 확률변수(분산 모를 때) -> 검정통계량 : 스튜던트  $t$ 분포의  $t$ 통계량
  4. 분산에 대한 검정 -> 검정통계량 : 표본분산을 정규화한 값 (자유도가  $n-1$ 인 카이제곱분포를 따름)

## 6) 유의확률

- 유의확률이 작다 ==> 귀무가설을 지지하지 않음. 대립가설 채택
- 검정통계량을 기준으로 가장자리의 면적을 잰다. (면적이 작을 수록 탈 중심)
- 이름은 '확률'이 불지만, 개념 상 거리의 개념 (거리가 멀 수록 면적이 작다.)
- 유의확률 : 귀무가설이 맞음에도 불구하고, 현재 검정통계량값과 같은 혹은 혹은 대립가설을 더 옹호하는 검정통계량 값이 나올 확률
  - \* 거리가 짧다 = 면적이 넓다 = 귀무가설 지지
  - \* 거리가 멀다 = 면적이 좁다 = 귀무가설 지지하지 않음



## 7) 단측검정 유의확률

09.04

검정과 유의확률

- 증명하고자 하는 대립가설이 부등식인 경우, 대립가설을 옹호하는 검정통계량 값이 나올 확률을 구할 때
- 특정한 한 방향의 확률만을 구해야 함. 이때 사용

### ex) 9.4.21 (모수가 클 때, 검정통계량도 정비례 증가 가정)

만약 증명하고자 하는 대립가설이 부등식인 경우에는 그 대립가설을 옹호하는 검정통계량값이 나올 확률을 구할 때 특정한 한 방향의 확률만을 구해야 한다. 이를 단측검정(one-side test, single-tailed test)이라고 한다.

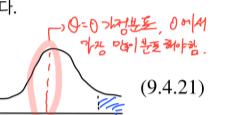
모수가 특정한 값보다 크다는 것을 증명하고 싶다면 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \theta = \theta_0, \quad H_a : \theta > \theta_0 \quad (9.4.21)$$

반대로  $\theta$ 가 특정한 값보다 작다는 것을 증명하고 싶다면 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \theta = \theta_0, \quad H_a : \theta < \theta_0 \quad (9.4.22)$$

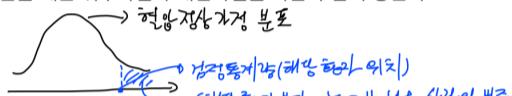
단측검정의 유의확률은 다음과 같이 구한다. 여기에서는 모수가 클 때 검정통계량도 정비례해서 같이 커진다고 가정한다. **가는 영역!**



### ex) 9page하단

어떤 환자의 혈압이 고혈압이라는 것을 증명하고 싶을 때는 귀무가설과 대립가설을 다음과 같이 놓는다.

- 귀무가설: '혈압이 정상이다'
- 대립가설: '고혈압이다'



이 검정에서 혈압 검사 결과를 통계량 분포로, 해당 환자의 혈압을 검정통계량으로 사용하여 계산한 우측유의확률이 0.02% 이 나왔다고 하자. 이는 정상인 중에서 혈압이 해당 환자의 혈압보다 더 높게 나온 사람은 0.02%뿐이었다는 뜻이다.

## 8) 유의수준과 기각역

- p-value가 얼마나 작아야 귀무가설 지지하지 않는 것인가
- 유의확률이 유의수준보다 작으면, 귀무가설 기각, 대립가설 채택
- 관례적으로 1,5,10%를 유의수준으로 설정
- 기각역 : 유의수준에 대해 계산된 검정통계량. 기각역을 알고 있다면, 검정통계량을 직접 기각역과 비교 -> 검정통계량이 기각역 안에 포함된다면, 귀무기각
- (유의확률을 유의수준(1,5,10% 등)과 비교할 필요 X)

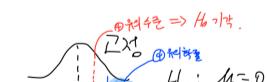
## 9) 검정 방법론 (10page)

1. 확률분포 및 모수 가정 -> 귀무가설 설정
2. 검정통계량 결정
3. 유의확률 및 유의수준 비교
4. 귀무가설 채택 or 기각

### 검정 방법론

검정의 기본적인 논리를 다시 정리하면 다음과 같다.

