An Introduction to R: Statistics, Graphics and Data Mining

Daniele Soria
d.soria@cs.nott.ac.uk
IMA tutorial, 2 March 2010

Outline
- Why R?
- Download and install
- Data manipulation
- Basic statistics
- Graphics
- Clustering

Why R?
- It allows users to add additional functionality by defining new functions.
- Its graphical facilities produce publication-quality graphs which can include mathematical symbols.
- And most importantly...
  IT’S FREE!!

Download and install
- [http://www.r-project.org](http://www.r-project.org)
- On the left select CRAN
- Scroll down to find UK and select the University of Bristol link
- Select the OS you are using
- For Windows select base
- Download R 2.10.1 for Windows and install it

Let’s get started
- In R-2.10.1 folder, several packages (libraries) are already present
- Command line interface
- Follow suggestion and type `demo()`
- Type `demo(graphics)`
- Script files may be written in R
  1. I use an external editor, called Crimson Editor
- Create
  - Variable
  - Vector
  - Sequence
  - Replicate a value
- Length, max, min, mean, summary
- Dealing with missing values
- Vectors and matrices

Basic stuff
- Create
  - Variable
  - Vector
  - Sequence
  - Replicate a value
- Length, max, min, mean, summary
- Dealing with missing values
- Vectors and matrices
Load data
- UCI Machine Learning Repository
  - Download IRIS data
- `read.table(options)`
- Dataset information
  - Names of variables
  - Size
  - Tables
  - `ifelse` command

Graphics
- Pie charts
- Histograms
  - Add mean and median vertical lines
  - Add legend
- Boxplots
- Matrix of scatter plots
- 3D plots

Data Mining
- Clustering
  - K-means
  - PAM
  - Fuzzy c-means
- Plots of clustering results
- Principal component analysis

KM method
- Minimization of the objective function
  \[ J(V) = \sum_{j=1}^{k} \sum_{i=1}^{c_j} \| x_i - v_j \|^2 \]
- \( \| x_i - v_j \| \): Euclidean distance between \( x_i \) and \( v_j \)
- \( c_j \): data points in cluster \( j \)
- \( v_j \) can be calculated as \( v_j = \frac{1}{c_j} \sum_{i=1}^{c_j} x_i \), \( j = 1, ..., k \)

KM method (cont.)

PAM method
- Based on the search for \( k \) representative objects (medoids) among the observations
- Minimum sum of dissimilarities among the observations to their closest medoid
- \( k \) clusters are constructed by assigning each observation to the nearest medoid
Additional resources

- R-Seek Search Engine
  - [http://www.rseek.org](http://www.rseek.org)


- R Reference Card
  - [http://cran.r-project.org/doc/contrib/Short-refcard.pdf](http://cran.r-project.org/doc/contrib/Short-refcard.pdf)

- Use Google to look for specific commands and functions

Thank you!

Questions?

Contact: d.soria@cs.nott.ac.uk