판별 분석

판별 분석

판별 분석은 두개 이상의 모집단에서 추출된 표본들이 지니고 있는 정보들을 이용하여 표본들이 어디에서 추출된건지 기준을 정하는 분석법을 말한다.

예를 들어 은행에서 채무자가 대출금을 갚을것인지 아닌지에 대한 여부를 판단하기 위해

과거에 대출금을 갚지 않은 사람의 정보 유형을 참고하여 담보 신청시 신청자의 정보 유형을 과거의 유형과 비교하여 판단할수 있겠습니다.

판별 변수의 주요 개념

1.판별 변수

판별 변수는 어떤 집단에 속하는지 판단 할때 사용되는 변수로서 독립변수중 판별력이 높은 변수를 뜻합니다.

판별 변수를 선택하는데 중요하게 생각해야 할것은 다른 독립변수들 간의 상관관계입니다.

상관관계가 높은 변수는 전부다 선택하지 않도록하고 상관관계가 적은것 끼리 판단해야 효과적인 판별함수를 만들 수 있습니다.

판별 변수의 주요 개념

2.판별 함수

판별함수는 판별변수들을 선형관계로 모은 것으로 집단의 수와 독립변수의 수중 작은 값만큼 도출할 수 있습니다. 판별 함수의 목적은 종속변수의 집단에 대한 예측력을 높이는데 있다. 판별분석이 이용되기 위해서는 각 개체가 어느 집단에 속해 있는지 알려져 있어야 하고 이에 대한 판별식을 만들어 판단하는 과정을 포함하게 된다.

3.판별점수

판별점수는 어떤 대상이 어떤 집단에 속하는지 판별하기 위하여 그 대상의 판별변수들의 값을 판별함수에 대입하여 구한 값을 뜻한다.

판별 분석의 단계

1.케이스가 속한 집단을 구분하는데 기여할수 있는 독립 변수를 찾습니다.

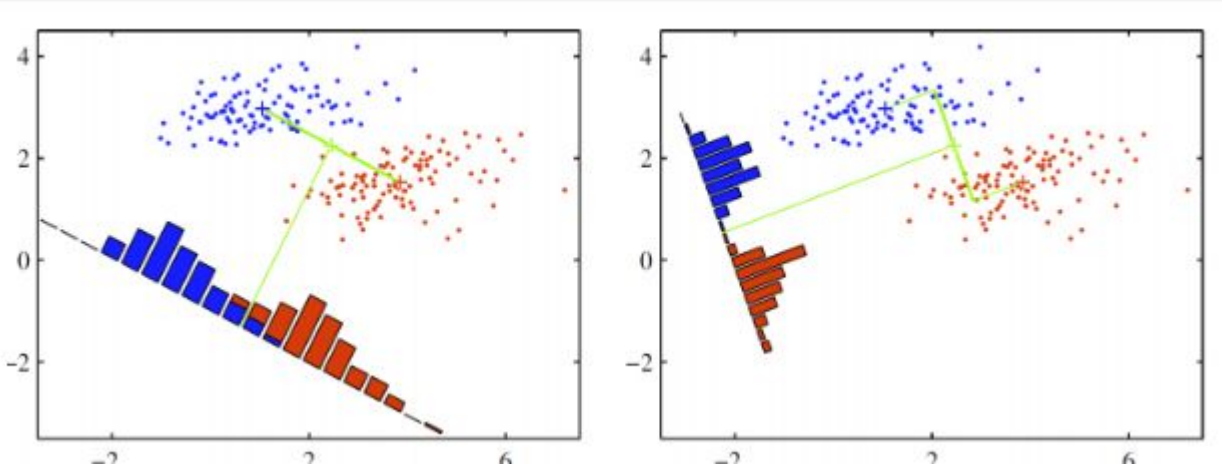
2.집단을 구분하는 기준이 되는 독립 변수들의 선형 결합 즉 판별 함수를 도출합니다

3.도출된 판별 함수에 의한 분류의 정확도를 파악합니다.

4.판별함수를 이용하여 새로운 케이스가 속하는 집단을 예측합니다.

판별 함수 예시

선형판별분석 (Linear Discriminant Analysis)



선형판별분석(Linear Discriminant Analysis)

그렇다면 두 범주를 잘 구분할 수 있는 직선은 어떤 성질을 지녀야 할까요?

사영 후 두 범주의 중심(평균)이 서로 멀도록, 그 분산이 작도록 해야할 겁니다. 왼쪽 그림을 오른쪽과 비교해서 보면 왼쪽 그림은 사영 후 두 범주 중심이 가깝고,

분산은 커서 데이터가 서로 잘 분류가 안되고 있는 걸 볼 수가 있습니다.

반대로 오른쪽 그림은 사영 후 두 범주 중심이 멀고, 분산은 작아서 분류가 비교적 잘 되고 있죠. **LDA**는 바로 이런 직선을 찾도록 해줍니다.

