Assignment 2 Multivariate Statistics 2023-2024

Task 1

Description of the data

For this task we use a selection of the MNIST data which are available on

http://yann.lecun.com/exdb/mnist/index.html

The file **mnist_task1.Rdata** contains the datasets train.data, train.target, test.data and test.target.

The data sets train.data and test.data each consist of 5000 28 x 28 images of 10 handwritten digits (0,1,2, …,9). The data sets train.target and test.target contain the class labels for cases in train.data and test.data, respectively.

Description task

a) Conduct principal components analysis on the covariance matrix of the training data (i.e. centered variables) and select the number of components so that the components account for 80% of the variance in the training data (=scenario 1) or for 90% of the variance in the training data (=scenario 2).

b) Compute the training and test error of the following classifiers

- LDA conducted on unstandardized principal components of scenario 1
- LDA conducted on unstandardized principal components of scenario 2
- QDA conducted on the unstandardized principal components of scenario 1
- QDA conducted on the unstandardized principal components of scenario 2
- KNN conducted on the unstandardized principal components of scenario 1
- KNN conducted on the unstandardized principal components of scenario 2
- Random Forest conducted on unstandardized principal components of scenario 1
- Random Forest conducted on unstandardized principal components of scenario 2
- HDDA conducted on all the centered variables in train.data using the common dimension models “AKJBKQKD” and “AKJBQKD” and in which you select the number of components using the method of Cattell with threshold=0.05.

Try to select the tuning parameters of the classifiers KNN and Random Forest so that test error is as low as possible.

c) Make an overview table that includes the training and test error of each classifier and visualize the results. Discuss the results of the analysis.

Task 2

Description of the data

For this task we use a selection of the MNIST data which are available on

http://yann.lecun.com/exdb/mnist/index.html

The file **mnist_task2.Rdata** contains the datasets data and target.

The data set **data** contains of 2000 28 x 28 images of 4 handwritten digits (0,1,2, and 3) and the data set **target** contains the corresponding class labels.

Description task

Center the variables of **data**.

(a) Investigate (using the Adjusted Rand Index) to what extent you can recover the true class labels using unsupervised clustering techniques on the data if you extract 4 clusters. Investigate the performance of the following methods to recover true class labels:

- Hierarchical clustering on squared Euclidean distances using the method of Ward
- K-means clustering
- common dimension hddc() models "AkjBkQkD" and "AkjBQkD" for which you select the number of components using the method of Cattell with threshold=0.05 and in which you use as starting point (1) the solution obtained with hierarchical clustering and (2) the solution obtained with k-means clustering (see (a)).

(b) Visualize the observed and predicted class labels for the clustering model with the best performance (i.e., the highest value of the Adjusted Rand Index) in the space of the first two principal components.

Discuss the results of the analysis.

Task 3

In a study 101 first-year psychology students indicated whether or not they would display each of 8 anger-related behaviors when being angry at someone in each of 6 situations (Kuppens et al., 2004). The 8 behaviors consist of 4 pairs of reactions that reflect a particular strategy to deal with situations in which one is angry at someone, namely, (1) fighting (fly off the handle, quarrel), (2) fleeing (leave, avoid), (3) emotional sharing (pour out one's heart, tell one's story), and (4) making up (make up, clear up the matter). The six situations are constructed from two factors with three levels: (1) the extent to which one likes the instigator of anger (like, dislike, unfamiliar), and (2) the status of the instigator of anger (higher, lower, equal). Each situation is presented as one level of a factor, without specifying a level for the other factor.

The package plfm (available at CRAN) contains the data. For a description of the data use help(**anger**). The binary person x situation x behavior array **anger$data** indicates whether or not a person would display a certain behavior in a certain situation. The situation x behavior matrix **anger$freq1** indicates how many persons would display a certain behavior in a certain situation.

Description task

a. Aggregate the binary person x situation x behavior array **anger$data** across situations and compute a person x behavior frequency matrix with elements that indicate how often a person would display a certain behavior across the 6 situations. Conduct hierarchical clustering with the method of Ward on squared Euclidean distances to cluster persons in the person x behavior frequency matrix. Save the cluster membership variable of the 2-cluster solution and interpret the centroid of the clusters.
b. Compute for each cluster a profile vector that indicates how often persons in the cluster would display a certain behavior. Create a frequency matrix in which you add the profile vectors of the two clusters as extra rows to the situation x behavior matrix **anger$freq1**.
c. Conduct a correspondence analysis on the created frequency matrix to analyze the situation x behavior associations. Consider the rows with the profile vectors of the clusters as supplementary row points. Discuss the results of the correspondence analysis.
d. Visualize the situations, behaviors and clusters in a two-dimensional biplot and discuss what you can conclude from the biplot.

Reference

Kuppens, P., Van Mechelen, I., and Meulders, M. (2004). Every cloud has a silver lining: Interpersonal and individual differences determinants of anger-related behaviors. *Personality and Social Psychology Bulletin, 30*, 1550-1564.

Submission of the assignment

For this assignment, one member of each team should upload the following files on Toledo:

- Report that includes the solutions of the tasks (word document or .pdf file). The length of the report is limited to **maximum 15 pages (including one title page)**.
- Script File with the R-code (.R file)

Report with solutions of the tasks

For each task show the R-code followed by the relevant analysis output or graphs generated by R, and discuss in sufficient detail the results of the analysis. Remark: include the R-code and R-output using an appropriate font (e.g., courier) and layout.

Script file with R-code

- Include for each task all the R-code
- Add comments to the R-code
- Write the code so that it can be used to replicate all the reported analyses