1. 데이터가 어떤 고정된 확률분포를 가지는 확률변수라고 가정한다. 예를 들어 통천은 베르누이분포를 따르는 확률변수의 표본이며 주식의 수익률은 정규분포를 따르는 확률변수의 표본이라고 가정한다.
2. 이 확률분포의 모수값이 특정한 값을 가진다고 가정한다. 이때 모수가 가지는 특정한 값은 우리가 검증하고자 하는 사실과 관련이 있어야 한다. 이러한 가정은 귀무가설이라고 한다. 예를 들어 동전이 공정한 동전이라고 주장하는 것은 베르누이 확률분포의 모수  $\theta$ 의 값이 0.5라고 가정하는 것과 같다. 주식이 손실을 보지 않는다는 것은 정규분포의 기댓값 모수  $\mu$ 가 0과 같거나 크다고 가정하는 것이다.
3. 만약 데이터가 주어진 귀무가설에 따른 표본이라면 이 표본 데이터를 특정한 수식에 따라 계산한 수치는 귀무가설에서 유도한 특정 확률분포를 따르게 된다. 이 수치를 검정통계량이라고 하며 검정통계량의 확률분포를 검정통계분포라고 한다. 검정통계분포의 종류 및 모수의 값은 처음에 정한 가설 및 수식에 의해 결정된다.
4. 주어진 귀무가설이 맞으면 표본 데이터에 대해서 실제로 계산된 검정통계량의 값과 같은 혹은 그보다 더 극단적인 (extreme) 또는 더 희귀한(rare) 값이 나올 수 있는 확률을 계산한다. 이를 유의확률이라고 한다.
5. 만약 유의확률이 미리 정한 특정한 기준값보다 작은 경우를 생각하자. 이 기준값을 유의수준이라고 하는데 보통 1% 혹은 5% 정도의 작은 값을 지정한다. 유의확률이 유의수준으로 정한 값보다도 작다는 말은 해당 검정통계분포에서 이 정통계치(혹은 더 극단적인 경우)가 나올 수 있는 확률이 아주 적다는 의미이므로 가장 근본이 되는 가설 즉, 귀무가설이 틀렸다는 의미이다. 따라서 이 경우에는 귀무가설을 기각한다.
6. 만약 유의확률이 유의수준보다 크다면 해당 검정통계분포에서 이 검정통계치가 나오는 것이 불가능하지만은 않다는 의미므로 귀무가설을 기각할 수 없다. 따라서 이 경우에는 귀무가설을 채택한다.



09.05

## 실습 정리 노트북 참고

Scipy.stats

활용한 검정

## 엔트로피

## 1) 엔트로피의 정의

- 엔트로피 : 확률분포가 가지는 정보의 확신도 혹은 정보량을 수치로 표현한 것  
확률밀도가 특정 값에 몰려있으면 -> 엔트로피가 작음  
확률밀도가 골고루 퍼져있으면 -> 엔트로피가 큼

- 엔트로피는 확률분포함수를 입력으로 받아 숫자를 출력하는 범함수로 정의

(10.1.1, 10.1.2)

확률변수  $Y$  가 카테고리분포와 같은 이산확률변수이면 다음처럼 정의한다.

$$H[Y] = - \sum_{k=1}^K p(y_k) \log_2 p(y_k) \quad (10.1.1)$$

이 식에서  $K$ 는  $X$ 가 가질 수 있는 클래스의 수이고  $p(y)$ 는 확률질량함수다. 확률의 로그값이 항상 음수이므로 음수 기호를 붙여서 양수로 만들었다.확률변수  $Y$  가 정규분포와 같은 연속확률변수이면 다음처럼 정의한다.

