개념

다변량 데이터 정의 Definition

2개 이상 확률변수는 열, 행은 관심 개체의 확률변수의 관측치로 구성된 다변량 데이터 형태이다.

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

- 다변량 분석에 사용되는 확률변수는 정량(측정)변수를 원칙으로 한다.
- 다변량 분석은 함수모형에 의한 예측, 변수의 유사성으로 변수 차원을 축소하거나 개체의 유사성에 의해 개체를 분류하므로 정량 데이터 형태만 가능하다.
- 예측모형의 경우 정성(질적, 범주)변수의 활용이 가능하다.

분석 개념

인과 관계 [변수 관계]

인과 관계(casual relationship)에서 원인이 되는(영향을 주는) 변수를 설명 변수(exploratory var.) 혹은 독립 변수(independent var.)라 하고 결과나 영향을 받는 변수를 종속 변수(dependent var.) 혹은 반응 변수(response var.)라 한다. 종속 변수는 Y, 설명 변수는 X로 표시한다. 분산 분석에서는 설명 변수를 처리 효과, 요인으로 불리어진다.

인과 관계는 이론적, 경험적 타당성에 근거하여 연구 목적에 설정되는 것이지 자료 분석 후 인과 관계가 설정되는 것은 아니다.

상관관계 [변수 관계]

- 변수들간의 상관 구조(공분산 구조)를 활용하여 변수들 간의 함수(직선) 관계를 탐색하거나
- 변수들간 유사성(상관계수 크기)를 이용하여 변수의 차원을 줄이는 방법이다.
- 빅데이터에서는 차원축소(diemnsion reduction) 방법으로 불리는데 열(변수)의 차원을 줄이는 방법 뿐 아니라 행 차원도 줄이는 SVM 방법이 있다.
- 고전적인 다변량 차원축소는 열의 차원만 축소하였지만 데이터의 양이 많아지는 빅데이터에서는 행의 차원도 줄이는 기법이 적용되고 있다.

개체의 유사성

개체들의 유사성을 변수로 측정하여 유사한 개체끼리 분류하는 방법이다. 사전 정의된 그룹에 의해 판별모형을 유도하고 새로운 개체 판별하는 판별 discriminant 과 그룹의 사전 정의 없이 분석자가 임의의 그룹(개수 포함)으로 나누는 분류 classification이다.

다변량 분석

예측모형 prediction model

종속변수 Y, 설명변수 X's 함수관계에 대한 탐색한다. $\underline{y} = X\underline{b} + \underline{e}$, 평균벡터 $\underline{\mu} = X\underline{b}$

종속변수 데이터가 가질 수 있는 확률분포 형태는 다음과 같다.

[위키피디아]

Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\boldsymbol{\beta} = g(\mu)$	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$	$\mu = \mathbf{X}\boldsymbol{\beta}$
Exponential Gamma	real: $(0, +\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-1}$	$\mu = -(\mathbf{X}\boldsymbol{\beta})^{-1}$
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$	$\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{n-\mu}\right)$	
Categorical	integer: $[0, K)$ K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1	outcome of single K-way occurrence		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences			

Probit Model, Predicted Probabilities and Estimated Effects — Key Concept 11.2

Assume that Y is a binary variable. The model

$$Y = \beta_0 + \beta_1 + X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u$$

with

$$P(Y = 1|X_1, X_2, \ldots, X_k) = \Phi(\beta_0 + \beta_1 + X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)$$

Logit Regression — Key Concept 11.3

The population Logit regression function is

$$P(Y=1|X_1,X_2,\ldots,X_k)=F(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k)$$
$$=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k)}}.$$

차원축소 dimension reduction

행렬 데이터 $X_{n\times p}$의 행 차원과 열 차원을 변수의 유사성 혹은 축소한다.

PCA

원데이터 행렬		주성분 행렬		차원축소	
$X_{N\times P}$ N=300(5분*1초) P=199(측정변수 개수)	=	$Y_{N\times P}=L'_{N\times P}X_{N\times P}$ L : 부하행렬	분해	$Y_{N\times K}$ (분석행렬) 주요주성분 (80%규칙) K=17(주성분 개수)	$Y_{N\times(P-K)}$ 잔차주성분

SVM

원데이터 행렬		분해		차원축소 방법
$X_{T\times P}$	=	$U_{T\times T}\Sigma_{(T\times P)}V_{P\times P}$	분해	1) THIN SVD 2) Compact SVD 3) truncated SVD
$X^*=(X_{T\times P}-\overline{X})$ 단위 표준화 필요		U : 왼쪽 특이값 직교행렬 Σ : 특이값 대각행렬 V : 오른쪽 특이값 직교행렬		

