Assignment 2 Multivariate Statistics

Task 1

Description of the data

The data set for this task (**dtrust.Rdata**) is extracted from Round 6 of the European Social Survey (ESS Round 6, 2012). Besides the variable "country", the data set includes the responses of 4839 persons of three countries (The Netherlands, Poland and United Kingdom) on 10 items. Table 1 lists the items that are included in the data set.

The items in columns 2-8 measure on a 10-point scale to what extent persons have trust in institutions (a higher score indicates more trust). The items in columns 9-11 measure to what extent persons are satisfied with the present state of the economy, the national government and the way democracy works in the country (0=extremely dissatisfied, …, 10=extremely satisfied).

Table 1

variable	label
country	country
trust_cntryparliament	Trust in country's parliament
trust_legalsystem	Trust in the legal system
trust_police	Trust in the police
trust_politicians	Trust in politicians
trust_politicalparties	Trust in political parties
trust_EUparliament	Trust in the European Parliament
trust_UN	Trust in the United Nations
satisf_economycntry	How satisfied with present state of economy in country
satisf_nationalgovernment	How satisfied with the national government
satisf_democracycntry	How satisfied with the way democracy works in country

Question

Use the candisc() function (on all observations) to investigate to what extent you can classify respondents in their country using the 10 standardized items as predictors. Discuss the output of this canonical discriminant analysis. Plot the observations of the three countries and the 10 standardized items in discriminant space, and interpret the plot.

Task 2

Description of data

The dataset **Caravan.Rdata** (van der Putten & Van Someren, 2000) includes 84 variables that describe sociodemographic characteristics and product ownership for 5822 customers.

The dependent variable "Purchase" equals 1 if the customer purchased a car insurance policy and 0 otherwise. Note: the original data set includes 86 variables of which 2 variables were omitted. For more information about the attributes included in the data set, see https://liacs.leidenuniv.nl/~puttenpwhvander/library/cc2000/data.html.

Question

Investigate to what extent you can predict whether a person will purchase a car insurance policy using the 83 centered predictor variables. Compare the performance of the following classifiers: (1) LDA, (2) PCA+ LDA, (3) QDA, (4) PCA+QDA, (5) HDDA, (6) bagging, and (7) random forests using two classification scenario's:

(a) Account for different prior probabilities and equal classification costs

(b) Account for different prior probabilities and unequal classification costs: C(no purchase |purchase)=20*C(purchase|no purchase)

Make an overview table that includes training and test classification performance (hit rate, sensitivity, specificity) for each classifier and for the two classification scenario's and discuss the results of the analysis. Remark: use LOOCV predictions to estimate the test performance for LDA, PCA+LDA, QDA, PCA+QDA, HDDA and use OOB predictions to estimate the test performance of bagging and random forests.

Task 3

Description of data

Rosenberg and Kim (1975) studied the similarity of 15 kinship terms. Students sorted the kinship terms on the basis of their similarity into groups. Each student generated a dissimilarity matrix where a pair of kinship terms was coded as 1 if the terms were sorted in different groups and as 0 if the terms were sorted in the same group. The elements of the data set **dissim.Rdata** are dissimilarities that indicate the percentage of students that did not group a pair of terms together. In addition, the data set **features.Rdata** includes two external variables that describe the kinship terms, namely **generation** (-2 = two back, -1 = one back, 0 = same generation, 1 = one ahead, 2 = two ahead), and **degree** (1 = first, 2 = second, etc.).

Questions

a. Use smacofSym() to conduct MDS with 2 dimensions and assuming different measurement levels (i.e., ratio, interval, mspline, ordinal) for the observed dissimilarities.
b. Evaluate the goodness of fit of solutions with different measurement levels using stress-1, and by computing stress norms with the functions randomstress() and permutation(). Discuss which solution you would select. Investigate the stability of the selected solution using the Jackknife.
c. Use an MDSbiplot to project the standardized external variables generation and degree in the configuration plot of the selected solution, and interpret the results of the analysis.

References

ESS Round 6: European Social Survey Round 6 Data (2012). Data file edition 2.4. NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC. doi:10.21338/NSD-ESS6-2012

Rosenberg, S., & Kim, M. P. (1975). The method of sorting as a data gathering procedure in multivariate research. Multivariate Behavioral Research, 10, 489‑502.

van der Putten, P. & van Someren, M. (2000). CoIL Challenge 2000: The Insurance Company Case. Sentient Machine Research: Amsterdam. Also a Leiden Institute of Advanced Computer Science Technical Report 2000-09.

Submission of the assignment

For this assignment, one member of each team should upload the following files on Toledo before January 10 at 9 pm:

- Report with answers to questions of the tasks (word document or .pdf file). The length of the report is limited to **maximum 16 pages (including one title page)**.
- Script File with the R-code (.R file)

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. Remark: 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