$$H[Y] = - \int_{-\infty}^{\infty} p(y) \log_2 p(y) dy \quad (10.1.2)$$

- $p(y)$ , 확률값 = 0 이면, 엔트로피 = 0

(로그값이 정의되지 않지만, 엔트로피 공식 상 0\*무한대 꼴 = 0으로 처리)(10.1.3)

엔트로피 계산에서  $p(y) = 0$ 인 경우에는 로그값이 정의되지 않으므로 다음과 같은 극한값을 사용한다.

$$\lim_{p \rightarrow 0} p \log_2 p = 0 \quad (10.1.3)$$

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

## 2) 엔트로피의 성질

- 엔트로피의 최소값은 0, 최대값은 K이다.

(최소일 때 : 확률변수가 결정론적. 특정한 하나의 값이 나오는 모두 쓸린 경우)

(최대일 때 : 이산확률변수가 가질 수 있는 값이  $2^K$ ,각 값에 대한 확률값이 모두 같은  $1/2^K ==>>$  엔트로피는 K)

## 3) 엔트로피의 추정

- 이론적인 pdf가 없을 때, 주어진 데이터에서 pdf를 추정 -> 엔트로피 계산

- `scipy.stats ==>>` 엔트로피 구하는 패키지 제공

ex) `sp.stats.entropy(p, base=2)`

## 4) 지니 불순도

- 엔트로피와 유사한 값 출력. 하지만, log를 취하지 않아 엔트로피 계산보다 계산량이 적음(10.1.5)

엔트로피와 유사한 개념으로 **지니불순도(Gini impurity)**라는 것이 있다. 지니불순도는 엔트로피처럼 확률분포가 어느쪽에 치우쳐있는가를 재는 척도지만 로그를 사용하지 않으므로 계산량이 더 적어 엔트로피 대용으로 많이 사용된다. 경제학에서도 사용되지만 지니계수(Gini coefficient)라는 다른 개념이라는 점에 주의해야 한다.

$$G[Y] = \sum_{k=1}^K P(y_k)(1 - P(y_k)) \quad (10.1.15)$$

## 5) 가변길이 인코딩

## 엔트로피

- 엔트로피 : 본래 통신분야에서 데이터가 가지고 있는 정보량을 계산하기 위해 고안

- 자주 나오는 글자, 데이터는 인코딩 시, 작은 수로 인코딩!

ex) A = '0', B = '10', C = '110', D = '111'

- 그 결과, 인코딩 결과 글자수가 줄어드는 효과

## 6) 엔트로피 최대화

(정보가 가장 적은)

•  $6^2$  와  $E[x] = 0$  이 주의질 때, 엔트로피가 가장 큰 학률분포는 ?  $\Rightarrow$  정규분포  $N(0, 6^2)$



비이즈 측정에서, 사실상



무정보 prior로 학률 ↑↑

\* Why 정규분포가 가장 엔트로피가 큰 분포인가?

$(0, 6^2)$

$$H[p(x)] = - \sum_{\text{클래스}} p(x) \log_2 p(x)$$



제한 조건 최적화  
(Equality constraint)

$$\begin{cases} \int p(x) dx = 1 \\ \int x p(x) dx = 0 \quad (\because 기대값 = 0) \\ \int (x-0)^2 p(x) dx = 6^2 \quad (\because 분산 = 6^2) \end{cases}$$



$p(x) =$  정규분포 PDF  $(N(0, 6^2))$

## 10.2

### 조건부 엔트로피

#### 1) 결합 엔트로피

##### - 결합pdf로 정의한 엔트로피 (10.2.1-2)

결합엔트로피(joint entropy)는 결합확률분포를 사용하여 정의한 엔트로피를 말한다.

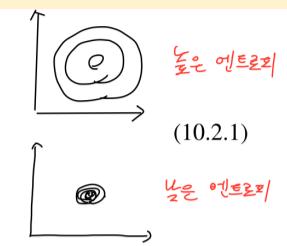
이산확률변수  $X, Y$ 에 대해 결합엔트로피는 다음처럼 정의한다.

$$H[X, Y] = - \sum_{i=1}^{K_X} \sum_{j=1}^{K_Y} p(x_i, y_j) \log_2 p(x_i, y_j) \quad (10.2.1)$$

이 식에서  $K_X, K_Y$ 는 각각  $X$ 와  $Y$ 가 가질 수 있는 값의 개수고  $p$ 는 확률질량함수다.

