최소최대와 분류과정

정의

- 개요

1. 검정과 관련된 기각역의 최소최대를 결정하는 방법
2. 어떤 함수 $\delta = u(X_1, \ldots, X_n)$이라고 정의하자.

1) 이 함수는 어떤 검정 $H_0 : \theta = \theta_1$ VS $H_1 : \theta > \theta_2$를 검정하는 함수이다.
2) 이 때, 이 함수와 관련된 손실함수를 정의하자. 즉, 함수 δ에 대하여

(1) $\mathcal{L}(\theta, \delta = \theta_1) = 0$ 이고,$\mathcal{L}(\theta, \delta = \theta_0) = 0$ 이다.

(2) 하지만, 오답인 경우에 이 손실함수는

- $\mathcal{L}(\theta, \delta = \theta_1) > 0$ 이고,$\mathcal{L}(\theta, \delta = \theta_0) > 0$ 이다.

3) 이 때, 이 예측함수 δ를 기각역의 개념과 연결지어서

(1) $H_0 : u(X_1, \ldots, X_n) \in C^c$ 이면 $\theta = \theta_0$

(2) $H_1 : u(X_1, \ldots, X_n) \in C$ 이면 $\theta = \theta_1$

정의

- 개요

3. 한편, 역으로 이를 기각역 c에 따른 검정으로 변환할 수 있다. 즉

1) $\mathcal{L}(\theta, \delta = [\theta_1, \theta_0]) \rightarrow \mathcal{L}(\theta, C = [C, C^c])$

2) 즉 위 함수를 다시 풀어보면

(1) $X_1, \ldots, X_n \in C$이면 θ_1로 분류하고

(2) $X_1, \ldots, X_n \in C^c$이면 θ_0으로 분류한다.

4. 이 때, 손실함수의 기댓값인 위험함수를 정의하면
$R(\theta, C) = \int_C \mathcal{L}(\theta, \theta_1)\mathrm{L}(\theta; \mathrm{n}) + \int_{C^c} \mathcal{L}(\theta, \theta_0)\mathrm{L}(\theta; \mathrm{n})$

1) 만약 정답이 $\theta = \theta_1$ 일때, $\int_C \mathcal{L}(\theta, \theta_1)\mathrm{L}(\theta; \mathrm{n})$는 소거되고

2) 만약 정답이 $\theta = \theta_0$ 일때, $\int_{C^c} \mathcal{L}(\theta, \theta_0)\mathrm{L}(\theta; \mathrm{n})$가 소거된다.

3) 즉, 옳은 선택을 한 부분의 위험값은 0으로 평가된다.

정의

- 개요

5. 이제, 목표는 $\max\ [R(\theta_1, C), R(\theta_2, C)] = R$ 이라고 정의할 때,
   이를 최소화하는 C를 찾는것이다.

1) 이는 손실함수 $\mathcal{L}(\theta_0, \theta_1)$ 나 $\mathcal{L}(\theta_1, \theta_0)$을 최소화하거나
2) 그 트레이드 오프 관계인 $\int_C \mathrm{L}(\theta_0; \mathrm{n}) = \int_{C^c} \mathrm{L}(\theta_1; \mathrm{n})$ 인 지점에서 최소화된다.
3) 이 때, C는 C={$u(X_1, \ldots, X_n)$: $\frac{\mathrm{L}(\theta_1; X_1, \ldots, X_n)}{\mathrm{L}(\theta_2; X_1, \ldots, X_n)} \leq \mathrm{k}$} 일 때, 임의의 k가 C를 결정짓고

(1) 이 때, 그 C가$\mathcal{L}(\theta_0, \theta_1) \int_C \mathrm{L}(\theta_0; \mathrm{n}) = \mathcal{L}(\theta_1, \theta_0) \int_{C^c} \mathrm{L}(\theta_1; \mathrm{n})$를 결정짓는다.

정의

- 분류문제의 응용

1. 어떤 관측값들을 X,Y 둘 중 하나의 확률변수에서 추출됐는지 알아내는 문제가 주어졌다고 가정하자.
2. 이 때, 이 두 확률변수 X,Y가 결합 PDF $\frac{f(x,y;\theta_0)}{f(x,y;\theta_1)} \leq k$일 때 θ_1에 의존하는 분포 예를 들면 Y로 분류한다.

예제

- 분류문제의 응용

1. $[x, y]$를 모수 $\mu_1, \ \mu_2, \sigma_1^2, \sigma_2^2, \rho$를 갖는 이변량 정규분포의 쌍 X,Y의 관측값이라고 하자.

2. 결합 PDF는

$$f(x, y; \mu_1, \ \mu_2, \sigma_1^2, \sigma_2^2) \ = \frac{1}{(2\pi)^{\frac{n}{2}} \ \sigma_1 \sigma_2 \sqrt{(1-\rho^2)}} \ \exp\left(-\frac{1}{2} q\right)$$

이 때 $q = \frac{1}{1-\rho^2} [\left(\frac{x-\mu_1}{\sigma_1^2}\right)^2 - 2\rho \ \left(\frac{x-\mu_1}{\sigma_1}\right) \left(\frac{y-\mu_2}{\sigma_2}\right) + \ \left(\frac{y-\mu_2}{\sigma_2^2}\right)^2]$

1) 부등식 $\frac{f(x, y; \mu_x, \ \mu_2, \sigma_1^2, \sigma_2^2)}{f(x, y; \mu'_1, \ \mu'_2, \sigma'^2_1, \sigma'^2_2)} \leq k$는 다음과 같다.

(1) $-\frac{1}{2}\left[q(x, y; \mu_1, \ \mu_2) - q\left(x, y; {\mu'}_1, \ {\mu'}_2\right)\right] \leq \log(k)$

(2) 정리하면

$$\frac{1}{1-\rho^2}\left\{\left[\frac{\mu_1 - {\mu'}_1}{\sigma_1^2} - \frac{\rho\left({\mu'}_2 - \mu_2\right)}{\sigma_1 \sigma_2}\right] x + \left[\frac{\mu_2 - {\mu'}_2}{\sigma_2^2} - \frac{\rho\left({\mu'}_2 - \mu_2\right)}{\sigma_1 \sigma_2}\right] y\right\} \leq \log(k) + \frac{1}{2}\left[q(0,0; \mu_1, \ \mu_2) - q\left(0,0; {\mu'}_1, \ {\mu'}_2\right)\right]$$

(3) 위 식을 간단히 한 ax+by ≤c에서 특정 c값보다 작으면 $f(x, y; \mu'_1, \ \mu'_2, \sigma'^2_1, \sigma'^2_2)$에서 추출한 것으로 본다.