UNIVERSITY OF SOUTHERN CALIFORNIA
MARSHALL SCHOOL OF BUSINESS
DATA SCIENCES AND OPERATIONS DEPARTMENT
FALL 2014

DSO 530 – APPLIED MODERN STATISTICAL LEARNING METHODS

COURSE DETAILS

Professor	Dr. Abbass Sharif
Office	BRI 400-E
Email	asharif@marshall.usc.edu
Web	http://www.alsharif.info
Office Hours	Monday and Wednesday from 10:00am to 11.00 am

COURSE OBJECTIVES

This course aims to go far beyond the classical statistical methods, such as linear regression, that are introduced in GSBA 524 (Applied Managerial Statistics). As computing power has increased over the last 20 years many new, highly computational, regression, or "Statistical Learning", methods have been developed. In particular the last decade has seen a significant expansion of the number of possible approaches. Since these methods are so new, the business community is generally unaware of their huge potential. This course aims to provide a very applied overview to such modern non-linear methods as *Generalized Additive Models, Decision Trees, Boosting, Bagging* and *Support Vector Machines* as well as more classical linear approaches such as *Logistic Regression, Linear Discriminant Analysis, K-Means Clustering* and *Nearest Neighbors*.

We will cover all of these approaches in the context of Marketing, Finance and other important business decisions. At the end of this course you should have a basic understanding of how all of these methods work and be able to apply them in real business situations. With the explosion of "Big Data" problems, statistical learning has become a very hot field in many scientific areas as well as marketing, finance and other business disciplines. People with statistical learning skills are in high demand!

To this end, approximately one third of the class time is dedicated to in class labs where the students will work through the latest methods we have covered, on their own laptops, under the supervision of the instructor. These labs will ensure that every student has a full understanding of the practical, as well as the theoretical, aspects of each method.

Several of the approaches we will cover in this course are new even in the statistics community. Hence at the end of this course you will have truly innovative and important applied skills to market and differentiate yourself with.

CASES/DATA SETS STUDIED

We will cover all the following cases/data sets in the course plus some additional interesting applications.

- **Case 1. PREDICTION OF FUTURE MOVEMENTS IN THE STOCK MARKET**: Recently several of the statistical learning methods we discuss in this course (such as Boosting and Bagging) have been used to predict future values of financial markets. Such methods have obvious potential economic implications. We will investigate the performance of these methods on daily movements of the S&P500. We show that, while there is a weak signal, there are clearly some non-linear patterns that we can potentially exploit to predict whether the market will increase or decrease on each given day.
- **Case 2. PREDICTING INSURANCE PURCHASE:** This is a very large data set recording whether a given potential customer purchased insurance or not. For each customer we have a record of 80 different characteristics and we wish to predict which customers are most likely to purchase

insurance. Overall, only 6% of potential customers actually buy the insurance so if we randomly choose people to target our success rate is very low. However, using the methods from this course to target specific people we raise the success rate to around 30% (a five fold improvement).

Case 3. DIRECT MARKETING: This case involves a real dataset of direct mailings to potential donors of a not-for-profit organization. We wish to predict which people are most likely to respond so that the campaign can be better targeted. However, there is an extra wrinkle to this problem because those people that are most likely to respond also tend to give the least while, among those that are unlikely to respond, those that do tend to give the most. Hence we ultimately want to predict dollar giving.

Case 4. HOUSING VALUATIONS: This case involves understanding which variables (both macro and micro) affect housing valuations. For example what is the effect of house size, lot size, neighborhood, number of bedrooms, mortgage rates etc. on the value of a property.

Case 5. MARKETING OF ORANGE JUICE: This is a detailed data set with observations from many customers purchasing one of two brands of OJ. For each transaction many variables are recorded including, prices of each brand, promotions, discounts, which store the purchase was made at etc. The aim is to build a model for predicting which type of OJ a customer will purchase and which variables have an impact on the decision.

Case 6. EMAIL SPAM: Detecting whether an email is a SPAM based on relative frequencies of the 57 most commonly occurring words.

Case 7. Online Search Engine Marketing: We will see how companies try to increase their product sales and acquire more customers through the Internet by analyzing Google Adwords.