연속확률변수  $X, Y$ 에 대해 결합엔트로피는 다음처럼 정의한다.

$$H[X, Y] = - \int_x \int_y p(x, y) \log_2 p(x, y) dx dy \quad (10.2.2)$$



이 식에서  $p$ 는 확률밀도함수다.

#### 2) 조건부 엔트로피

- 조건부 엔트로피 : 해당 확률변수가 다른 변수의 예측에 도움이 되는지 확인하는데 활용

\* $H[Y|X] ==> \text{낮다면, } X\text{라는 변수가 } Y\text{변수 예측에 도움됨을 의미}(X\text{가 고정될 때, } Y\text{의 분포가 쓸림})$

##### - 조건 값이 정해진 경우 (10.2.3)

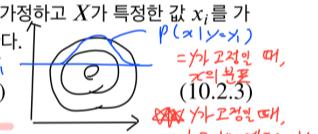
##### - 조건 값이 정해지지 않은 경우 (10.2.4)

\*조건 변수의 값에 따라 엔트로피가 달라짐

->각 조건변수 값에 대한 엔트로피의 가중평균 시행(가중치 = 조건변수의 확률값)

조건부엔트로피의 정의는 다음과 같이 유도한다. 확률변수  $X, Y$ 가 모두 이산확률변수라고 가정하고  $X$ 가 특정한 값  $x_i$ 를 가질 때의  $Y$ 의 엔트로피  $H[Y | X = x_i]$ 는 다음처럼 조건부확률분포의 엔트로피로 정의한다.

$$H[Y | X = x_i] = - \sum_{j=1}^{K_Y} p(y_j | x_i) \log_2 p(y_j | x_i) \quad (10.2.3)$$



조건부엔트로피는 확률변수  $X$ 가 가질 수 있는 모든 경우에 대해  $H[Y | X = x_i]$ 를 가중평균한 값으로 정의한다.

$$\begin{aligned} H[Y | X] &= \sum_{i=1}^{K_X} p(x_i) H[Y | X = x_i] \\ &= - \sum_{i=1}^{K_X} \sum_{j=1}^{K_Y} p(y_j | x_i) p(x_i) \log_2 p(y_j | x_i) \\ &= - \sum_{i=1}^{K_X} \sum_{j=1}^{K_Y} p(x_i, y_j) \log_2 p(y_j | x_i) \end{aligned} \quad (10.2.4)$$

~~그러나 고정된 때, Y의 분포는 고정된 때, X의 조건부 엔트로피가 높다면, Y는 고정된 때, X의 조건부 엔트로피가 낮다면, Y는 그때 유용하지 못함.~~

#### 3) 예측에 도움이 되는 경우

- 확률변수  $X$ 의 값을 고정(선택) 시,  $Y$ 의 분포가 확 쓸리도록 할 수 있는가( $Y$ 의 엔트로피가 낮아지는가)

= 조건부  $Y$ 의 엔트로피가 작은가 ==>  $X$ 는  $Y$ 예측에 도움이 된다.

-  $H[Y|X] = 0 \iff X\text{가 정해지는 순간, } Y\text{도 정해진다. 최소의 조건부 엔트로피}$

## 10.3

### 교차 엔트로피와

### 쿨백-라이블러 발산

## 1) 교차 엔트로피 (10.3.1-2)

### - 확률분포를 인수로 받음 (엔트로피, 결합엔트로피, 조건부엔트로피와의 차이점)

두 확률분포  $p, q$ 의 교차엔트로피(cross entropy)  $H[p, q]$ 는

이산확률분포의 경우에는 다음과처럼 정의한다.

$$H[p, q] = - \sum_{k=1}^K p(y_k) \log_2 q(y_k) \quad (10.3.1)$$

연속확률분포의 경우에는 다음과처럼 정의한다.

$$H[p, q] = - \int_y p(y) \log_2 q(y) dy \quad (10.3.2)$$

교차엔트로피는 지금까지 공부한 엔트로피, 결합엔트로피, 조건부엔트로피와 다르게 확률변수가 아닌 확률분포를 인수로 받는다는 점에 주의하라.

