

10.1 10.2 두 개 그룹의 판별

2019년 6월 16일 일요일 오전 3:11

집단에 대한 정보로부터 집단을 구별할 수 있는 판별함수 또는 판별규칙을 만들고 새로운 개체에 대해 어느 집단에 속하는지를 판별하여 분류하는 다변량 기법

모집단 = 그룹

그룹간의 차이를 크게 해주는 그룹 판별함수를 구하고자함.

판별함수는 그룹간의 거리를 최대한 멀리하도록 만든 변수들의 선형결합식이 됨

판별함수를 구한 후에는 판별함수를 이용해 기존의 개체들을 분류하여 오분류율을 계산함.

새로운 개체에 대해서는 판별함수를 이용하여 속하는 그룹을 추정할 수 있음.

판별과 분류분석의 목적

- 1) 몇개의 알려진 그룹으로부터 그룹의 특성을 나타내주고 구별해주는 함수를 결정함.
- 2) 결정된 판별함수를 이용하여 새로운 관측치를 판별하여 개체를 분류함.

10.2.1 Fisher의 방법

두개의 그룹을 비교하고 판별할 수 있도록 변수들의 선형조합으로 판별함수를 구하고자함.

다변량벡터를 일변량 변수로 변환하여 그룹을 판별하는 방법. 정규성 가정은 필요하지 않음.

그룹을 판별하기위해 G_1 , G_2 를 가능한 한 떨어져 구별되게 하는 선형판별함수 $Y = \ell'X$ 를 구하고자함.

두 그룹은 공통 공분산행렬을 가짐

각 그룹에서 판별함수의 기대값과 분산, 분산을 고려한 두 그룹의 평균 차이, Fisher의 선형판별함수는 p259

표본에 대한 Fisher의 선형판별함수, 표본평균벡터, 표본공분산행렬, 합동공분산행렬은 p260

각 그룹의 평균점은 $\bar{y}_1 = \hat{\ell}'\bar{X}_1$, $\bar{y}_2 = \hat{\ell}'\bar{X}_2$ 이고 이들의 중간점은 $\hat{m} = \frac{1}{2}(\bar{y}_1 + \bar{y}_2)$ 이 됨.

새로운 관측벡터 X_0 에 대한 판별함수와 그 판별함수를 이용한 판별규칙은 p260

Fisher의 선형판별함수에 의해 새로운 y축이 형성되어 두 그룹의 중간점을 기준으로 두 그룹이 나누어 분포됨.

10.2.2 다변량 정규분포를 따르며 두 그룹의 공분산행렬이 같은 경우

$f(X|G_i)$: G_i 그룹으로부터 발생한 X 의 확률밀도함수

P_1 과 P_2 는 각각 X 가 G_1 , G_2 그룹에서 발생할 사전확률. ($P_2 = 1 - P_1$)

정리 10.1 오분류 확률을 최소화하는 최적 분류규칙 p263

만약 $P_1 = P_2$ 이면 최적 분류규칙 p263

이는 최대우도를 이용한 규칙이 됨.

특히 X 가 다변량 정규분포를 따를 경우. (p264)

확률밀도함수, 최적 분류규칙, 추정량 사용

새로이 관측된 벡터 X 가 속하는 그룹에 대한 판별규칙 => 선형 판별 규칙

만약 $P_1 = P_2$ 이면, 정규성으로부터 유도된 선형판별규칙은 Fisher의 판별규칙과 같아짐

그리고 선형판별규칙을 이용한 분류규칙은 군사적으로 최적인 판별규칙이 됨.

10.2.3 다변량 정규분포를 따르며 두 그룹의 공분산행렬이 다른 경우

우도비, 이차판별함수, 최적 분류규칙은 p266

표본을 이용할때 이차판별함수의 표본함수는 p266

표본함수를 이용한 분류규칙은 군사적으로 최적규칙은 아님

실제 판별분석에서 판별함수를 구하고자 할 때, 그룹들이 공동공분산행렬을 가진다하여

귀무가설을 기각하지 못할 경우에는 선형판별함수 사용하고

귀무가설을 기각하게 될 경우에는 이차판별함수를 선택할 수 있음.

10.3 세 개 이상 그룹의 판별

2019년 6월 16일 일요일 오전 3:11

그룹이 3개 이상인 경우 g 개 그룹의 차이를 가장 크게 하도록 하는 변수들의 선형결합식을 찾는것.

분산분석법 이용.

10.3.1 Fisher의 방법

각 그룹 평균은 $\bar{X}_i = \alpha' \bar{x}_i$ 이 됨.

각 그룹평균을 분리해내는 식의 기준을 g 개 그룹에 대해 확장하기 위하여 분산분석에서의 그룹간 행렬 B 와 그룹 내 행렬 E 를 이용해 표현함.

판별함수 $Z_1 = \alpha' Y$ 는 그룹평균들의 차이를 가장 크게 해주는, 즉 그룹을 구별해주는 함수가 됨.

새로운 개체에 대한 분류규칙은 p270

10.3.2 공분산행렬이 모두 같은 경우

확률밀도함수를 아는 경우 최적 판별규칙 p271

(만약 그룹에 대한 사전확률이 모두 같으면 위 판별규칙은 우도함수만을 이용하게 되므로 최대우도를 이용한 규칙이 됨)

x가 다변량 정규분포를 따를 경우

선형함수와 판별규칙은 p271

이 판별규칙은 근사적으로 최적인 분류규칙이 됨.

