kernlab: Machine Learning with kernels in R

Alexandros Karatzoglou[1]
joint work with Alex Smola and Kurt Hornik

[1]INSA de Rouen
LITIS
France

December 12, 2008

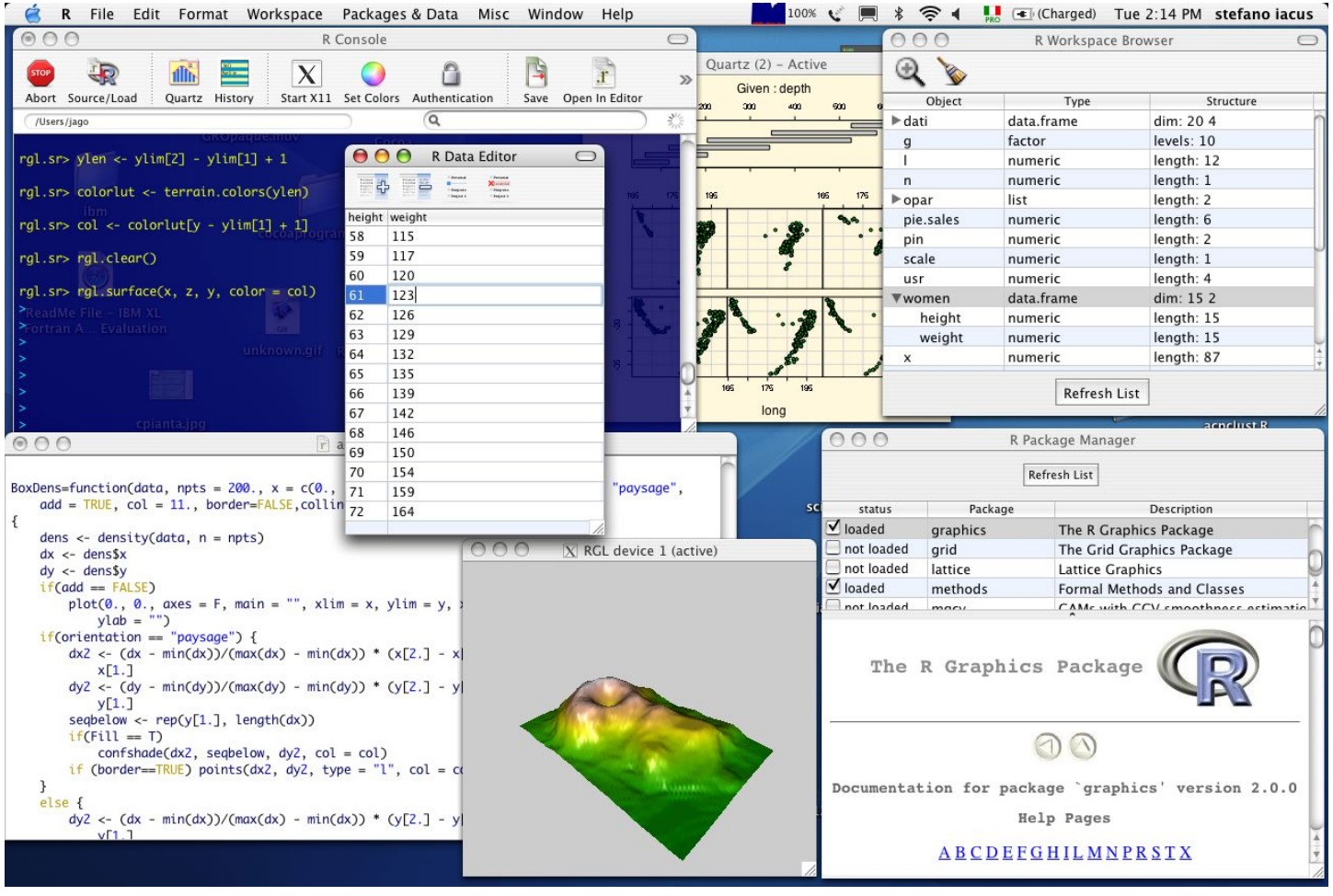
Outline

kernlab

kernlab R package:

- Contains a wide range of kernel-based Machine Learning methods
- Uses modern Open Source R environment
- Extensible by exploiting the inherent modularity of kernel methods
- GPL 2