### - 분류모형의 성능 측정 : 교차 엔트로피

\*  $H[p, q]$ ,  $p, q$ 는 pdf(보통)  
정답의 예측 확률인  
정답의 예측 확률인

교차엔트로피는 분류모형의 성능을 측정하는데 사용된다.  $Y$ 가 0 또는 1이라는 값인 가지는 이진분류문제를 예로 들어보자.

$p$ 는  $X$ 값이 정해졌을 때 정답인  $Y$ 의 확률분포다. 이진분류문제에서  $Y$ 는 0 또는 1이다. 따라서  $p$ 는

- 정답이  $Y = 1$ 일 때,  $p(Y = 0) = 0, p(Y = 1) = 1$  (10.3.5)
- 정답이  $Y = 0$ 일 때,  $p(Y = 0) = 1, p(Y = 1) = 0$  (10.3.6)

따라서 확률분포  $p$ 와  $q$ 의 교차엔트로피는

- 정답이  $Y = 1$ 일 때,  $H[p, q] = -\cancel{p(Y = 0)} \log_2 q(Y = 0) - p(Y = 1) \log_2 q(Y = 1) = -\log_2 \mu$  (10.3.8)
- 정답이  $Y = 0$ 일 때,  $H[p, q] = -p(Y = 0) \log_2 q(Y = 0) - \cancel{p(Y = 1)} \log_2 q(Y = 1) = -\log_2(1 - \mu)$  (10.3.9)

이 값은 분류성능이 좋을수록 작아지고 분류성능이 나쁨수록 커진다. 이유는 다음과 같다.

- $Y = 1$ 일 때는  $\mu$ 가 작아질수록 즉, 예측이 틀릴수록  $-\log_2 \mu$ 의 값도 커진다.
- $Y = 0$ 일 때는  $\mu$ 가 커질수록 즉, 예측이 틀릴수록  $-\log_2(1 - \mu)$ 의 값도 커진다.

따라서 교차엔트로피값은 예측의 틀린정도를 나타내는 오차함수의 역할을 할 수 있다.

N 개의 학습 데이터 전체에 대해 교차엔트로피 평균을 구하면 다음 식으로 표현할 수 있다. 이 값을 로그손실(log-loss)이라고 한다.

$$\text{log loss} = -\frac{1}{N} \sum_{i=1}^N (y_i \log_2 \mu_i + (1 - y_i) \log_2 (1 - \mu_i)) \quad (10.3.10)$$

같은 방법으로 이진분류가 아닌 다중분류에서도 교차엔트로피를 오차 함수로 사용할 수 있다. 다중분류문제의 교차엔트로피 손실함수를 카테고리 로그손실(categorical log-loss)이라고 한다.

$$\text{categorical log loss} = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K (\mathbb{I}(y_i = k) \log_2 p(y_i = k)) \quad (10.3.11)$$

위 식에서  $\mathbb{I}(y_i = k)$ 는  $y_i$ 가  $k$ 인 경우에만 1이고 그렇지 않으면 0이 되는 지시함수(indicator function)다.  $p(y_i = k)$ 는 분류모형이 계산한  $y_i = k$ 일 확률이다.

