판별분석

Jinseog Kim
Dongguk University
jinseog.kim@gmail.com

2017-09-04

판별분석-Discriminant analysis

- 몇개의 알려진 그룹으로부터 그 그룹을 구분해 주는 함수를 결정한다.
- 결정된(추정된) 판별함수를 이용하여 새로운 관측치를 분류한다.

통계적 의사결정

- $\mathbf{X}_i, i = 1, \ldots, n$를 m차원 랜덤벡터, y_i를 그룹변수
- $\mathbf{X}$가 주어진 경우 y값을 예측하기 위한 함수 $f(\mathbf{X})$를 어떻게 찾을 수 있을까?
- 손실함수 (loss function) - 손실함수의 예로 제곱오차(squared error loss)를 고려하자.

$$L(Y, f(\mathbf{X})) = (Y - f(X))^2.$$

- $f(\mathbf{X})$가 손실함수(L)에 대한 예측오차(Expected squared prediction error) 최소화 하여야 한다.

$$EPE(f) = E(Y - f(X))^2 = \int (y - f(x))^2 Pr(dx, dy).$$

통계적 의사결정

- 예측오차는 조건부 기대값을 이용하면,

$$EPE(f) = E_X\{E_{Y|X}[(Y - f(X))^2|X]\}.$$

- 예측오차를 최소화하는 f는 X가 주어질때 Y의 조건부 기대값

$$\arg_f \min EPE(f) = \arg_f \min E_X\{E_{Y|X}[(Y - f(X))^2|X]\} = E(Y|X = x).$$

통계적 의사결정

- 만일, $y \in \{1, 2, \ldots, K\}$라고 하고, 손실함수를 0-1 loss를 고려한다면,

$$\arg\min_f EPE(f) = P(Y \neq f(x)|X = x)$$

- 오분류률을 Bayes error라고 한다.

LDA and QDA

- $Y \in \{1, 2, \ldots, K\}$라고 하고,
- $Y = k$에서 얻어진 관측치의 분포 - 다변량 정규 분포

$$f_k(x_1, \ldots, x_p) = f_k(\mathbf{x}) = (2\pi)^{-p/2} |\Sigma_k|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_k)' \Sigma_k^{-1} (\mathbf{x} - \mu_k)\right).$$

- 사전 확률 - Y가 그룹 k에 속할 확률

$$\pi_k = P(Y = k), \quad \sum_{k=1}^{K} \pi_k = 1.$$

- $Y = k$가 주어질 때, X의 조건부 확률

$$P(\mathbf{X} = \mathbf{x} | Y = k) = f_k(\mathbf{x}).$$

LDA and QDA

- 사후확률 분포 (posterior probability density) - Bayes 정리에 의하여

$$Pr(Y = k|\mathbf{X} = \mathbf{x}) = \frac{f_k(\mathbf{x})\pi_k}{\sum_k^K f_k(\mathbf{x})\pi_k} \propto f_k(\mathbf{x})\pi_k, \quad k = 1, \ldots, K.$$

- 자료 $\mathbf{X}$가 어느 그룹에 속하는지를 판별한 때, 사후확률이 가장 큰 그룹에 할당하는 규칙을 Bayes rule, 즉,

$$\hat{f}(\mathbf{x}) = \arg\max_k \Pr(Y = k|\mathbf{X} = \mathbf{x}) = \arg\max_k f_k(\mathbf{x})\pi_k.$$

LDA and QDA

- 위 식을 최대화하는 것은 log변환된 식을 최대화 하는 것과 동일

$$\max_k \left\{ -\frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k) - \frac{1}{2}\ln|\Sigma_k| + \ln(\pi_k) \right\}$$

- 위 함수는 이차함수가 되므로 이를 이차판별함수라고 부르고, 이차판별함수를 이용한 판별분석을 이차판별분석(QDA)
- 만일 $\Sigma_1 = \ldots = \Sigma_K = \Sigma$라고 하면, 즉, 다변량 확률밀도함수가 아래와 같다고 하면,

$$f_k(\mathbf{x}) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_k)'\Sigma^{-1}(\mathbf{x}_k-\mu_k)\right), k = 1, \ldots, K$$

LDA and QDA

$$\max_k Q_k(\mathbf{x}) = \max_l \left\{ -\frac{1}{2}(\mathbf{z} - \mu_k)^T \Sigma^{-1} (\mathbf{x} - \mu_k) + \ln(\pi_k) \right\}$$

$$= \max_k \left\{ \mathbf{x}^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu^T \Sigma^{-1} \mu_k + \ln(\pi_k) \right\}$$

- 1차함수를 최대화하는 k를 찾으면 된다. 이 때 다음 함수를 선형판별함수라고 부른다.

$$L_k(\mathbf{X}) = \mathbf{X}^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu^T \Sigma^{-1} \mu_k + \ln(\pi_k)$$

R에서의 판별분석

- R은 선형판별분석과 이차판별분석을 위하여 `lda()`와 `qda()` 두 가지 함수를 제공하고 있다. 이 함수들은 MASS 라이브러리안에 내장되어 있다.