# STATISTICA APPLICATA E ANALISI DEI DATI

## *Argomenti tesine a.a.2021-2022*

### Topics related to computer science

**Reticulate (assegnata)**

Illustrate the R package "Reticulate", providing an R interface to Python. The package has a website (https://rstudio.github.io/reticulate/),with plenty of documentation. Further documentation is available at the CRAN repository (https://cloud.r-project.org/web/packages/reticulate/index.html).

**rpy2 (assegnata)**

Illustrate the usage of the "rpy2" package for running R from python. The starting point is the webpage of the package (https://rpy2.github.io/index.html) and, in particular, an introduction is available on https://rpy2.github.io/doc/latest/html/introduction.html#. Further documents can be found on the internet.

**DBI (assegnata)**

Describe the usage of the R package "DBI", which provides a database interface definition for communication between R and relational database management systems. The package is illustrated in several vignettes available at https://cran.r-project.org/web/packages/DBI/index.html. A Github website provides further details and references (https://github.com/r-dbi/DBI).

**rrecsys (assegnata)**

Illustrate the usage of the "rrecsys" package, which defines an environment for evaluating recommender systems. Recommendation systems are intensively used for online business and marketing. A good starting point is the Github webpage https://github.com/ludovikcoba/rrecsys. The package is also available on the CRAN, with several vignettes available (https://cran.r-project.org/web/packages/rrecsys/index.html). There are many illustrations of the package available on the internet, such as http://ceur-ws.org/Vol-1688/paper-12.pdf.

**recommenderlab**

Illustrate the usage of the "recommenderlab" package, which provides research infrastructure to test and develop some recommender algorithms. A good starting point is the Github webpage https://github.com/mhahsler/recommenderlab containing also a link to the CRAN package vignette https://cran.r-project.org/web/packages/recommenderlab/vignettes/recommenderlab.pdf. There are several illustrations of the package available on the internet.

**Advanced R**

The book “Advanced R” provides a very useful description of the language, and it is freely available at https://adv-r.hadley.nz. Using the text as reference, illustrates some advanced features of R. Among the possible choices, some portions of Part II (on functional programming) or Part III (on object-oriented programming) would fit in well with the task.

## Applications

### dtwSat

Illustrate the usage of the “dtwSat” package which provides an implementation of the time-weighted dynamic time warping method for land cover mapping using sequence of multi-band satellite images. The package supports the full cycle of land cover classification using image time series. A description of the package is available at https://www.jstatsoft.org/article/view/v088i05.

### bupaR (assegnata)

The “bupaR” suite of packages (https://www.bupar.net/index.html) provides several tools for the handling and analysis of business process data. Illustrate the usage of the main package “bupaR”, along with some examples of usage of the other packages of the suite.

### stplanr

Illustrate the usage of the “stplanr” package which provides tools for transport planning with an emphasis on spatial transport data and non-motorized modes. The CRAN webpage https://cran.r-project.org/web/packages/stplanr/index.html contains plenty of documentation, consisting in several vignettes and some links.

### hoopR (assegnata)

Illustrate the usage of the “hoopR” package which provides functions to quickly obtain clean and tidy men's basketball play by play data. Some vignettes are available at the Github web page of the package (https://saiemgilani.github.io/hoopR/articles/index.html) and further documentation is available at the CRAN repository https://cran.r-project.org/web/packages/hoopR/index.html.

### PerformanceAnalytics

Illustrate the usage of the “PerformanceAnalytics” package which provides a wide collection of econometric functions for performance and risk analysis in a number of financial and economic frameworks. Describe and apply some of the functions of the package. Some vignettes are available at the CRAN web page (https://cran.r-project.org/web/packages/PerformanceAnalytics/index.html) and further documentation is available at the Github web page https://github.com/braverock/PerformanceAnalytics.

### quantmod (assegnata)

Describe the “quantmod” package which is designed to assist the quantitative trader in the development, testing and deployment of statistically based trading models. The documentation is available on the CRAN

repository and on the package web page http://www.quantmod.com/. Further documentation is provided in the Github web page https://github.com/joshuaulrich/quantmod.