## 2) 교차 엔트로피를 사용한 분류성능 측정

### - 교차 엔트로피 값 : 예측의 틀린 정도를 나타냄 (오차함수의 역할)

### - 로그 손실(이진분류) : N개의 학습데이터 전체에 대해 구한 교차엔트로피 평균(10.3.10)

$N$  개의 학습 데이터 전체에 대해 교차엔트로피 평균을 구하면 다음 식으로 표현할 수 있다. 이 값을 로그손실(log-loss)이라고 한다.

$$\text{log loss} = -\frac{1}{N} \sum_{i=1}^N (y_i \log_2 \mu_i + (1 - y_i) \log_2 (1 - \mu_i)) \quad (10.3.10)$$

### - 카테고리 로그 손실(다중분류) : 다중분류 문제의 교차엔트로피 손실함수(10.3.11)

같은 방법으로 이진분류가 아닌 다중분류에서도 교차엔트로피를 오차 함수로 사용할 수 있다. 다중분류문제의 교차엔트로피 손실함수를 카테고리 로그손실(categorical log-loss)이라고 한다.

$$\text{categorical log loss} = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K (\mathbb{I}(y_i = k) \log_2 p(y_i = k)) \quad (10.3.11)$$

위 식에서  $\mathbb{I}(y_i = k)$ 는  $y_i$ 가  $k$ 인 경우에만 1이고 그렇지 않으면 0이 되는 지시함수(indicator function)다.  $p(y_i = k)$ 는 분류모형이 계산한  $y_i = k$ 일 확률이다.

## 10.3

교차 엔트로피와

쿨백-라이블러 발산

### 3) 쿨백-라이블러 발산

- 쿨백-라이블러 발산 : 두 확률분포  $p(y), q(y)$ 의 분포모양이 얼마나 다른지를 숫자로 계산한 값
- 정의 (10.3.12-13)

쿨백-라이블러 발산(Kullback-Leibler divergence)은 두 확률분포  $p(y), q(y)$ 의 분포모양이 얼마나 다른지를 숫자로 계산한 값이다.  $KL(p||q)$ 로 표기한다.

이산확률분포에 대해서는 다음처럼 정의한다.

$$\begin{aligned} KL(p||q) &= H[p, q] - H[p] \\ &= \sum_{i=1}^K p(y_i) \log_2 \left( \frac{p(y_i)}{q(y_i)} \right) \end{aligned} \quad (10.3.12)$$

연속확률분포에 대해서는 다음처럼 정의한다.

$$\begin{aligned} KL(p||q) &= H[p, q] - H[p] \\ &= \int p(y) \log_2 \left( \frac{p(y)}{q(y)} \right) dy \end{aligned} \quad (10.3.13)$$

- 상대 엔트로피 라고도 불림 : 교차엔트로피 -  $p$ 분포의 엔트로피(기준이 되는  $p$  분포)

\* 쿨백-라이블러 발산  $> 0$ , 두 분포가 같을 때 쿨백-라이블러 발산 값 = 0

- 쿨백-라이블러 발산 : 거리 개념이 아님. 확률분포  $q$ 가 기준확률분포  $p$ 와 얼마나 다른지를 나타내는 값.

\* 두 확률분포의 위치가 달라지면, 값이 달라짐  $KL(p||q) \neq KL(q||p)$

### 4) 가변길이 인코딩과 쿨백-라이블러 발산

- 확률분포  $q$ 의 모양이  $p$ 의 모양과 다른 정도를 정량화한 것.
- 잘못된 분포  $q$ 로 인코딩 결과 글자수 - 제대로 분포  $p$ 로 인코딩 결과 글자수 =  $q$ 와  $p$ 의 분포 모양이 다른 정도를 정량화
- 예제 구현 및 10.3.2 연습문제 풀이

## 1) 상호정보량

## 상호정보량

- 상호정보량 : 상관관계를 대체할 수 있는 값
- 비선형적 상관관계를 알아내는 데 도움
- 상호정보량 = 결합pdf와 주변pdf곱의 쿨백-라이블러 발산 값
- 두 확률변수가 독립이면, 상호정보량 = 0 = 쿨백-라이블러 발산 값 (1page상단)

