

# Acc 522 Statistical Methods for Forensic Accounting & Assurance

Instructor: Jagdish S. Gangolly  
Department of Accounting & Law  
State University of New York at Albany

## 1 Catalog Description

Exploratory descriptive data analysis using the language R. Basic graphics commands in R. Descriptive data exploration and statistical modeling. Data preprocessing for Datamining. Classification: Induction of Decision trees, Association Rules in Large Databases. Multivariate Methods: Clustering & other multivariate statistical methods. Anomaly detection. **Prerequisite:** *Msi 220 or Mat 108 or equivalent.*

## 2 Class Conduct

The course consists of lectures, analysis of datasets, solution of problems, discussion of homework, discussion of research papers, and book assignments in the general area of Forensic and Assurance practice in the accounting profession.

You are expected to do the readings well ahead of the class. Class time is to be used for the clarification of any doubts that you may have. Do not expect to merely listen to the instructor and gain knowledge. Applied statistics is a practical field backed by robust theory. A good understanding of the theory and its use in practice is essential to excel in the field.

To some extent, this is a hands-on course, and you are required to demonstrate competence in the topics covered in order to receive an acceptable grade. There will be occasional homework assignments and quizzes in addition to a test. You also will be required to discuss assigned research papers before the class and submit briefs.

During the second half of the course you will be required to present a series of papers that I assign. You also will be required to submit a brief, not to exceed 2 pages, on the paper(s) you present.

From time to time I shall give pop quizzes to make sure you have an understanding of the materials that we will cover. I also might give homework which is graded.

A small group project where you will analyse a dataset that you collect, relevant to forensic accounting & assurance.

### 3 Software

The course will use the open source software R system in command line mode extensively. You can download and install on your computer a free copy of the software R from CRAN (Comprehensive R Archive Network) on the internet.

### 4 Course Objectives

A good understanding of the following in the context of forensic and assurance practices in the accounting profession: exploratory data analysis, the language R, decision tree induction and related algorithms, association rule mining and the relevant algorithms, hierarchical as well as partitioning algorithms for clustering, rudiments of multivariate methods such as discriminant analysis, multi-dimensional scaling, and principle components analysis. Use of Benford's Law in Forensic Accounting. Analysis of emails in fraud detection using Social Network Analysis.

### 5 Textbooks

- *The R Book*, Michael J. Crawley (Wiley, 2007), ISBN-10: 0470510242 or ISBN-13: 978-0470510247. ( **C** in the tentative schedule below).
- *Introduction to Data Mining*, Pang-Ning Tan, Michael Steinbach, and Vipin Kumar (Addison Wesley, 2005), ISBN-10: 0321321367 or ISBN-13: 978-0321321367. ( **TSK** in the tentative schedule below).

I shall be covering around 8 selected chapters from the R Book. R Book is an excellent book to have on the desk of any forensic accountant. However, if you find the cost prohibitive, you can rely on classnotes, any of the R tutorials on the internet, and borrow the book from friends.

In addition to the above, I shall assign additional readings from time to time.

## 6 Grading

The final course grade is dependent on the following factors:

- 100 points: Test (In class open book/notes. Details to be announced)
- 75 points: Class presentations and written briefs
- 100 Points: Group Project & written report
- 50 points: Quizzes & Homework (when assigned)
- 25 points: Class participation
- 350 points: Total points (max)

The final course grade is strictly relative, based on the total points scored. The grades, once assigned cannot be changed except in case of errors in grading. Under no circumstances is it possible to do extra credit work to improve the grade.

## 7 Tentative Schedule

- – **Aug 25, 27**
  - **Theme:** Introduction to Descriptive Statistics, Datamining, and R
  - **Readings:** <http://www.albany.edu/acc/courses/acc522fall2008/acc52208252008.ppt>
- **Sept 1: No Class**
- – **Sept 3, 8, 10**
  - **Theme:** Essentials of the R Language
  - **Topics:** Built-in functions and operators; vectors and vector functions; attach, subscripts and indices; sequences, sorting, ranking, ordering; matrices, frames, arrays, and their manipulation; character strings, lists, sets, dates, TIMES, AND pattern matching.
  - **Readings:** C: Ch. 1, 2. Omit pp.47–72.
- – **sept 15, 17**
  - **Theme:** Data Input
  - **Topics:** scan function, read.table, setting working directory, checking files from command line, reading dates and times from files, readLines
  - **Read:** C: Ch. 3.
- – **Sept 22, 24**

- **Theme:** Data Frames
- **Topics:** Subscripts and indices, sorting dataframes, using logical conditions to select rows from dataframes, omitting rows containing missing values, eliminating duplicate rows, selecting variables based on their attributes, merging dataframes, summarising the content of the dataframes
- **Read:** C: Ch.4
- **Sept 29, Oct 1: No Class**
- – **Oct 6,8**
  - **Theme:** Graphics
  - **Topics:** Plots with two variables, scatter plot, bar plot, box plot, histograms, index plots, time series plots, pie charts, pairs and coplot functions, trellis graphics, bubble plots
  - **Read:** C: Ch.5. Omit pp.146-148
- – **October 13, 15 Topic:** *Tables, Introduction to Datamining*
  - **Topics:***Summary tables, table of counts, converting table into dataframes and conversion of tables into dataframes. Distance measures, etc.*
  - **Read:** *C: Ch. 6. Selected parts of TSK: Ch 1-3 to be assigned*
- – **Theme:** Classification: Decision Trees and Model Evaluation
  - **Topics:** General approach to solving classification problems, Algorithm for decision tree induction, Model overfitting and evaluating performance.
  - **Read:** TSK: Ch 4. C: Ch. 21
- – **Oct 27,29**
  - **Theme:** Association Rule Mining
  - **Topics:** Frequent itemset generation, Rule generation, Compact representation of frequent itemsets, Evaluation of association patterns
  - **Read:** TSK: Ch 6.
- – **Nov 3,5**
  - **Theme:** Classification: Clustering **Topics:** Hierarchical (Agglomerative, Divisive) and Partitioning (k-means, Partitioning among medoids)
  - **Read:** TSK: Ch. 8

- – **Nov 10,12**
  - **Theme:** Other Multivariate Methods: A Survey
  - **Topics:** Principal Components Analysis, Discriminant Analysis, Multi-Dimensional Scaling, etc.
  - **Read:** C: Ch.23
- – **Nov 17, 19**
  - **Topic:** E-mails and Social Network Analysis
  - **Read:**
    - \* *Email Archive Analysis Through Graphical Visualization*, Wei-Jen Li, Shlomo Hershkop, and Salvatore J. Stolfo, **VizSEC/DMSEC04, October 29, 2004**
    - \* *Detecting Unusual Email Communication*, P.S. Keila and D.B. Skillicorn (2005)
    - \* *Behavior-Based Modeling and Its Application to Email Analysis*, Salvatore J. Stolfo, Shlomo Hershkop, Chia-Wei Hu, Wei-Jen Li, **ACM Transactions on Internet Technology**, Vol. 6, No. 2, May 2006, Pages 187221.
- **Nov 24: TEST**
- **Nov 26: No Class**
- – **December 1**
  - **Theme:** Benford's Law in Forensic Accounting
  - **Read:** To be assigned
- – **December 3,8**
  - **Theme:** Student presentations