10.3.3 공분산행렬이 모두 다를 경우

x가 다변량 정규분포를 따를 경우

표본으로부터의 추정량을 이용한 이차형식함수, 판별규칙은 p272

이 판별규칙은 근사적으로 최적인 분류규칙이 됨.

그룹에 대한 사전확률이 모두 같으면 이차형식함수에서 $\ln p$ 가 사라짐

10.4 오분류율 계산

2019년 6월 16일 일요일 오전 3:11

분류함수의 능력을 판단하기 위한 오분류의 확률.

오류율과 정확한 분류율을 이용함.

10.4.1 재대입 분류에 의한 오류율 계산

데이터로부터 유도된 판별함수를 다시 데이터에 적용하는 재대입 분류에 의해 오분류율을 계산할 수 있음.

표를 만들어서 계산한 명백한 오류율, 정확한 분류율은 P_{274}

재대입 분류에 의한 오류율은 실제 bias 보다 적게 계산될 수 있음.

10.4.2 표본분할에 의한 오류율 계산

표본을 두 부분으로 나누어 훈련표본은 판별함수를 만드는데 이용하고 타당성 검사 표본은 만든 판별함수를 이용해 분류한 후 판별함수의 판별 능력을 평가함.

추정된 오류율은 불편추정량이 됨.

단점1 : 표본을 두 부분으로 나누어야하므로 비교적 큰 표본의 크기가 요구

단점2 : 실제로 사용할 판별함수에 대해서는 평가할 수 없음

10.4.3 교차타당성에 의한 오류율 계산

한개만의 표본을 제외한 나머지 표본으로 판별함수를 계산함

구해진 판별함수를 이용해 제외되었던 표본을 분류함.

전체 표본의 크기만큼 시행하여 오류율을 구함.

교차타당성에 의한 오류율은 재대입분류로 구한 오류율보다 커지는 경향이 있음.

10.5 10.6 판별함수의 표준화 & 유의성 검정

2019년 6월 16일 일요일 오전 3:11

판별함수에 기여하는 변수의 상대적 비중은 판별함수계수에 나타남.

변수값들의 단위가 다르면 동등한 비교 불가능

단위가 다를 때는 표준화변수를 이용하여 판별함수를 구한 후 판별함수의 계수를 비교함.

두 개의 그룹이 있으며 공분산 행렬이 같다고 할 수 있는 경우는 p276

여러 개의 그룹이 있으며 공분산행렬이 같다고 할 수 있는 경우는 p277

Fisher의 판별함수는 집단 또는 그룹간의 평균차가 최대가 되도록 판별함수를 구하는 것이며 그룹간의 평균 (벡터) 차이가 없다면 판별함수를 이용할 필요가 없게 됨 => 판별함수의 유의성 검정

판별함수의 유의성 검정에는 정규성 가정이 필요함.

10.6.1 두 개 그룹의 경우 판별함수의 유의성 검정 (p278)

두 그룹의 분리를 위해 두 그룹 간 거리가 최대가 되도록 만든 판별함수의 계수벡터

판별함수의 계수벡터의 유의성에 대한 귀무가설

두 그룹의 평균비교에 대한 가설검정은 Hotelling T-square 검정을 이용함.

10.6.2 여러 개 그룹의 경우 판별함수의 유의성 검정 (p278)

10.7 판별함수의 변수선택 : 단계별 변수 선택

2019년 6월 16일 일요일 오전 3:12

10.7.1 변수선택방법

적절한 변수선택을 통해 판별함수를 찾으면 판별차원을 줄일 뿐만 아니라 오분류율을 낮출 수 있음.

- 1) 전진적 선택 (forward selection) : 그룹간의 거리를 최대로 하는 한 개의 변수를 선택한 후 그 다음으로 그룹간의 거리를 최대로 하는 다른 한 개의 변수를 선택함. 변수가 더 이상 선택되지 않을때까지 진행함.
- 2) 후진적 선택 (backward selection) : 모든 변수로부터 시작함. 부분 F-검정통계량의 값이 가장 작은 변수부터 제거해가며 더 이상 변수가 제거되지 않을때까지 진행함.
- 3) 단계적 선택 (stepwise selection) : 전진적 선택 방법과 후진적 선택 방법의 혼합형. 변수가 선택되는 때 단계마다 선택된 변수가 기여하는 바를 검정하여 결정함. 각 단계에서 판별함수를 계산하지는 않으며 실질적으로 매단계 MANOVA 과정을 수행하는것. 변수 선택 과정이 끝나면 선택된 변수를 이용하여 판별함수를 구함. 이와 같이 변수를 선택하여 판별함수를 만드는 작업을 단계적 판별분석이라함.

10.7.2 부분 F-통계량을 이용한 단계별 변수 선택 (p282)

일변량에 대한 분산분석의 F-통계량을 비교하여 가장 큰 값을 갖는 변수가 첫번째로 선택됨.

선택된 변수에 대한 wilks lambda를 구함. 각 변수에 대해 가장 큰 부분 F-값을 갖는 변수를 다음으로 선택함.

더 이상 변수가 선택되지 않거나 제거되지 않는 단계에서 과정은 멈추게됨.