

통계학개론 REPORT

6팀



과목 : 통계학개론

교수님 : 심재창 교수님

학과 : 컴퓨터공학과, 창의융합학부

조원 : 20191128 이재경(조장), 20191119 김소망

20191362 양운태, 20191129 이주협 2011 서현직

1. 군집분석

정의

군집 분석은 비지도학습(Unsupervised learning) 분석 기법 중 하나입니다. 쉽게 말해서, 사전 정보 없이 자료를 컴퓨터에게 주고, “유사한 대상끼리 묶어보아라!” 라고 명령을 내리는 분석 방법입니다. 따라서 군집분석에서는 어떤 변수를 컴퓨터에게 입력하느냐가 중요하다고 볼 수 있습니다.

예시

K-MEANS 기법

컴퓨터에 할당된 자료를 K개의 군집으로 묶는 알고리즘으로 각 무리의 거리 차이의 분산을 최소화하는 방식으로 작동된다.

변수를 통해 컴퓨터가 직접 무리를 묶는 것이기 때문에 구하고자 하는 데이터의 변수를 올바르게 설정하는 것이 가장 중요하고 변수의 단위를 표준화하는 것 또한 중요하다.

K-MEANS 기법이 활용되는 예시로는 서울 각지에 분포되어 있는 A마트에 들어가는 제품들을 관리할 때 그 제품들을 납품하기 전 필요한 장소의 수와 위치를 K의 무리로 묶는다고 생각하면 된다. 만약 K를 1로 설정할 경우 창고를 모든 A마트 위치의 중앙에 두는 것이 이동 및 납품에 용이하고 만약 창고를 세 군데 두고 싶다면 K의 값을 3으로 설정하여 알고리즘을 돌려 물류를 쉽게 유통할 수 있는 최적의 장소를 세 군데에 창고를 설치할 수 있게 하는 것과 같다.

2. 판별분석

정의

분류모형을 구축할 때 사용하는 **선형 지도학습 분류법(Supervised Claafier)**이다.

자료의 그룹 정보를 가장 잘 구별해낼 수 있는 **설명변수의 선형결합**을 찾는데 사용된다.

소속집단을 알고 있는 데이터를 이용하여 모형을 만들어 소속집단을 모르는 한 데이터의 집단을 결정하는 기법이다.

예시

선형판별분석(Linear Discriminant Analysis, LDA)

자료의 그룹 정보를 가장 잘 구별해낼 수 있는 **설명변수의 선형결합**을 찾는데 사용된다.

분류모형을 구축할 때 사용하는 **선형 지도학습 분류법(Supervised Claafier)**이다.

선형판별분석의 예시로는 온타리오 호수에 사는 어종에 대한 자료를 통해서 (ex 블루길, 아미아, 잉어, 골드아이, 입큰송어 등을 학습시킨 후)

fish.data에서 100마리를 추출하여 물고기 종류 별로 선형판별하는 방법이 있다.

3. 우리들이 생각하는 군집분석과 판별분석의 공통점과 차이점

군집분석과 판별분석의 공통점은 클러스터링, 즉 무리로 묶어서 분석한다는 점이 있다.

차이점으로는 가장 큰 특징이 지도학습이나 비지도학습이냐가 가장 중요한 부분인 것 같은데

군집분석의 경우에는 컴퓨터가 사전정보 없이 데이터를 직접 무리로 나누어서 유사한 대상으로 클러스터링한다는 점이다.

컴퓨터가 직접 학습을 하여 클러스터링을 하기 때문에 변수를 올바르게 설정하는 등 적절한 방향성을 갖고 변수를 설정하는 것이 중요하고 또 변수의 단위의 오차가 발생하면 안되기 때문에 변수의 단위 통일 또한 중요하다고 생각한다.

판별분석의 경우에는 추출된 표본들이 가지고 있는 정보들을 통해서 무리로 묶어서 이 표본들이 어느 집단에 속해 있는 것인지를 결정해줄 수 있다는 점이 가장 큰 특징인 것 같다.

소감

심재창 교수님의 가르침 덕에 R언어의 흥미와 희열을 느꼈습니다. 군집분석과 판별분석이 잘 이해가 가지 않았지만 6팀과 함께 서로 모르는 부분을 보완해주고 알려주니 혼자 학습하는거 보다 배의 학습효과를 얻었습니다. 이렇게 재밌는 통계 수업을 들을 수 있어 안동대학교에 잘 왔다고 생각합니다. 교수님 사랑합니다♥