STAT 6560
Graphical Methods

Spring Semester 2009

Project 2

Ying Jin

Utah State University

Department of Mathematics and Statistics

3900 Old Main Hill

Logan, UT 84322–3900

Introduction to animation package

- This package is invented by Yihui Xie in 2007.
- "This package consists of various functions for animations in statistics, covering many areas such as probability theory, mathematical statistics, multivariate statistics, nonparametric statistics, sampling survey, linear models, time series, computational statistics, data mining and machine learning. These functions might be of help in teaching statistics and data analysis." (From R documentation for this package: `http://cran.r-project.org/web/packages/animation/animation.pdf`)
- This package contains many functions performing the demonstrations of statistical issues. Such as central limit theorem, law of large numbers, the estimation of *pi* by the demonstration of Buffon'needls. Please refer to the next section for examples.

3 examples from the animation packagae

- "Animation". This is a demonstration of how animation works within R graphical device: `http://animation.yihui.name/animation:start`.
- "Buffon's needle". This animation gives the estimate of *pi*. There are three graphs created by this demonstration: the top-left one is a simulation of the scenario, the topright one is to help us understand the connection between dropping needles and the mathematical method to estimate *pi*, and the bottom one is the result for each dropping.The mathematical background can be found at `http://en.wikipedia.org/wiki/Buffons_needle`.
- "least.squares". This is a simple demonstration of the meaning of least squares in univariate linear regression. With either the intercept or the slope changing, the lines will be moving in the graph and corresponding residuals will be plotted. We can finally see the best estimate of the intercept and the slope from the residual plot.

Please refer to the R code at the end for the design choices for these graphics.

Teaching demonstration

- Estimating the natural base e. In Math 1050 class on campus, I have to introduce a very important mathematical concept: natural base e. I can use the following animation to demonstrate why e is not a specific value, it is approaching to some bound. This is a good introductory example of teaching the concepts of asymptotes.

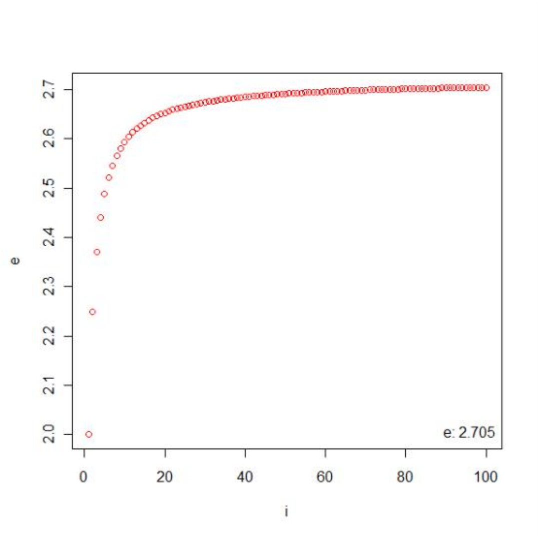

Figure 1: Estimating e

- The change of error rate in classification using the method of randomForest. One of the parameters in randomForest which will change the error rate is the number of trees we grow. The default number of trees if 500. The animation here is to demonstrate how the error rate changes when the number of trees we grow changes.

- "Swans data". There are 8 variables and 135 observations. The continuous variables are circum, weight, wingspan, cwidth and cheight. The three categorical variables are species, sex and age. I try to use this data set to demonstrate animations in randomForest and Kmeans cluster analysis.

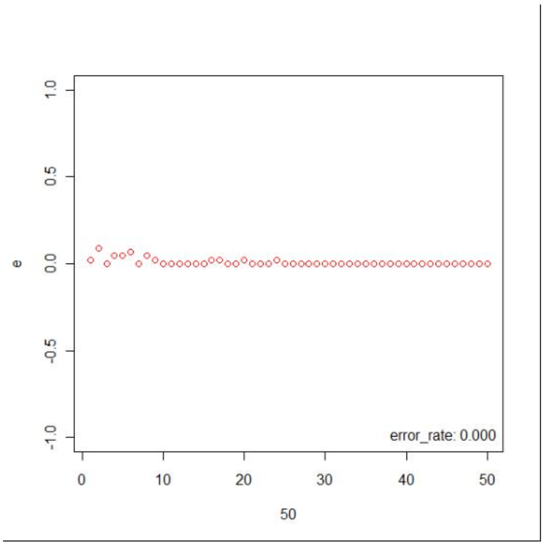

Figure 2: randomForest

- "kmeans.ani". This function is used to finding cluster centers, computing distances between sample points, and redefining cluster membership.The kmeans algorithm aims to partition the data into k groups such that the sum of squares from data to the assigned cluster centers is minimized. At the minimum, all cluster centers are at the mean of their Voronoi sets (the set of data points which are nearest to the cluster center). For the swans data, I selected the most two interesting predictor variable "circum" and "wingspan" to do the cluster analysis. The classifier is species. Since we only have two levels for the class variable, so the center parameter in the function should be 2.

- "Interpretation". Please refer to the R output for details.

Links to data files and R-code

- `http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/jin_ying_project2_swans.txt`

- `http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/jin_ying_project2_Rcode.R`

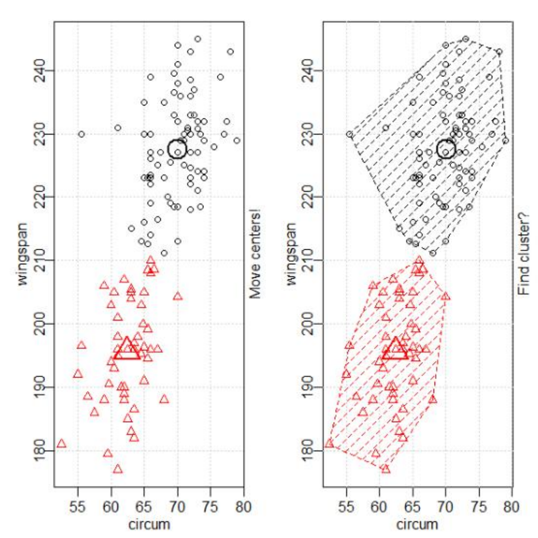

Figure 3: Kmeans