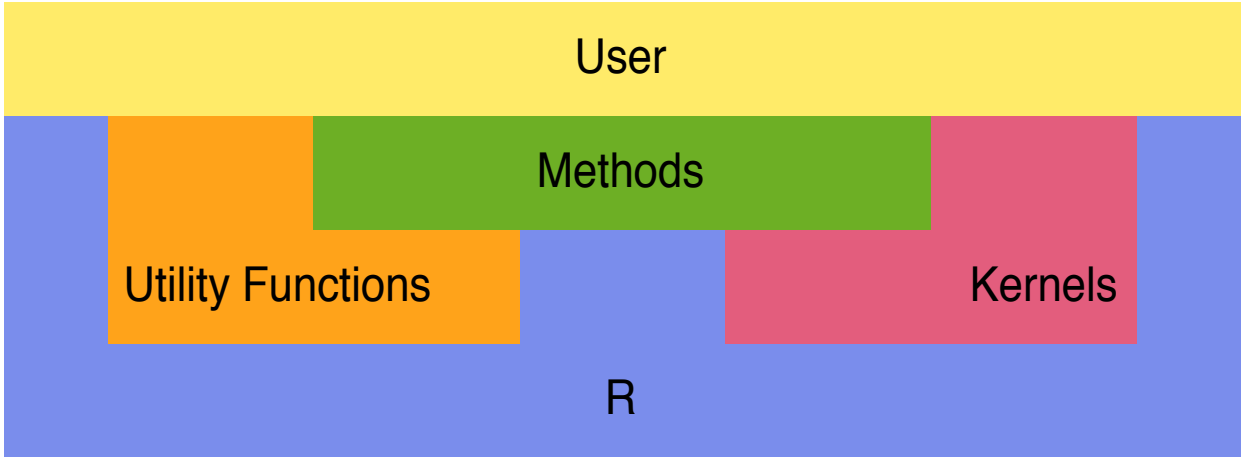
R


R File Edit Format Workspace Packages & Data Misc Window Help
R Console
Abort Source/Load Quartz History Start X11 Set Colors Authentication Save Open In Editor
rgl.sr> ylen <- ylim[2] - ylim[1] + 1
rgl.sr> colorlut <- terrain.colors(ylen)
rgl.sr> col <- colorlut[y - ylim[1] + 1]
rgl.sr> rgl.clear()
rgl.sr> rgl.surface(x, z, y, color = col)
R Data Editor
height weight
58 115
59 117
60 120
61 123
62 126
63 129
64 132
65 135
66 139
67 142
68 146
69 150
70 154
71 159
72 164
Quartz (2) - Active
Given : depth
long
R Workspace Browser
Object Type Structure
dati data.frame dim: 20 4
g factor levels: 10
l numeric length: 12
n numeric length: 1
opar list length: 2
pie.sales numeric length: 6
pin numeric length: 2
scale numeric length: 1
usr numeric length: 4
women data.frame dim: 15 2
height numeric length: 15
weight numeric length: 15
x numeric length: 87
Refresh List
BoxDens=function(data, npts = 200., x = c(0.,
add = TRUE, col = 11., border=FALSE,collin
{
dens <- density(data, n = npts)
dx <- dens$x
dy <- dens$y
if(add == FALSE)
plot(0., 0., axes = F, main = "", xlim = x, ylim = y,
ylab = "")
if(orientation == "paysage") {
dx2 <- (dx - min(dx))/(max(dx) - min(dx)) * (x[2.] - x
x[1.]
dy2 <- (dy - min(dy))/(max(dy) - min(dy)) * (y[2.] - y
y[1.]
seqbelow <- rep(y[1.], length(dx))
if(Fill == T)
confshade(dx2, seqbelow, dy2, col = col)
if (border==TRUE) points(dx2, dy2, type = "l", col = c
}
else {
dy2 <- (dx - min(dx))/(max(dx) - min(dx)) * (y[2.] - y
"paysage",
RGL device 1 (active)
R Package Manager
Refresh List
status Package Description
loaded graphics The R Graphics Package
not loaded grid The Grid Graphics Package
not loaded lattice Lattice Graphics
loaded methods Formal Methods and Classes
The R Graphics Package
Documentation for package `graphics' version 2.0.0
Help Pages
A B C D E F G H I L M N P R S T X

R

- Environment for statistical data analysis, inference and visualization.
- Ports for Unix, Windows and MacOSX
- Highly extensible through user-defined functions
- Generic functions and conventions for standard operations like plot, predict etc.
- $\sim$ 1200 add-on packages contributed by developers from all over the world
- e.g. Multivariate Statistics, Machine Learning, Natural Language Processing, Bioinformatics (Bioconductor).
- Interfaces to C, C++, Fortran, Java

kernlab Package Content



- 9 kernel functions, 4 kernel expression functions
- 18 kernel-based ML methods:
  - Regression, Quantile regression, Classification and Novelty detection (SVM, RVM, Gaussian Processes)
  - Clustering (kernel *k*-means, Spectral Clustering)
  - Ranking and dimensionality reduction (kernel PCA, kernel CCA, etc.) kernel-based two sample test (KMMD)
- Utility Functions:
  - Quadratic Problem solver
  - Incomplete Cholesky Decomposition

Some kernel functions in `kernlab`

- Linear kernel
- Gaussian radial basis kernel
- Hyperbolic tangent kernel
- String kernel
- Polynomial kernel
- ...

Runtime

Inputs

- `kernlab` methods take kernel function and data or kernel matrix as input
- Function parameters are given as options

Outputs

- Return objects that contain model parameters.
- Objects can be used with generic functions such as predict, plot etc.

Performance

- Nystrom Method for eigendecomposition based methods (e.g. spectral clustering). Eigenvectors can be calculated using only part of the kernel matrix yielding speedups and better scaling $O(n^3) + O(nN)$.
- Build in kernel function computations are vectorized
- C++ on performance bottlenecks
- Use of scalable algorithms
- ...

Features

Most function implement additional features such as e.g.

- Cross-validation (svm, GP's)
- Model-selection in particular hyper-parameter estimation
- Data Scaling
- Extensive documentation of every function and options

Demo