SET UP FILE COURSEWORK PROJECT

1. Download the zip file.
2. Unzip the file and load the files into the existing directory.
3. Install the libraries that are specified in the below table.

|  |  |  |  |
| --- | --- | --- | --- |
| **LIBRARY NAME** | **VERSION** | **DESCRIPTION** | **DOCUMENTATION** |
| pacman | 0.5.1 | Tools to more conveniently perform tasks associated with add-on packages. pacman conveniently wraps library and package related functions and names them in an intuitive and consistent fashion. It seeks to combine functionality from lower level functions which can speed up workflow. | https://cran.r-project.org/web/packages/pacman/pacman.pdf |
| ggplot2 | 3.3.3 | A system for 'declaratively' creating graphics, based on ``The Grammar of Graphics''. You provide the data, tell 'ggplot2' how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details. | https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf |
| caret | 6.0-88 | Misc functions for training and plotting classification and regression models. | https://cran.r-project.org/web/packages/caret/caret.pdf |
| scatterplot3d | 0.3-41 | Plots a three dimensional (3D) point cloud. | https://cran.r-project.org/web/packages/scatterplot3d/scatterplot3d.pdf |
| ROSE | 0.0-3 | The package provides functions to deal with binary classification problems in the presence of imbalanced classes. Synthetic balanced samples are generated according to ROSE (Menardi and Torelli, 2013). Functions that implement more traditional remedies to the class imbalance are also provided, as well as different metrics to evaluate a learner accuracy. These are estimated by holdout, bootstrap or cross-validation methods. | https://cran.r-project.org/web/packages/ROSE/ROSE.pdf |
| randomForest | 4.6-14 | Classification and regression based on a forest of trees using random inputs, based on Breiman (2001) . | https://cran.r-project.org/web/packages/randomForest/randomForest.pdf |
| e1071 | 1.7-6 | Functions for latent class analysis, short time Fourier transform, fuzzy clustering, support vector machines, shortest path computation, bagged clustering, naive Bayes classifier, generalized k-nearest neighbour ... | https://cran.r-project.org/web/packages/e1071/e1071.pdf |
| Mlmetrics | 1.1.1 | A collection of evaluation metrics, including loss, score and utility functions, that measure regression, classification and ranking performance. | https://cran.r-project.org/web/packages/MLmetrics/MLmetrics.pdf |
| pROC | 1.17.0.1 | Tools for visualizing, smoothing and comparing receiver operating characteristic (ROC curves). (Partial) area under the curve (AUC) can be compared with statistical tests based on U-statistics or bootstrap. Confidence intervals can be computed for (p)AUC or ROC curves. | https://cran.r-project.org/web/packages/pROC/pROC.pdf |
| ROCR | 1.0-11 | ROC graphs, sensitivity/specificity curves, lift charts, and precision/recall plots are popular examples of trade-off visualizations for specific pairs of performance measures. ROCR is a flexible tool for creating cutoff-parameterized 2D performance curves by freely combining two from over 25 performance measures (new performance measures can be added using a standard interface). Curves from different cross-validation or bootstrapping runs can be averaged by different methods, and standard deviations, standard errors or box plots can be used to visualize the variability across the runs. The parameterization can be visualized by printing cutoff values at the corresponding curve positions, or by coloring the curve according to cutoff. All components of a performance plot can be quickly adjusted using a flexible parameter dispatching mechanism. Despite its flexibility, ROCR is easy to use, with only three commands and reasonable default values for all optional parameters. | https://cran.r-project.org/web/packages/ROCR/ROCR.pdf |
| dplyr | 1.0.6 | A fast, consistent tool for working with data frame like objects, both in memory and out of memory | https://cran.r-project.org/web/packages/dplyr/dplyr.pdf |
| kernlab | 0.9-29 | Kernel-based machine learning methods for classification, regression, clustering, novelty detection, quantile regression and dimensionality reduction. Among other methods 'kernlab' includes Support Vector Machines, Spectral Clustering, Kernel PCA, Gaussian Processes and a QP solver. | https://cran.r-project.org/web/packages/kernlab/kernlab.pdf |
| naivebayes | 0.9.7 | In this implementation of the Naive Bayes classifier following class conditional distributions are available: Bernoulli, Categorical, Gaussian, Poisson and non-parametric representation of the class conditional density estimated via Kernel Density Estimation. Implemented classifiers handle missing data and can take advantage of sparse data. | https://cran.r-project.org/web/packages/naivebayes/naivebayes.pdf |