상호정보량(mutual information)은 결합확률밀도함수  $p(x, y)$ 와 주변확률밀도함수의 곱  $p(x)p(y)$ 의 쿨백-라이블러 발산이다. 즉 결합확률밀도함수와 주변확률밀도함수의 차이를 측정하므로써 두 확률변수의 상관관계를 측정하는 방법이다. 만약 두 확률변수가 독립이면 결합확률밀도함수는 주변확률밀도함수의 곱과 같으므로 상호정보량은 0이 된다. 반대로 상관관계가 있다면 그만큼 양의 상호정보량을 가진다.

$$MI[X, Y] = KL(p(x, y) || p(x)p(y)) = \sum_{i=1}^K p(x_i, y_i) \log_2 \left( \frac{p(x_i, y_i)}{p(x_i)p(y_i)} \right) \quad (10.4.3)$$

$$MI = KL(p(x, y) || p(x)p(y))$$

$p(x, y) = p(x)p(y)$       거리       $p(x, y) \neq p(x)p(y)$

독립      ↗      상관 0  
상관 ↓      상관 ↑

독립에 얼마나 가까운가,  
상관에 얼마나 가까운가,  
↓  
 $p(x, y)$  와  $p(x)p(y)$  의  
모양이 다른 정도를 정량화  
 $\Leftrightarrow KL(p(x, y) || p(x)p(y))$   
 $\Leftrightarrow$  독립이면 0

## 2) 최대정보 상관계수

- 연속확률변수의 표본 데이터에서 상호정보량 추출
- 연속확률변수는 상호정보량을 직접 추출할 수 있는 방법은 없음
- 대신, 구간을 나눠 카테고리 변수로 전환 (연속확률변수 -> 이산확률변수)
  - \* 히스토그램을 통해 유한개의 구간(bin)으로 나누어 확률분포함수 추정  
=> 카테고리 확률분포로 변환
  - \* 구간을 나누는 방법을 다양하게 시도 후 -> 각각의 상호정보량 중 가장 큰 값 선택 -> 정규화  
=> 최대정보 상관계수 (MIC, Maximum Information Coefficient)

• 엔트로피  $\Rightarrow$  정보의 분산 정도!  
 ↓  
 (지지부) (확률)

$$H[Y] = - \sum_{i=1}^{k_Y} P(Y_i) \log_2 P(Y_i)$$

↑  
 ↑  
 조건부 엔트로피



↑  
 조건부 엔트로피  
 $y=y_0$

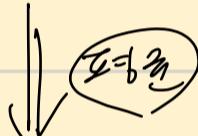
:  $y=y_0$ 라는 조건일 때,  
 $X$ 의 분산 정도가 크면  
 (엔트로피가 크면)  
 조건 ( $y=y_0$ )은 무의미!

$$H[X|Y] = - \sum_{i=1}^{k_X} \sum_{j=1}^{k_Y} P(x_i, y_j) \log_2 P(x_i, y_j)$$

$$H[X|Y=y_0] = - \sum_{i=1}^{k_X} P(x_i|y=y_0) \log_2 P(x_i|y=y_0)$$

↓  
 $H[X|Y] = H[X|Y=y_0]$  가 증명됨  
 $= \sum_{i=1}^{k_Y} H[X|Y=y_i] P(y_i)$

↓  
 교차 엔트로피



KL divergence

: binary classification의 loss function으로!  
 :  $KL(P||Q)$   
 P와 Q의 분포가 얼마나 다른지

↑  
 $H[P, Q] = - \sum_{i=1}^k P(y_i) \log_2 Q(y_i)$

↑  
 $\log$  loss =  $- \frac{1}{N} \sum_{i=1}^N (y_i \log_2 u_i + (1-y_i) \log_2 (1-u_i))$

↓  
 상호 정보량  
 :  $I(P(X,Y)) = H[X] - H[X|Y]$   
 $KL(P(x,y) || p(x)p(y))$

↓  
 $KL$  divergence,  $KL(P||Q)$   
 $= H[P, Q] - H[P]$

MI  
 $= KL(p(x,y) || p(x)p(y))$

최대 정보 상관계수  
 : 비선형 분포의 상관정도 확인  
 $KL$  중 최대값을 정극화