**ggenealogy**

Describe the "ggenealogy" package that provides tools for searching through genealogical data and generating basic statistics with the aim of drawing a genealogy output. A useful source is the paper https://www.jstatsoft.org/article/view/v089i13. Further documentation can be found in the Github page https://github.com/lindsayrutter/ggenealogy and in the CRAN repository https://cran.r-project.org/web/packages/ggenealogy/index.html.

**phonics**

Describe the "phonics" package that provides several functions and algorithms for indexing words and for word matching across a variety of English language use cases. A useful source is the paper https://www.jstatsoft.org/article/view/v095i08. Further documentation can be found in the Github page https://github.com/k3jph/phonics-in-r and in the CRAN repository https://cran.r-project.org/web/packages/phonics/index.html.

**fastnet (assegnata)**

Illustrate the usage of the "fastnet" package, which provides tools for scaling and speeding up the simulation and analysis of large-scale social networks. A useful source is the paper https://www.jstatsoft.org/article/view/v096i07. Further documentation can be found in the CRAN repository https://cran.r-project.org/web/packages/fastnet/index.html.

**multinet (assegnata)**

Describe the "multinet" package that provides several functions for the creation/generation and analysis of multilayer social networks where a common set of actors are connected through multiple types of relations. A useful source is the paper https://www.jstatsoft.org/article/view/v098i08. Further documentation can be found in CRAN repository https://cran.r-project.org/web/packages/multinet/index.html.

**darts (assegnata)**

Illustrate the usage of the R package "darts", which gives tools for finding an optimal personalized strategy for playing darts. A good starting point is the webpage http://www.stat.cmu.edu/~ryantibs/darts/ and the research paper available at http://www.stat.cmu.edu/~ryantibs/papers/darts.pdf.

**trackeR (assegnata)**

Illustrate the usage of the R package "trackeR", which provide an infrastructure for analizing running and cycling data from GPS-enabled tracking devices. Vignettes are available at https://cran.r-project.org/web/packages/trackeR/vignettes/trackeR.pdf and https://cran.r-project.org/web/packages/trackeR/vignettes/TourDetrackeR.html. After extraction and appropriate manipulation of the training or competition attributes, the data are placed into session-based and unit-aware data objects of class 'trackeRdata'. The information in the resulting data objects can then be visualised, summarised, and analysed through corresponding flexible and extensible methods.

**radiant**

Illustrate the usage of the R package "radiant", which is a browser-based interface for business analytics in R. A vignette is available at https://cran.r-project.org/web/packages/radiant/vignettes/programming.html. The webpage of the software is https://github.com/radiant-rstats/radiant, containing a link to a page with documentation and tutorials (https://radiant-rstats.github.io/docs/). A simple starting tutorial can be downloaded at https://radiant-rstats.github.io/docs/tutorials.html.

## Statistical learning

**arules**

Illustrate the R package "arules" package, which provides the infrastructure for representing, manipulating and analyzing transaction data and patterns with the aim of finding frequent item sets and association rules. Association rules are commonly used in marketing. A vignette is available at https://cran.r-project.org/web/packages/arules/vignettes/arules.pdf and a tutorial can be found at http://michael.hahsler.net/research/arules_RUG_2015/demo/.

**vcd**

Illustrate the usage of the R package "vcd" for the visualization of categorical data. Documentation is available at the CRAN repository, https://cran.r-project.org/web/packages/vcd/index.html, where two vignettes can be found.

**stream**

Illustrate the usage of the R package "stream" for data stream modelling, focusing on the associated data mining tasks, such as clustering and classification. A vignette is available at https://cran.r-project.org/web/packages/stream/vignettes/stream.pdf.

**DataVisualizations**

Illustrate the usage of the R package "DataVisualizations", which is a collection of various visualization methods, useful for exploratory data analysis. A vignette is available at https://cran.r-project.org/web/packages/DataVisualizations/vignettes/DataVisualizations.html.

**HDclassif**

Illustrate the usage of the R package "HDclassif", which is devoted to the clustering and the classification of high-dimensional data. Consider the paper available at https://www.jstatsoft.org/article/view/v046i06, focusing in particular on clustering procedures.