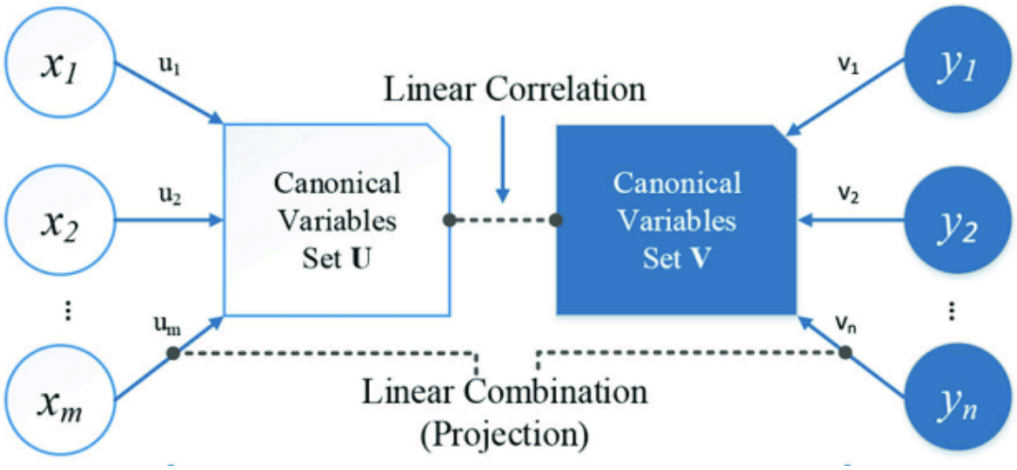
[PCA vs. SVD 차이]

PCA 차원축소는 원데이터의 공분산행렬(공분산구조)을 이용하지만 SVD 방법은 원데이터를 직교행렬, 특이값 대각원소로 분해하는 방법

축약된 주성분변수는 부하 크기에 의해 변수의 속성을 파악할 수 있으나 SVD는 가능하지 못함

정준상관분석 canonical correlation

서로 다른 변수 군으로부터 상관계수 0(서로 독립) 변수군을 만들어 변수군간 상관분석을 실시한다. 변수의 개수를 줄이는 한 방법이다.



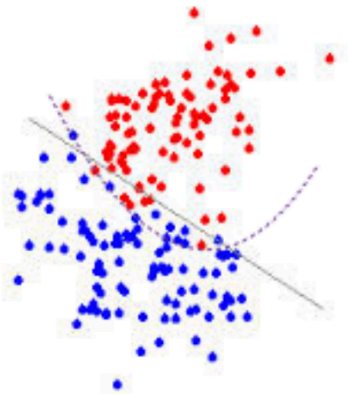
판별 및 분류

개체 판별 discrimination

판별분석은 자료 수집 시 이미 그룹이 나누어져 있어(빨간원, 파랑원 개체) 이를 가장 잘 판별하는 판별규칙(측정변수의 거리에 의해 개체의 유사성을 측정함)을 도출하여 새로운 개체의 군집을 판별하는 방법이다.

개체분류 classification

군집분석에서는 개체의 그룹에 대한 정보 없이 (사전에는 △□◎ 구분이 없이 동일하나 분석 후 나누어 짐) 유사성이 가까운 개체들끼리 계층적으로 묶어 가거나 군집의 개수를 정하여 군집의 중심점을 이용하 여 개체를 군집화 하는 방법이다.



판별분석

군집분석

다변량 분석 정의

	주성분	요인	판별	군집	정준상관
변수 관계 탐색	S	D	N	N	S
자료 탐색	D	S	N	S	N
새 변수 만들기	Yes	Yes	No	No	Yes
개체 분류	No	No	Yes	Yes	No
변수 그룹	P	P	N	N	D
차원 줄이기	D	P	N	N	N

Sometimes, Definitely, Never, Possible, Rarely

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{21} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{n1} \end{bmatrix}$$

개체분류

- 군집분석 clustering analysis : 유사한 개체들을 동일 군집 분류
- 판별분석 discriminant analysis : 새로운 개체를 가장 적절한 특정 집단으로 분류

차원축약

- 주성분분석 principle component analysis : 변수들의 상관관계를 활용하여 변수 개수 축약 – 새로운 변수, 주성분 변수
- 요인분석 factor analysis : 유사한 변수들을 군집으로 분류