12장 판별분석

1. 서론

판별분석 - 분류기법, 로지스틱 회귀분석과 같이 분류와 프로파일링에 쓰이는 전통적 통계기법

앞에서본 다른 분류 기법들과 유사

2. 클래스로 부터 관측치에 이르는 거리

클래스 - 관측치 간의 거리측정이 필요 - 가장 가까운 클래스에 배정

기초적인 방법: 유클리드 거리 (Euclidean distance)

$$D_{Euclidean}(X, \bar{X}) = \sqrt{(X_1-\bar{X}_1)^2 + \cdots + (X_p-\bar{X}_p)^2}$$ → 단점 존재

① 거리측정이 입력변수에 대한 측정단위 선택에 따라 달라짐

↳ 연간 소득을 천 달러 기준에서 달러로 바꾸면 거리도 달라짐

② 변수의 변동성을 반영하지 못함

↳ 각 변수들의 변동성 (표준편차, 분산)을 반영하지 못함

⇒ 표준화된 값 (Z-score 등)을 이용해야함

③ 변수들 간의 상관관계를 무시

변수가 많을경우 특히 중요 - 상관관계에 의해서 변수 중요도가 바뀔 수 있음

⇒ 이러한 단점 해결책으로 통계적 거리 (Statistical Distance) 혹은 마할라노비스 거리 (Mahalanobis) 사용

↳ P개의 변수들 사이의 공분산 행렬 S

$$D_{Sta}(X, \bar{X}) = [X-\bar{X}]' S^{-1} [X-\bar{X}]$$

$$= [(x_1-\bar{x}_1), (x_2-\bar{x}_2), \cdots, (x_p-\bar{x}_p)] S^{-1} \begin{bmatrix} x_1-\bar{x}_1 \\ x_2-\bar{x}_1 \\ \vdots \\ x_p-\bar{x}_p \end{bmatrix}$$

$[X-\bar{X}]'$ → 열벡터를 행벡터로 변환 S^{-1} → S의 역행렬

↳ 통계적 거리는 예측변수들의 평균값뿐만 아니라 예측변수 값들의 분산성과 상관관계도 고려 (공분산 행렬을 계산에 포함)

↳ 예측변수의 평균(중심점)과 변수들 간의 공분산 계산 필요

⇒ 판별분석은 통계적 거리를 기반으로 분리선 (예측변수가 세 개 이상인 경우에는 분리초평면)을 찾는다.

↳ 모든 클래스의 평균들으로부터 동일 거리에 위치

3. 피셔의 선형 분류함수

Linear Classification Function = '클래스 간 변동성 대 클래스 내 변동성'의 비율을 최대화 시키는 선형 함수

Fisher Discrimination Function — 클래스 간 변동성 대 클래스 내 변동성 비율로 최적화하는 선형 판별함수

⇒ 각 클래스 내의 관측치들은 동질적이나 서로 다른 클래스끼리는 이질적이게 분류하는 선을 찾는게 목표

↳ 관측치가 주어졌을 때, 각 클래스까지의 근접성을 점수로 계산하고 이들 중 가장 높은 분류점수

(가장 가까운 통계적 거리)를 주는 클래스로 관측치로 분류

↳ 분류함수의 개수는 클래스의 수와 동일

또 다른 방법은 관측치가 클래스에 속할 확률을 구하고, 이 중에서 가장 높은 확률을 갖는 클래스로 할당

m개의 클래스가 있는 경우 관측치 i가 클래스 k에 속할 확률 계산

↳ 먼저 모든 분류점수 $C_1(i), C_2(i), \cdots, C_m(i)$ 계산 후

$P[\text{관측치 } i \text{가 클래스 } k \text{에 속할 경우}] = \frac{e^{C_k(i)}}{e^{C_1(i)} + e^{C_2(i)} + \cdots + e^{C_m(i)}}$ 이용해 결합

4. 판별분석의 분류성능

판별분석은 분류점수에 이르는 2가지 주요 가정에 의존

① 모든 클래스의 특성치들은 다변량 정규분포로부터 얻어진 것으로 가정

↳ 이러한 가정이 만족된다면, 다른 분류방법보다 더 효과적인 분류방법

가정 만족 시에 더 적은 데이터로도 분류가능

② 예측변수들 간의 상관관계는 특정 클래스 내에서나 다른 클래스에서도 동일

↳ 각 클래스 별로 상관관계 행렬을 계산하고 이를 비교함으로써 확인 가능

↳ 클래스 별 상관관계가 많이 다르면 변동이 큰 클래스로 분류하는 경향

↳ 상관관계 구조가 유의하게 다르고 데이터 세트가 매우 크다면

이차 판별분석 (quadratic discriminant Analysis)을 사용하는 것이 대안

모델 구축 전에 정규성이나 상관관계에 대한 탐색적 분석을 수행하는 것이 필요.

분류 정확도는 Confusion matrix나 lift chart 이용해 비교가능

5. 사전확률

지금까지의 판별분석 — 오분류를 최소화하는 것으로 가정

↳ 앞에 방법들은 클래스 별로 제시되는 관측치의 수가 동일한 것으로 가정 → 만약 서로 다르다면?

관측치들이 제시될 확률이 클래스 별로 다르면 평균오차율을 줄이기 위해 분류함수 수정 필요.

↳ P_j가 클래스 j에 속할 사전확률 (미래확률)

그 후 $\log(P_i)$를 분류함수에 상수로 추가

6. 서로 다른 오분류 비용

클래스 간의 오분류 비용이 현저하게 다를 경우 오분류의 기대 비용을 최소화해야 한다.

ㄴ 클래스 1에 속하는 멤버를 오분류하는 비용을 q_1, 클래스 2를 오분류하는 비용을 q_2라 할 때,

$\log(q_1)$, $\log(q_2)$를 각각의 분류함수에 상수로 더해주면 된다.

ㄴ 사전확률과 오분류 비용을 다 고려하기 위해서 $\log(P_1 q_1)$을 더해주면 됨.

ㄴ 실제로 오분류 비용을 계산하는 것보다 그 비율을 (q_2/q_1)을 계산하는 것이 더 쉽다.

⇒ 클래스 2의 분류함수에 $\log(q_2/q_1)$만 더해주면 됨.

8. 판별분석의 장단점

주로 통계적 분류기법으로 여겨짐

판별분석의 사용과 성능은 다중선형 회귀분석과 유사 → 장단점도 유사

ㄴ 판별분석은 예측변수의 최적 가능치를 찾는 과정을 포함

선형 회귀분석에서의 가중치는 반응(종속)변수와 관련 ↔ 판별분석에서는 클래스들을 분리시키는 것과 관련

두 분석 모두 최소제곱법 사용하는 추정방법 이용, 결과로 얻은 추정치들은 국소 최적값에 영향을 받지않고 강건.

ㄴ 기본과정은 정규성(normality) 판별분석에서는 예측변수들이 다변량 정규분포를 따른다고 가정

현실적으로 지켜지지 않는 경우가 많지만, 이런 문제에 강건

너무 한쪽으로 치우친 경우에는 로그변환 등을 시행 ⇒ 성능향상이 가능

ㄴ 분류방법으로서의 장점: 단일 예측변수의 기여도에 대한 추정치를 제공

계산과정이 간단·간결하고 데이터세트가 작을 때 유용