Assignment 2 Multivariate Statistics 2025-2026

Task 1

Description data

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

https://www.kaggle.com/datasets/crawford/emnist

The file **task1.Rdata** contains 6 data sets.

The data set **train.data.s1** consists of 9600 28 x 28 images of 4 letters (D, G, O, Q), and the data set **train.target.s1** contains the corresponding class labels.

The data set **train.data.s2** consists of 1600 28 x 28 images of 4 letters (D, G, O, Q), and the data set **train.target.s2** contains the corresponding class labels.

The data set **test.data** consists of 1600 28 x 28 images of 4 letters (D, G, O, Q), and the data set **test.target** contains the corresponding class labels.

Note that "s1" and "s2" respectively represent scenarios where we have a larger or a smaller training set available, and that all data sets are balanced as class labels have equal proportions.

Description task

a) For each scenario, 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 90% of the variance in the training data.
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
- Multinomial logistic regression on the unstandardized principal components of scenario 1
- Multinomial logistic regression on the unstandardized principal components of scenario 2
- Multinomial logistic regression on the unstandardized principal components and the squared unstandardized principal components of scenario 1

- Multinomial logistic regression on the unstandardized principal components and the squared unstandardized principal components of scenario 2
- Gradient boosting conducted on unstandardized principal components of scenario 1
- Gradient boosting conducted on unstandardized principal components of scenario 2
- HDDA conducted on all the centered variables in the training set of scenario 1 and using the common dimension model "AKJBKQKD" in which you select the number of components using the method of Cattell with threshold=0.05.
- HDDA conducted on all the centered variables in the training set of scenario 2 and using the common dimension model "AKJBKQKD" in which you select the number of components using the method of Cattell with threshold=0.05.

Use all available training observations of the scenario to estimate training error. Use all the observations in the test set to estimate the test error.

For LDA and QDA use the lda() and qda() functions of the `MASS` package. For KNN use the knn() function of the `class` package. For multinomial logistic regression, use the multinom() function of the package `nnet`. For Gradient boosting use the functions xgb.cv() and xgb.train() of the `xgboost` package. For HDDA use the hdda() function of HDclassif package. Select the tuning parameter K of KNN so that the test error is as low as possible. Select meaningful tuning parameters for the gradient boosting.

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

Task 2

Description data

The file **beer.Rdata** contains a ranking of 12 beers for 399 persons. A ranking of 1 means that the person prefers the beer most, whereas a ranking of 12 indicates the person prefers the beer least.

Description task

a) Use a hierarchical clustering on squared Euclidean distances using the Method of Ward to cluster the 12 beers. Draw the dendrogram and discuss the results.
b) Split the data set in a training set and a validation set by assigning odd-numbered observations to the training set and by assigning even-numbered observations to the validation set. Compute, for the training set and for the validation set, the following clustering solutions with 3 and 4 clusters
    - Hierarchical clustering on squared Euclidean distances using the method of Ward

- The best k-means solution using 100 random starting points
- Hierarchical clustering on squared Euclidean distances using the method of Ward followed by k-means using the centroids of the hierarchical clustering as a starting point

Next compute, for each of the 6 solutions (i.e., 3 methods x 2 values for the number of clusters), the stability of the cluster solution in a split half-approach using the adjusted rand index as criterion.

Apply the clustering method with the highest value of the adjusted rand index to the entire 399 x 12 matrix to obtain a final clustering solution.

c) Conduct two-dimensional ordinal row-conditional unfolding on the rankings of the beers. Visualize the final clustering solution in the obtained unfolding configuration by using colored dots for the persons. Discuss what you can conclude about the preferences of the clusters from the configuration plot and from the centroids of the clusters.

Submission of the assignment

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

- Report with answers to questions of both 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)
- Completed code of conduct on the use of genAI for each team member

Report with answers to the questions

For each question 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 to answer the question. In case you use similar R-code on different data sets (e.g. scenario 1 and 2 in task1, clustering solutions with different numbers of clusters), it is sufficient if you include the code for one of the cases in the report. Include the R output using an appropriate font (e.g., courier) and layout.

Script file with R-code

- Include for each question all the R-code of the fitted models
- Add comments to the R-code
- Write the code so that it can be used to replicate all the reported analyses
- If certain lines of the code were generated using GenAI, add a comment about this.