MERRIMACK COLLEGE
GIRARD SCHOOL OF BUSINESS

MERRIMACK COLLEGE
SCHOOL OF SCIENCE & ENGINEERING

Statistics and Statistical Learning

Overview

This course will focus on a set of tools for data analysis but more importantly predictive modeling. Tools can be supervised and unsupervised.

Supervised statistics involves inputs and outputs, such that problems with predicting an output based on one or more inputs.

Unsupervised statistical learning involves the inclusion of inputs but no specific outputs. Problems are dealing with learning and identifying relationships from such data sets.

We will discuss real-world data sets in this class. Here are some examples:

- Wage Data: Wage prediction based on age, education and year. The response variable is a quantitative output. Another term commonly used in the data science world is “data with labels.”
- Stock Market Data (“Data with Labels”): Stock index qualitative prediction (up/down) on a given day by using the five days percentage changes in the index (quantitative inputs).
- Gene Expression Data: We only observe input variables with no corresponding output. The data have 6,830 gene expression measurements for each of 64 cell lines. The goal is to determine whether there are groups among the cell lines based on their gene expression measurements.

In summary, this course will address:

- Prediction and exploratory analysis
- Big Data = Volume + Velocity + Variety
- Data Matrix = Big n, Big p (observations, predictors)
- Supervised + unsupervised + semi-supervised learning (example: text mining)

We will also review linear and non-linear regression models, classification models, cross validation and bootstrap methods and explore tree methods. R (Python modules such as SciKit).

Learn Machine Learning in Python (if preferred by students) will be utilized to build models and produce statistical results.

The textbook's website: www.statlearning.com includes a wealth of resources.

The MASS library and ISLR package in R will be installed for the accessibility of all the data sets that need to be utilized for homework and projects.

We will now revisit some basic matrix notation and introduction to R.

Notation and simple matrix algebra – why? Because, real life data sets can get quite big! Big n (number of observations) and big p (many predictors).

- N x P matrix whose (i,j) th element is x_ij
- Transpose of a matrix
- Vectors
- Addition and multiplication of matrices

Chapter 2 Lab: Review and Introduction to R

R can be downloaded from http://cran.r-project.org/

- Basic Commands:

  Funcname(input1, input2)

  Vectors <- c(1,2,3,4)

  Length( ) of a vector

  Ls(): list of all of the objects

  Rm( ): delete objects not wanted

  Matrix()

  Sqrt() the square root function

  Rnorm(): vector of random variables

  Cor(): correlations

  Mean(), var(), std(): descriptive statistics

- Graphics

- Indexing Data
- Loading Data