ST514/DS805 Multivariate Statistical Analysis
Poster Presentation Project

Instructions

This poster (size A1) should be submitted per group via itslearning as a PDF file before 14/05/2023.

In this activity, it is the intention to analyse a chosen data set with the techniques discussed in the course, with as ultimate objective the development of a good classification model. The content in the poster could include the following ingredients:

1. A brief summary of the data (among other things: background information, attributes (which can be served as categorical and which can be potential classifiers)) etc.
2. Check necessary assumptions, e.g. normality, homogeneity of covariance matrices
3. Data transformation if necessary
4. Selection of optimal classification rule
5. An additional classification rule for further comparison
6. Evaluation of the classification rules proposed. You may choose to use APER and/or $\hat{E}(AER)$ and/or ROC.
7. Selection of classifier.
8. Conclusion.

Data sets

- Your own data set that is suitable for the classification purpose
- Multiple sclerosis data, AMSA, Johnson and Wichern (see exercise 1.14 and exercise 11.23)

- Crude oil data, AMSA, Johnson and Wichern (see exercise 11.30)
- Data on Brands of Cereal (see exercise 11.34 and Table 11.9)
- Real estate sales data provided in realestate.txt. Further information see screenshot provided in Figure 1.
- Breast tissue data, UCI Machine Learning Repository,
  `http://archive.ics.uci.edu/ml/datasets/Breast+Tissue`

  Restrict the number of classes to four, as described on the above webpage.

Data Set C.7 Real Estate Sales

The city tax assessor was interested in predicting residential home sales prices in a mid-western city as a function of various characteristics of the home and surrounding property. Data on 522 arms-length transactions were obtained for home sales during the year 2002. Each line of the data set has an identification number and provides information on 12 other variables. The 13 variables are:

Variable Number	Variable Name	Description
1	Identification number	1–522
2	Sales price	Sales price of residence (dollars)
3	Finished square feet	Finished area of residence (square feet)
4	Number of bedrooms	Total number of bedrooms in residence
5	Number of bathrooms	Total number of bathrooms in residence
6	Air conditioning	Presence or absence of air conditioning: 1 if yes; 0 otherwise
7	Garage size	Number of cars that garage will hold
8	Pool	Presence or absence of swimming pool: 1 if yes; 0 otherwise
9	Year built	Year property was originally constructed
10	Quality	Index for quality of construction: 1 indicates high quality; 2 indicates medium quality; 3 indicates low quality
11	Style	Qualitative indicator of architectural style
12	Lot size	Lot size (square feet)
13	Adjacent to highway	Presence or absence of adjacency to highway: 1 if yes; 0 otherwise

1	2	3	4	5	6	7	8	9	10	11	12	13
1	360000	3032	4	4	1	2	0	1972	2	1	22221	0
2	340000	2058	4	2	1	2	0	1976	2	1	22912	0
3	250000	1780	4	3	1	2	0	1980	2	1	21345	0
...	...	...	...	...	...	...	...	...	...	...	...	...
520	133500	1922	3	1	0	2	0	1950	3	1	14805	0
521	124000	1480	3	2	1	2	0	1953	3	1	28351	0
522	95500	1184	2	1	0	1	0	1951	3	1	14786	0

Figure 1: Information of real estate sales data.