COURSE MATERIALS

Statistical Software:

Clearly a statistics package is essential for such an applied course. There are very few packages that can implement all of the different approaches we will cover. Of those that can, most are extremely expensive. The one that we will use in this course is R. R has several advantages. In addition to supporting all of the statistical learning methods we will cover, it is also the package of choice for research statisticians. This means that it is at the cutting edge with respect to new methods. R is also an extremely flexible program. For example, one can use R to write ones own functions to format data or implement new procedures. Finally, R is free so you can easily use it at any company that you may end up at! R can be downloaded from http://www.r-project.org/

Also make sure to install ISLR package, which includes the datasets used in the course book. http://cran.r-project.org/web/packages/ISLR/index.html

Rstudio is a recommended interface for the R software. It is also free, and it runs on Windows, Mac, and Linux operating systems. http://www.rstudio.org

Students are expected to bring their laptops to class during all class sessions (labs and non labs).

Course Book:

The book we will be using for this course covers the technical side of statistical learning with less emphasis on mathematical details. The title is: *"An Introduction to Statistical Learning with Applications in R"* by James, Witten, Hastie, and Tibshirani. The book's website is http://www-bcf.usc.edu/~gareth/ISL/index.html. Also, USC has subscription to Springer, so you should be able to access the book online http://link.springer.com/book/10.1007/978-1-4614-7138-7/page/1.

I would also recommend that you obtain a copy of "*An Introduction to R*" by Venables and Smith which we will use as a manual for learning R. You can either purchase the book ($13 on Amazon!) or download it for free from http://cran.r-project.org/doc/manuals/R-intro.pdf.

EVALUATION (i.e. Grades)

In line with the applied nature of this class, a large portion of the assessment will be made through homework. There will be approximately 8 homework assignments. The homework will contain some theory questions but the majority of the material will involve implementing the different methods that we cover in class using the computer package. There will also be a presentation on a group project and two in class tests. There will be no final exam. The breakdown of grades will be:

Homework (6-8 Homework)	35%
Midterm Exam	25%
Project	15%
Final Exam	25%

STUDENTS WITH DISABILITIES

Any student requesting academic accommodations based on a disability is required to register with Disability Services and Programs (DSP) each semester. A letter of verification for approved accommodations can be obtained from DSP. Please be sure the letter is delivered to me as early in the semester as possible. DSP is located in STU 301 and is open 8:30 am - 5:00 pm, Monday through Friday. The phone number for DSP is 213 740-0776.

COURSE OUTLINE (TENTATIVE)

This is a **tentative** view of the course outline.

Class 1. **Course Introduction**

- Introduction to Modern Statistical Learning Approaches
- Summary of different methods we will cover in the course
- What is Statistical Learning?
    - Inference vs. Prediction
    - Supervised vs. Unsupervised Learning Problems
    - Regression vs. Classification

Class 2. **Lab Class 1: Introduction to R**

- Basic Commands
- Graphics
- Indexing Data
- Loading Data

Class 3. **Assessing the Accuracy of a Statistical Learning Method**

- Less Flexible vs. More Flexible Methods
- Training vs. Test Error Rates
- Nearest Neighbors Methods
- Bayes Classifier
- Bias/Variance ideas

Class 4. **Review of Linear Regression**

- Linear Regression Model
- Using Least Squares to Fit the Model
- Testing Statistical Significance
- Dealing with Categorical Variables

Class 5. **Lab Class 2: Linear Regression**

- Using the *lm()* Function to Fit Linear Regression Models in R

Class 6. **Logistic Regression**

- Using the Logistic Function for Classification
- Estimating Regression Coefficients
- Estimating Probabilities

Class 7. **Linear Discriminant Analysis**

- Bayes Theorem for Classification
- Estimating the Bayes Classifier
- Confusion Matrices
- Quadratic Discriminant Analysis

Class 8. **Lab Class 3: Logistic Regression and LDA**

- Using the *glm()* Function to Fit Logistic Regression Models in R
- Using the *lda()* and *qda()* Functions to Fit LDA in R

Class 9. **Resampling Methods**

- Cross Validation
- The Bootstrap

Class 10. **Lab Class 4: The Cross-Validation and the Bootstrap**

- The validation set approach
- LOOC Validation
- K-Fold Cross Validation

Class 11. **Variable Selection**

- Best Subset Regression
- Leave Out Samples
- BIC and AIC
- Cross Validation
- Illustrations on Real Estate Data

Class 12. **Lab Class 4: kNN, Best Subset Regression**

- Using the *knn()* Function to Implement Nearest Neighbors
- Using the *regsubsets()* Function to Implement Best Subset Regression