| PRROC | 1.3.1 | Computes the areas under the precision-recall (PR) and ROC curve for weighted (e.g., softlabeled) and unweighted data. In contrast to other implementations, the interpolation between points of the PR curve is done by a non-linear piecewise function. In addition to the areas under the curves, the curves themselves can also be computed and plotted by a specific S3- method. References: Davis and Goadrich (2006) ; Keilwagen et al. (2014) ; Grau et al. (2015) | https://cran.r-project.org/web/packages/PRROC/PRROC.pdf |
| data.table | 1.14.0 | Fast aggregation of large data (e.g. 100GB in RAM), fast ordered joins, fast add/modify/delete of columns by group using no copies at all, list columns, friendly and fast character-separated-value read/write. Offers a natural and flexible syntax, for faster development. | https://cran.r-project.org/web/packages/data.table/data.table.pdf |
| plotly | 4.9.3 | Create interactive web graphics from 'ggplot2' graphs and/or a custom interface to the (MITlicensed) JavaScript library 'plotly.js' inspired by the grammar of graphics. | https://cran.r-project.org/web/packages/plotly/plotly.pdf |
| rattle | 5.4.0 | The R Analytic Tool To Learn Easily (Rattle) provides a collection of utilities functions for the data scientist. A Gnome (RGtk2) based graphical interface is included with the aim to provide a simple and intuitive introduction to R for data science, allowing a user to quickly load data from a CSV file (or via ODBC), transform and explore the data, build and evaluate models, and export models as PMML (predictive modelling markup language) or as scores. A key aspect of the GUI is that all R commands are logged and commented through the log tab. This can be saved as a standalone R script file and as an aid for the user to learn R or to copy-and-paste directly into R itself. | https://cran.r-project.org/web/packages/rattle/rattle.pdf |
| plyr | 1.8.6 | A set of tools that solves a common set of problems: you need to break a big problem down into manageable pieces, operate on each piece and then put all the pieces back together. For example, you might want to fit a model to each spatial location or time point in your study, summarise data by panels or collapse high-dimensional arrays to simpler summary statistics. The development of 'plyr' has been generously supported by 'Becton Dickinson'. | https://cran.r-project.org/web/packages/plyr/plyr.pdf |
| scales | 1.1.1 | Graphical scales map data to aesthetics, and provide methods for automatically determining breaks and labels for axes and legends. | https://cran.r-project.org/web/packages/scales/scales.pdf |
| MASS | 7.3-54 | Functions and datasets to support Venables and Ripley, ``Modern Applied Statistics with S'' (4th edition, 2002). | https://cran.r-project.org/web/packages/MASS/MASS.pdf |
| lattice | 0.20-44 | A powerful and elegant high-level data visualization system inspired by Trellis graphics, with an emphasis on multivariate data. Lattice is sufficient for typical graphics needs, and is also flexible enough to handle most nonstandard requirements. See ?Lattice for an introduction. | https://cran.r-project.org/web/packages/lattice/lattice.pdf |
| tidyr | 1.1.3 | Tools to help to create tidy data, where each column is a variable, each row is an observation, and each cell contains a single value. 'tidyr' contains tools for changing the shape (pivoting) and hierarchy (nesting and 'unnesting') of a dataset, turning deeply nested lists into rectangular data frames ('rectangling'), and extracting values out of string columns. It also includes tools for working with missing values (both implicit and explicit). | https://cran.r-project.org/web/packages/tidyr/tidyr.pdf |
| repr | 1.1.3 | String and binary representations of objects for several formats / mime types. | https://cran.r-project.org/web/packages/repr/repr.pdf |
| rpart.plot | 3.0.9 | Plot 'rpart' models. Extends plot.rpart() and text.rpart() in the 'rpart' package. | https://cran.r-project.org/web/packages/rpart.plot/rpart.plot.pdf |

1. R environment needs to load these libraries to run the coursework.R file.

Select Tools > Install Packages…

In the empty “Packages” box type : pacman

1. Ensure that the “Install dependencies” box is ticked. Select Install

and with any luck and the University of Surrey security settings allow –

the console will show the progress of the library being downloaded and

then unzipped and installed.