**CBDA**

Illustrate the usage of the R package “CBDA”, which is devoted to classification concerning Big Data. A tutorial is available at https://cran.r-project.org/web/packages/CBDA/vignettes/Guide-to-CBDA.html, including links to dataset and applications.

**apcluster**

Illustrate the usage of the R package “apcluster” for affinity propagation clustering. There is a package vignette at https://cran.r-project.org/web/packages/apcluster/vignettes/apcluster.pdf and further documentation can be found on the internet.

**caret**

Illustrate the usage of the “caret” package for predictive modeling. There is a short introduction to it at the package vignette (https://cran.r-project.org/web/packages/caret/vignettes/caret.html), and the package has its own webpage http://topepo.github.io/caret/index.html, that includes several documents. Among the latter, the article http://www.jstatsoft.org/v28/i05/paper presents a survey on the main functionalities of the software.

**rminer (classification)**

Illustrate the usage of the R package “rminer”, focusing con classification models. The “rminer” package provides a set of R functions to perform classification and regression. Additional documentation is available at http://www3.dsi.uminho.pt/pcortez/rminer.html and a useful tutorial is http://repositorium.sdum.uminho.pt/bitstream/1822/36210/1/rminer-tutorial.pdf.

**rminer (regression)**

Illustrate the usage of the R package “rminer”, focusing con regression models. The “rminer” package provides a set of R functions to perform classification and regression. Additional documentation is available at http://www3.dsi.uminho.pt/pcortez/rminer.html and a useful tutorial is http://repositorium.sdum.uminho.pt/bitstream/1822/36210/1/rminer-tutorial.pdf.

**ranger**

Illustrate the usage of the “ranger” package for fast and scalable estimation of random forests, available at https://github.com/imbs-hl/ranger (the package is also available on the CRAN repository). The software is illustrated in https://www.jstatsoft.org/article/view/v077i01. Make also a comparison with other packages for random forests, such as “randomForest”.

**dbscan (assegnata)**

Illustrate the usage of the R package “dbscan” for density-based clustering. Two vignettes documenting the software are available at the CRAN webpage https://cran.r-project.org/web/packages/dbscan/ and the package has also a Github page https://github.com/mhahsler/dbscan. A useful source is the paper https://www.jstatsoft.org/article/view/v091i01.

**xgboost**

Extreme Gradient Boosting is an ensemble method representing a quite useful alternative to basic decision trees methods. Illustrate the usage of the "xgboost" package, which provides a rather efficient implementation of the method. Some illustrative vignettes are available at https://cran.r-project.org/web/packages/xgboost/, with more material available on the internet.

**VIM (assegnata)**

Illustrate the usage of the "VIM" package for visualisation and imputation of missing values. The package is illustrated in several vignettes available at https://cran.r-project.org/web/packages/VIM/index.html. A Github website provides further details and references (https://github.com/statistikat/VIM).

**ClusteR**

Illustrate the usage of the "ClusteR" package which provides a set of methods for cluster analysis, such as Gaussian Mixture Models, K-Means, K-Medoids and Affinity Propagation Clustering. A vignette is available at https://cran.r-project.org/web/packages/ClusterR/vignettes/the_clusterR_package.html and further details can be found on a Github web page https://github.com/mlampros/ClusterR.

**mclust**

Illustrate the usage of the "mclust" package which implements Gaussian mixture modelling for model-based clustering, classification and density estimation. A vignette is available at https://cran.r-project.org/web/packages/mclust/vignettes/mclust.html and further details can be found on a Github web page https://mclust-org.github.io/mclust/.

**Statistical models**

**glmnet**

Illustrate the main features of the R package "glmnet", with particular regard to ridge regression and the Lasso. A vignette available at https://cran.r-project.org/web/packages/glmnet/vignettes/glmnet.pdf and a useful source is the paper https://www.jstatsoft.org/article/view/v033i01. For further documentation see sections 6.2 and 6.6 of "An Introduction to Statistical Learning", available at http://www-bcf.usc.edu/~gareth.

