**Department of Statistics**

**Quaid-i-Azam University, Islamabad**

**Linear Models Dr.Zahid Asghar**

This course aims to provide a theoretical applied overview to General Linear Models, Model Selection (Best subset selection Using Ridge and Shrinkage estimators), 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 Finance, Economics, Marketing, Medical and other related areas. With the explosion of “Big Data” problems, it has become all the important to learn modern applied statistical learning.

**Course Book:**

The book we will be using for this course covers the technical side of statistical learning without a compromise 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://wwwbcf.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>.

**Reference Books:**

Elements of Statistical Learning by James, Witten, Hastie, and Tibshirani

Mastering Metrics by Angrist

Econometrics by Baltagi , 2nd Edition , Springer Verlag (1999)

**Evaluation:**

Homework (3-4 Homework) 10-15%

Quick Quizzes 5%

Midterm Exam 20-25%

Project 10%

Final Exam 50%

**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
* Assumptions

**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 (Finite Sample Theory)**

* Cross Validation
* The Bootstrap

**Class 10. Lab Class 4: The Cross-Validation and the Bootstrap (Finite Sample Theory)**

* 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. General 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 Relationships5**
* **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 27-30 for Project Presentations and Review of Concepts.**