선형판별분석 (Linear Discriminant Analysis)

$$y = \overrightarrow{w}^T \overrightarrow{x}$$

$$m_1 = \frac{1}{N_1} \sum_{n \in C_1} x_n$$

$$m_2 = \frac{1}{N_2} \sum_{n \in C_2} x_n$$

p차원의 입력벡터 **x**(변수 **p**개) 를 **w**라는 벡터(축)에 사영시킨 후 생성되는 **1**차원상의 좌표값(스칼라)를 아래와 같이 **y**라고 정의합니다. 각각 **N1**개와 **N2**개의 관측치를 갖는 **C1**과 **C2** 두 범주에 대해 원래 입력공간(**2**차원)에서 각 범주의 중심(평균) 벡터도 아래와 같이 **m1, m2**라고 정의합니다.

선형판별분석(Linear Discriminant Analysis)

첫번째로 사영후 두 범주의 중심이 멀리 떨어지도록 하는 벡터를 찾아야합니다

m2−m1 = wT * (m2−m1)

mk = wT * mk

사영후 각 범주안에 속하는 집단안의 분산은 작을수록 좋습니다. 분산을 구하는 식은 다음과 같습니다

$$s_k^2 = \sum_{n \in C_k} (y_n - m_k)^2$$

선형판별분석(Linear Discriminant Analysis)

앞에 두 과정을 동시에 진행 할려면 다음과 같은 과정을 거칩니다.

두 범주 중심을 분자, 두 범주의 분산을 분모에 넣고 이 식을 최대화

$$J(w) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2} = \frac{w^T S_B w}{w^T S_W w}$$

$$S_B = (m_1 - m_2)(m_1 - m_2)^T$$

$$S_W = \sum_{n \in C_1} (x_n - m_1)(x_n - m_1)^T + \sum_{n \in C_2} (x_n - m_2)(x_n - m_2)^T$$

선형판별분석(Linear Discriminant Analysis)

목적함수 J(w)은 w에 대해 미분한 값이 0이 되는 지점에서 최대값을 가집니다. 아래 식과 같습니다.

$$(w^T S_B w) S_W w = (w^T S_W w) S_B w$$

약간의 식을 정리하면 다음과 같은 식이 나옵니다\

$$S_W w = \lambda S_B w$$
$$S_B{}^{-1} S_W w = \lambda w$$

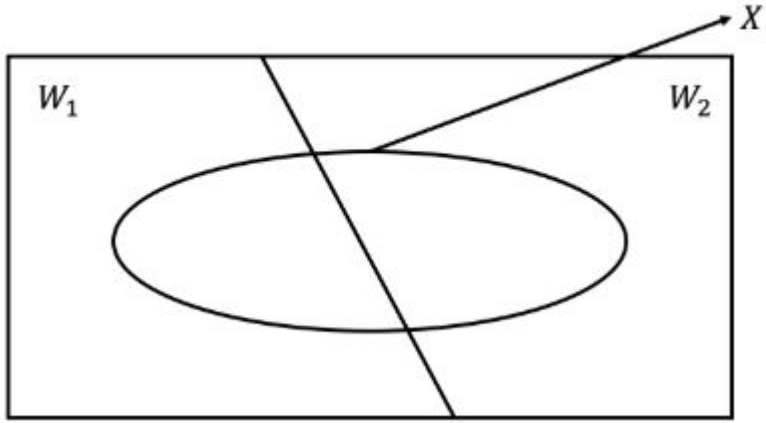
새로운 축 w는 SB의 역행렬과 SW를 내적한 행렬의 고유벡터라는 이야기 입니다.

선형판별분석(Linear Discriminant Analysis)

새로운 데이터(**x**)가 주어지면 이를 **w**와 내적해 각각의 스코어를 낼 수 있습니다. 그 스코어가 일정값보다 크면 C_1범주, 작으면 C_2 범주로 분류를 하게 됩니다.

선형 판별 분석

베이즈 정리를 이용한 선형판별분석 접근



$$P(W_i|x) = \frac{P(x|W_i)P(W_i)}{P(x)}$$
$$= \frac{P(x|W_i)P(W_i)}{P(x|W_1)P(W_1) + P(x|W_2)P(W_2)}$$

선형 판별 분석

새로운 데이터 **x**가 등장하면 **LDA** 모델은 **P(W1|x)**와 **P(W2|x)**를 각각 구합니다.
둘 중 전자가 크면 **W1**으로, 후자가 크면 **W2**로 분류를 하게 되는 방식입니다.
사후확률인 **P(Wi|x)**는 새로운 데이터 **x**가 주어졌을 때(=정답 범주를 모를 때=예측할 때) **Wi**일 확률, 즉 검증데이터가 특정 범주에 할당하기 위한 스코어를 의미합니다.

범주 분류를 위한 확률(스코어)를 내어주는 함수를 판별함수(**discriminant function**)라고 합니다.

LDA의 중요한 가정은 데이터 분포가 다변량 정규분포(**multivariate normal distribution**)을 따른다는 사실입니다.

많은 자연, 사회현상이 정규분포를 따르고 있고 그 예측력 또한 많은 실험을 통해 검증된 연속확률분포입니다. 다변량 정규분포의 파라메터는 평균(벡터)와 공분산(행렬)입니다. 이 두 파라메터와 정규분포 확률함수로 **P(x|WiWi)**를 구할 수 있고, 이를 우리가 이미 알고 있는 사전확률 **P(Wi)**과 함께 계산해 범주를 판별한다는 것이 **LDA**를 베이지안 관점에서 접근하는 방식의 핵심입니다.