Incorporating R functionality

Ashish Mahabal

KDD IG session

IVOA Pune InterOp

20 Oct 2011

http://www.ivoa.net/cgi-bin/twiki/bin/view/IVOA/IvoaKDDguide

- **1: What is 'data mining' and why is it important in astronomy?**
- **5: Which algorithm to use?**
- **7: Algorithms and techniques astronomers could benefit from but seldom use**
- **9: Worked example**

Algorithms astronomers could benefit from

Subject areas include:

- Applied mathematics
- Artificial intelligence
- Computer science
- Signal processing
- Statistics

Algorithms astronomers could benefit from

- **Bootstrap aggregating (bagging)**
- **Mixture of experts**
- **Boosting**
- **Simulated annealing**
- **Semi-supervised learning**
- **Radial basis function networks**
- **Adaptive resonance theory**
- **Fuzzy c-means clustering**
- **Hidden Markov models**
- **Wavelets**

With help from R Ramnath (Kerala)

??bootstrap

- boot::nested.corr Functions for Bootstrap Practicals
- boot::boot Bootstrap Resampling
- boot::boot.array Bootstrap Resampling Arrays
- boot::boot.ci Nonparametric Bootstrap Confidence Intervals
- boot::cd4.nested Nested Bootstrap of cd4 data
- boot::censboot Bootstrap for Censored Data
- boot::freq.array Bootstrap Frequency Arrays
- boot::jack.after.boot Jackknife-after-Bootstrap Plots
- boot::linear.approx Linear Approximation of Bootstrap Replicates
- boot::plot.boot Plots of the Output of a Bootstrap Simulation
- boot::print.boot Print a Summary of a Bootstrap Object
- boot::print.bootci Print Bootstrap Confidence Intervals
- boot::saddle Saddlepoint Approximations for Bootstrap
- Statistics
- boot::saddle.distn Saddlepoint Distribution Approximations for
- Bootstrap Statistics
- boot::tilt.boot Non-parametric Tilted Bootstrap
- boot::tsboot Bootstrapping of Time Series

Bootstrap aggregating(bagging)/ Bootstrap Frequency Arrays.html

```
freq.array                package:boot                R Documentation

Bootstrap Frequency Arrays

Description:

     Take a matrix of indices for nonparametric bootstrap resamples and
     return the frequencies of the original observations in each
     resample.

Usage:

     freq.array(i.array)

Arguments:

 i.array: This will be an matrix of integers between 1 and n, where n
          is the number of observations in a data set.  The matrix will
          have n columns and R rows where R is the number of bootstrap
          resamples.  Such matrices are found by 'boot' when doing
          nonparametric bootstraps.  They can also be found after a
          bootstrap has been run through the function 'boot.array'.

Value:

     A matrix of the same dimensions as the input matrix.  Each row of
     the matrix corresponds to a single bootstrap resample.  Each
     column of the matrix corresponds to one of the original
     observations and specifies its frequency in each bootstrap
     resample.  Thus the first column tells us how often the first
     observation appeared in each bootstrap resample.  Such frequency
     arrays are often useful for diagnostic purposes such as the
     jackknife-after-bootstrap plot.  They are also necessary for the
     regression estimates of empirical influence values and for finding
     importance sampling weights.

See Also:

     'boot.array'
```

Fuzzyc-meansclustering/ FuzzyAnalysisClustering.html

```
fanny                  package:cluster                  R Documentation

Fuzzy Analysis Clustering

Description:

     Computes a fuzzy clustering of the data into 'k' clusters.

Usage:

     fanny(x, k, diss = inherits(x, "dist"), memb.exp = 2,
           metric = c("euclidean", "manhattan", "SqEuclidean"),
           stand = FALSE, iniMem.p = NULL, cluster.only = FALSE,
           keep.diss = !diss && !cluster.only && n < 100,
           keep.data = !diss && !cluster.only,
           maxit = 500, tol = 1e-15, trace.lev = 0)

Arguments:

       x: data matrix or data frame, or dissimilarity matrix, depending
          on the value of the 'diss' argument.

          In case of a matrix or data frame, each row corresponds to an
          observation, and each column corresponds to a variable. All
          variables must be numeric.  Missing values (NAs) are allowed.

          In case of a dissimilarity matrix, 'x' is typically the
          output of 'daisy' or 'dist'.  Also a vector of length
          n*(n-1)/2 is allowed (where n is the number of observations),
          and will be interpreted in the same way as the output of the
          above-mentioned functions.  Missing values (NAs) are not
          allowed.
```

?hist

Examples:

```
op <- par(mfrow=c(2, 2))
hist(islands)
utils::str(hist(islands, col="gray", labels = TRUE))

hist(sqrt(islands), breaks = 12, col="lightblue", border="pink")
##-- For non-equidistant breaks, counts should NOT be graphed unscaled:
r <- hist(sqrt(islands), breaks = c(4*0:5, 10*3:5, 70, 100, 140),
          col='blue1')
text(r$mids, r$density, r$counts, adj=c(.5, -.5), col='blue3')
sapply(r[2:3], sum)
sum(r$density * diff(r$breaks)) # == 1
lines(r, lty = 3, border = "purple") # -> lines.histogram(*)
par(op)

require(utils) # for str
str(hist(islands, breaks=12, plot= FALSE)) #-> 10 (~= 12) breaks
str(hist(islands, breaks=c(12,20,36,80,200,1000,17000), plot = FALSE))

hist(islands, breaks=c(12,20,36,80,200,1000,17000), freq = TRUE,
```

"Digging deeper and faster: algorithms for computationally limited problems in time-domain astronomy"Caltech, Pasadena, CA, Dec. 12-13, 2011

- faster computers or better algorithms (massive data sets)
- include gravitational-wave data analysis
- The rapidly developing field of synoptic sky surveys.
- This workshop is a part of the Keck Institute for Space Studies program.
- Review some of the work done to date
- Focus on some of the outstanding issues for the future studies
- The workshop is open to anyone interested, and there is no registration fee. However, a registration is mandatory, for the logistical purposes, as the venue size is limited. http://www.astro.caltech.edu/digging/The (evolving) program and further details will be posted on that website.