**effects**

Illustrate the usage of the R package "effects" for effect displays in linear models, generalized linear models and other models. The package is documented in the article http://www.jstatsoft.org/article/view/v008i15/effect-displays-revised.pdf, and the package has also some vignettes (https://cran.r-project.org/web/packages/effects/). Plenty of documentation is available on the internet, such as the presentation at https://socialsciences.mcmaster.ca/jfox/Courses/sem-goettingen/partial-residuals-effect-displays-notes-revised.pdf, and many other documents.

**ordinal**

Illustrate the use of the R package “ordinal” for regression models with ordinal data. A vignette is available at https://cran.r-project.org/web/packages/ordinal/vignettes/clm_article.pdf and a tutorial can be downloaded at https://cran.r-project.org/web/packages/ordinal/vignettes/clmm2_tutorial.pdf.

**sn**

Illustrate the R package “sn”, which concerns probability distributions of the skew-normal and of the skew-t families. Additional documentation is available at http://azzalini.stat.unipd.it/SN/. There is also a book on the skew-normal and the skew-t distributions, entitled “The Skew-Normal and Related Families” and available at the University of Udine library (Polo Economico-Giuridico, Via Tomadini 30).

**glmulti**

Illustrate the R package “glmulti”, which perform automatic model selection by considering a very large number of candidate glm models. A description of the package is available at https://www.jstatsoft.org/article/view/v034i12/v34i12.pdf.

**nlstools**

Illustrate the main features of the R package “nlstools”, which gives several tools for assessing the quality of fit of a gaussian nonlinear regression model. Besides the references mentioned in the package documentation, a useful source is the paper https://www.jstatsoft.org/article/view/v066i05.

**pls**

Illustrate the R package “pls”, which performs partial least squares and principal component regression. For useful documentation see chapter 6 of “Applied Predictive Modeling”, available at the University of Udine library (Polo Economico-Giuridico, Via Tomadini 30) and sections 6.3 and 6.7 of “An Introduction to Statistical Learning”, available at available at http://www-bcf.usc.edu/~gareth.

**betareg**

Illustrate the usage of the R package “betareg” for regression models with beta-distributed responses. Some documentation is available https://cran.revolutionanalytics.com/web/packages/betareg/vignettes/betareg.pdf and an interesting application is proposed in the paper http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2065320.

**regtools**

Illustrate the usage of the R package “regtools” for regression and classification. The package, available at https://cran.r-project.org/web/packages/regtools/index.html has a broad scope, therefore focus on some of the methods. Further documentation is available at https://github.com/matloff/regtools.

**cvms (assegnata)**

Describe the usage of the R package “cvms”, which provides cross-validation procedures for assessing the predictive ability of regression and classification models. The package is illustrated in several vignettes available at https://cran.r-project.org/web/packages/cvms/index.html. A Github website provides further details and references (https://github.com/LudvigOlsen/cvms).

**quantreg**

Describe the usage of the R package “quantreg”, which provides functions for inference and prediction with regard to quantile regression models. Consider, in particular, the content of the vignette available at https://cran.r-project.org/web/packages/quantreg/vignettes/crq.pdf. Further documentation is available at the Github page https://github.com/cran/quantreg.

**segmented**

Segmented regression is a flexible method for effective modelling in many fields, as described at https://en.wikipedia.org/wiki/Segmented_regression. Illustrate the “segmented” package for applying this technique. Further information is available on the CRAN repository (https://cran.r-project.org/web/packages/segmented/index.html) and on the internet as, for example, in https://rpubs.com/MarkusLoew/12164.

**strucchange**

Describe the usage of the R package “strucchange”, which provides functions for testing, monitoring and dating structural changes in linear regression models. The package is illustrated in the vignette available at https://cran.r-project.org/web/packages/strucchange/vignettes/strucchange-intro.pdf. A Github website provides further details and references (https://github.com/cran/strucchangeRcpp).

**The students have to send an email to** paolo.vidoni@uniud.it **in order to define the topic for the written report, if not already assigned. The assignment is on a first come, first served basis.**

Udine, 13.12.2021