Class 13. **Shrinkage and Dimension Reduction Methods**

- Ridge Regression
- LASSO
- Illustrations on the Real Estate Data
- Principal Components Regression
- Partial Least Squares

Class 14. **Lab Class 5: Shrinkage Methods**

- Ridge Regression Using the *lm.ridge()* Function
- LASSO Using the *lars()* Function
- Identifying Important Housing Variables

Class 15. **Review**

- Midterm review and cover what we didn't have time to cover earlier from the class material

Class 16. **Midterm**

Class 17. **Moving Beyond Linear Methods**

- Introduction to Non-Linear Regression
- Polynomial Regression
- Splines
- Illustrations on S&P and Simulated Data Sets

Class 18. **Generalized Additive Models**

- Extending Linear Regression to Allow For Non-Linear Relationships
- Extending Logistic Regression to Allow For Non-Linear Relationships

- Predicting Tomorrow's Change in the S&P Given Movements Over the Last Week

Class 19. **Lab Class 6: Polynomial Regression, Splines and GAM**

- Using the *poly()* Function to Implement Polynomial Regression
- Fitting Splines Using the *smooth.spline()* Function
- Producing a Generalized Additive Model Using the *gam()* Function.
- Illustrations on the S&P Data

Class 20. **Tree Methods**

- Decision Trees
- Regression vs. Classification Trees
- Pruning Trees

Class 21. **Bagging and Boosting**

- Ensemble Classifiers i.e. Using Multiple Classifications to Improve Prediction Accuracy
- The Bootstrap Method
- Using the Bootstrap to Produce a Bagged Classifier
- An Alternative Ensemble Classifier
- AdaBoost and Other Boosting Methods

Class 22. **Lab Class 7: Tree Methods**

- Using the *tree()* Function to Grow Regression and Classification Trees
- Using the *gbm* Package to Implement Boosting Procedures

Class 23. **Support Vector Machines (SVM)**

- The Support Vector Classifier
- Computing the SVM for Classification
- The SVM as a Penalization Method

Class 24. **Lab Class 9**

- Using the *svm()* Function to Produce a Support Vector Machine

Class 25. **Clustering Methods**

- K-means Clustering
- Hierarchical Clustering

Class 26. **Lab Class 10: Clustering**

- Using the *kmeans()* Function to Implement K-means Clustering
- Using the *hclust()* Function to Implement Hierarchical Clustering

Class 27. **Project**

Class 28. **Project Presentations**

Class 29. **Final Exam**

READINGS AND TENTATIVE DATES

(CHECK ON BLACKBOARD FOR DATE CHANGES)

Class	Day	Date	Readings	HW
1	Tues	Aug 26	2.1	
2	Thur	Aug 28	R Lab	HW 1 Assigned
3	Tues	Sep 02	2.2	
4	Thur	Sep 04	3.1- 3.3	HW 1 Due
5	Tues	Sep 09	R Lab	HW 2 Assigned
6	Thur	Sep 12	4.1- 4.3	
7	Tues	Sep 16	4.4, 4.5	HW 2 Due
8	Thur	Sep 18	R Lab	HW 3 Assigned
9	Tues	Sep 23	5.1- 5.2	
10	Thur	Sep 25	R Lab	HW 3 Due and HW 4 Assigned
11	Tues	Sep 30	6.1 - 6.2	
12	Thur	Oct 02	R Lab	HW 4 Due
13	Tues	Oct 07	6.3 - 6.4	HW 5 Assigned
14	Thur	Oct 09	R Lab	
15	Tues	Oct 14	Review	HW 5 Due
MIDTERM	Thur	Oct 16	2-6	
17	Tues	Oct 21	7.1-7.5	
18	Thur	Oct 23	7.6-7.7	
19	Tues	Oct 28	R Lab	HW 6 Assigned
20	Thur	Oct 30	8.1	
21	Tues	Nov 04	8.2	HW 6 Due
22	Thur	Nov 06	R Lab	HW 7 Assigned
23	Tues	Nov 11	9.1- 9.4	
24	Thur	Nov 13	R Lab	HW 7 Due and HW 8 Assigned
25	Tues	Nov 18	10.1-10.3	
26	Thur	Nov 20	R Lab	HW 8 Due
PROJECT	Tues	Nov 25	Project	
PRESENTATIONS **FINAL EXAM**	Tues Thur	Dec 02 Dec 04	Presentation Cumulative	