데이터마이닝(DataMining)

Chapter 11. 판별분석

1. 판별분석

- 변수 유도 기법 (Variable-directed techniques)은 p 개의 변수 $(x_1, x_2, \cdots, x_p)$들의 상관관계를 이용하여 주성분 변수(원 변수의 선형결합)를 도출하여 80% 규칙에 의해 p차원을 주성분 차원으로 축약하는 주성분 분석과 원 변수를 요인(factor)이라는 잠재 변인을 이용하여 상호 배타적으로 그룹화하는 방법인 요인분석으로 나눌 수 있다.

- 판별 분석(Discriminant Analysis)은 군집 분석(Clustering Analysis)과 함께 개체들에 대해 측정된 특성(변수) 값을 이용하여 개체를 판별하는 식을 유도하여 새로운 개체의 집단을 판별하거나 개체의 유사성을 계산하여 유사한 개체끼리 군집화하는 개체 유도 기법 (individual directed techniques)이다.

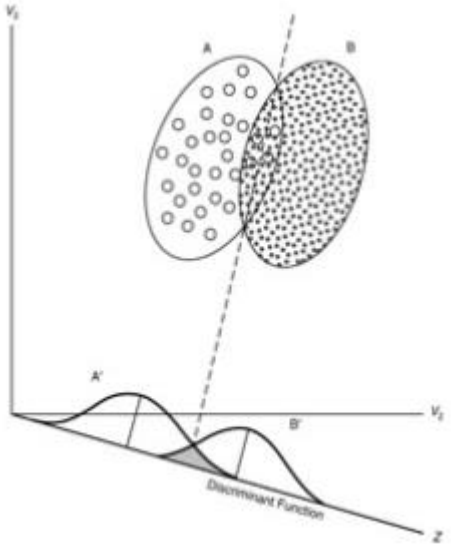
1. 판별분석

- 군집분석에서는 개체의 그룹에 대한 정보 없이 유사성이 가까운 개체들끼리 계층적으로 묶어 가거나 군집의 개수를 정하여 군집의 중심점을 이용하여 개체를 군집화 하는 방법이다.

- 판별분석은 자료 수집 시 이미 그룹이 나누어져 있어 이를 가장 잘 판별하는 판별규칙을 도출하여 새로운 개체의 군집을 판별하는 방법이다.

1. 판별분석

- 판별규칙은 모집단 그룹을 정확하게 분류하는 판별 변수의 함수식으로 판별규칙을 설정한다.
- 2개 모집단, 판별변수의 함수식으로 판별규칙을 얻었었다고 가정할 때 아래 분포의 빗금친 부분 중 B 모집단에 속한 개체를 판별규칙에 의해 A집단으로 분류한 오분류 확률, A 모집단에 속한 개체를 판별규칙에 의해 B집단으로 분류한 오분류 확률을 구할 수 있다.



2. 오분류

- Re-substitution 규칙 : 수집된 데이터로부터 얻은 판별식을 원 데이터에 적용하여 개체를 분류하여 오분류 비율을 구하는 방법이다.

- test 데이터 이용 : 데이터를 양분하여 한 개체 그룹으로부터 판별식을 유도하고, 이 판별식을 사용하여 다른 그룹의 개체를 분류하여 오분류 비율을 추정한다.

- Cross-validation 추정법 : Lachenbrush(1968)가 제안한 방법으로 가장 널리 사용된다. 첫 번째 개체 하나를 제외하고 판별식을 구하여 그 개체를 분류하고, 첫 번째 개체를 다시 넣고 두 번째 개체를 제외하고 판별식을 구한 후 두 번째 개체를 분류하는 방법을 반복하여 오분류 비율을 추정한다. 이 방법을 Jackknife 방법이라고도 한다.

3. 두 모집단에 대한 판별분석

- 판별 측정변수 벡터 : $\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \sim (\underline{\mu}, \Sigma)$
- 모집단1(π_1) : $f_1(\underline{x}) \sim MN\left(\underline{\mu_1}, \Sigma_1\right)$
- 모집단2(π_2) : $f_2(\underline{x}) \sim MN(\underline{\mu_2}, \Sigma_2)$
- 각 모집단으로부터 (n_1, n_2) 크기의 표본 개체를 추출하여 판별 측정변수를 관측하고, 각 표본의 평균벡터를 $(\bar{x}_1, \bar{x}_2)$, 공분산행렬을 (S_1, S_2)라 하자.

3. 두 모집단에 대한 판별분석

- P_i : 모집단 i의 사전확률
- $P(X \in R_i|\pi_j)$: 모집단 j에 속한 개체가 판별식에 의해 모집단 i로 판별할 확률
- $c(j|i)$: 모집단 i에 속한 개체를 모집단 j로 판별하여 발생하는 비용
- $P_i P(X \in R_i|\pi_i)$: 모집단 i에서 추출된 개체를 모집단 i로 분류
- $P_j P(X \in R_i|\pi_j)$: 모집단 i에서 추출된 개체를 모집단 i로 분류

3. 두 모집단에 대한 판별분석

- 오분류 기대비용 최소화
  - 오분류 기대비용=(오분류확률*비용함수)
  - $E = c(2|1)P_1P(X \in R_2|\pi_1) + c(1|2)P_2P(X \in R_1|\pi_2)$

$$\rightarrow R_1 : \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq \frac{P_2}{P_1}\frac{c(1|2)}{c(2|1)}$$

$$\rightarrow R_2 : \frac{f_2(\underline{x})}{f_1(\underline{x})} \geq \frac{P_1}{P_2}\frac{c(2|1)}{c(1|2)}$$

3. 두 모집단에 대한 판별분석

- 우도함수
  - 각 모집단이 다변량정규분포를 따른다면, 새로운 개체 $\underline{x}_0$에 대해

$$L\left(\underline{x}_0; \underline{x}_1, \widehat{\Sigma_1}\right) \geq L\left(\underline{x}_0; \underline{x}_2, \widehat{\Sigma_2}\right)$$

  을 통해 판별

3. 두 모집단에 대한 판별분석

- Fisher 선형 판별
  - 다변량 정규분포의 두 모집단이 동일한 공분산을 갖는다면 합동(통합)공분산을 통해

    추정 $\hat{\Sigma} = \frac{(n_1-1)S_1+(n_2-1)S_2}{(n_1+n_2-2)}$

  - 따라서 우도함수는 $\left(\underline{x}_1 - \underline{x}_2\right)'\hat{\Sigma}^{-1}\underline{x}_0 - \frac{1}{2}\left(\underline{x}_1 - \underline{x}_2\right)'\hat{\Sigma}^{-1}\left(\mu_1 - \underline{\mu}_2\right) \geq c,$

    $where\ c = ln\left(\frac{c(1|2)P_2}{c(2|1)P_1}\right)$

  - 두 모집단의 공분산이 동일하지 않은 경우

$$-\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}\left(\underline{x} - \underline{\mu_i}\right)' \Sigma_{\mathrm{i}}^{-1}\left(\underline{x} - \underline{\mu_i}\right) + \log(\mathrm{P}(\mathrm{X} \in R_i))$$