Final Exam: Multivariate Data Analysis (20636)

June 17, 2020 11:00 AM ~ June 18, 2020 13:00 (1:00 PM)

* Your data file (with your ID) contains data on breakrast cereals produced by difference American manufacturers. (One file is for analysis and the other one (with "–new") is for the prediction in question 4-d)
* You should answer each question as much detail as possible.
* You should write or type your answers. (Do not simply copy your R result.)
* R code and results should be included in the last part of the answer.
* You should turn in before **13:00 (1:00 PM), June 18, 2020**

1. Explore your data.

2. Test whether each manufacturer's means of 11 continuous variables are same or not.

3. a) Find principal components using the variance-covariance matrix, select *m*, and explain your *m* principal component. You should justify your choice of *m*.
b) Find principal components using the correlation matrix, select *m*, and explain your *m* principal component. You should justify your choice of *m*.
c) Compare the result in 3-a) and 3-b). Explain the differences between the results in details.

4. Assume that the variance-covariance matrices are all same in each manufacturer.
a) Find discriminants and describe a rule to classify manufacturers using discriminants.
b) Plot the discriminant scores in 2-dimensional discriminant space using different plotting symbols to identify manufacturers and explain your plot.
c) Predict manufacturers using your rule and compare the predicted manufacturers to the original manufacturers
d) With your classification rule, predict the manufacturer with data file start with "new".

5. a) Find 3 clusters with the hierarchical clustering - the complete linkage using the Euclidean distance and draw dendrogram.
b) Find 3 clusters using K-means algorithm and represent the result in the scatter plot with the first two principal components from 3-b)
c) Compare the result in 5-a) and 5-b)

Final Exam: Multivariate Data Analysis (20636)

June 17, 2020 11:00 AM ~ June 18, 2020 13:00 PM

* 각자의 학번이 들어있는 파일에는 미국의 다양한 제조사의 시리얼에 대한 정보가 들어있습니다.
두개의 파일 중 하나는 자료분석을 위한 것, 그리고 "-new"가 있는 파일은 문제 2-d)를 위한 것입니다.
* 각 문항의 답을 가능한한 자세히 적으시오.
* 각 문항에 답을 할 때에는 R 결과를 그대로 복사하지 말고 정리해서 쓰거나 typing 하시오.
* 각 문항의 답 밑에는 R code와 결과가 포함되어 있어야 함.
* **6월 18일 오후 1시까지 제출하시오.**

1. 자료를 살펴보시오

2. manufacturer 별로 11개의 연속변수들 평균이 같은지 다른지를 검정하시오.

3. a) 분산공분산 행렬을 이용하여 주성분을 찾으시오. m을 결정하고 m개의 주성분을 설명하시오. (m을 선택한 이유도 명시할 것)
b) 상관 행렬을 이용하여 주성분을 찾으시오. m을 결정하고 m개의 주성분을 설명하시오. (m을선택한 이유도 명시할 것)
c) 3-a) 과 3-b)의 결과를 비교하시오. 두 결과의 차이점에 대하여 상세히 설명하시오.

4. 각 manufacturer 별로 분산공분산행렬이 같다고 가정하자
a) manufacturer를 분류하기 위한 판별함수를 찾고 분류 규칙을 쓰시오
b) 2차원의 판별함수값을 이용하여 산점도를 그리고 설명하시오. (단, 산점도위에 manufacturer를 다르게 표시할 것)
c) 4-a)에서 찾은 규칙을 이용하여 manufacturer를 예측하고 자료의 manufacturer과 비교하시오.
d) 4-a)에서 찾은 규칙을 이용, "-new"가 있는 파일의 자료에 대한 manufacturer를 예측하시오.

5. a) 유클리디안 거리와 complete linkage를 이용하여 3개의 cluster를 찾으시오. Dendrogram도 그릴것.
b) K-means algorithm을 이용하여 3개의 cluster를 찾고 3-b)에서 구한 처음 2개의 principal component의 산점도 위에 3개의 cluster를 나타내시오.
c) 5-a) 와 5-b)의 결과를 비교하시오.