Andrew B. Nobel

STOR 565 Course Information

Class meetings: Tuesday and Thursday 3:30pm – 4:45pm in Hanes 120

Prerequisites: The material in 565 incorporates and makes use of ideas from numerous disciplines, including statistics, optimization, computer science, and mathematics. Students should have completed MATH 233, STOR 215 or MATH 381, and STOR 435, and should be familiar and comfortable with the material in these classes. In addition, students should have some knowledge of basic matrix algebra at the level of MATH 547. Familiarity with the material in STOR 415 (Optimization), STOR 455 (Methods of Data Analysis), and prior experience with basic programming is desirable, but is not required. A more detailed list of select prerequisite material is given below.

Registration: Enrollment and registration for the course is handled online.

Instructor: Andrew B. Nobel

Office: Hanes 308 Email: nobel@email.unc.edu

Nobel Office Hours: Mondays 2:00-3:20pm and Fridays 1-2pm

TA: Alexander Murph

Office: Hanes B-46 Email: acmurph@live.unc.edu

Murph Office Hours: Tuesdays 9:30-10:30am and Fridays 10:30-11:30am

Overview: Broadly speaking, machine learning is the study and development of statistical and computational methods that identify structure in large data sets, or that use existing data to make predictions about new (unseen or partially unseen) data. In most cases, machine learning approaches are based on general models and procedures that are not fine-tuned to the specific problem at hand. Machine learning theory and methods draw on ideas from statistics, optimization, and computer science, with key advances coming from researchers and techniques in each of these fields. Depending on how it is defined, machine learning encompasses or overlaps a number of active research areas, including data mining, the analysis of "big data", artificial intelligence, and deep learning.

Audience and Goals: STOR 565 is a theoretically oriented class that is intended to provide a broad-based introduction to statistical machine learning. The course is targeted to advanced undergraduate and masters students with interest and background in statistics, mathematics, or computer science. The class will present a select, but representative, sample of theory and methods that lie at the core of machine learning. Although the emphasis will be on more statistical material, the class will also deal with basic issues related to computation and optimization. Lectures will focus primarily on the theoretical background and description of different methods. Computer assignments completed outside of the lectures will introduce students to the R programming language, and some elementary methods of machine learning and exploratory data analysis.

Classroom Protocol: Please show up on time, as late arrivals tend to disturb those already present. Reading of newspapers and the use of laptops, tablets, and phones, is not permitted during class.

Primary Text: Introduction to Statistical Learning [ISL]: James et al (2013), Springer (free online, includes R labs). Available at http://www-bcf.usc.edu/~gareth/ISL/

Secondary Texts: Elements of Statistical Learning [ESL] : Hastie et al (2009) Springer (free online) and Machine Learning [MLPP]: A Probabilistic Perspective, by Kevin P. Murphy. (2012). MIT Press.

Attendance: Students are expected to attend all lectures. If you are unable to attend a lecture, please make plans to get the notes from another student in the class.

Homework and Computer Assignments: Homework and computer assignments will be posted on the course web page, and in most cases will be due once a week. The computer assignments are intended to introduce students to the basics of the R

programming language, and the basic statistical machine learning methods discussed in the lectures. The homework assignments are intended to cover and strengthen students' understanding of the theoretical material.

Homework Policy: Homework and computer assignments will be collected *at the beginning* of class on the day that they are due, so please be prepared to turn in your homework at that time. Each assignment will be graded: late/missed assignments will receive a grade of zero. All assignments will have equal weight. In computing a student's overall score for the course, their lowest homework score and lowest computer assignment score will be dropped. This provision is meant to cover exceptional situations in which a student is unable to turn in an assignment due to circumstances beyond his/her control. Students are expected to turn in every homework and computer assignment.

To receive full credit on the homework and computing assignments you should staple together the pages of each assignment in the correct order. Please write your name or initials on each page. In addition, for homework assignments, you should clearly label each problem, neatly show all your work (including your mathematical arguments), and give a clear account of your reasoning in English, using full sentences where appropriate. For computer assignments please avoid printing excessive output, e.g., printing out the entire data set.

You are allowed to discuss the homework and computer assignments with other students, but must prepare each assignment by yourself. Copying of another person's answers or code is not allowed. Any questions regarding the grading of homework or computer assignments should first be addressed to the TA. If you are absent from class when an assignment is returned, you can get your paper from the TA during their office hours.

Final Project: In addition to the homework and computer assignments, there will be a final project that will be towards the end of the semester. For the project students may apply methods covered in class to analysis of one or more data sets of interest, or may investigate in some detail a theoretical direction related to but not covered in the lectures. Students may work on the final project in teams. More details on possible final projects will be made available later in the semester.

Exams: There will be one in-class midterm exam, and a comprehensive final exam, also in-class. All exams will be closed book and closed notes, and without calculators. There will be no makeup exams. The midterm will be given near the middle of the semester. The final exam will be given at the date and time specified in the official University Final Exam Schedule.

Grading: Grading will be based on homeworks, computer assignments, a final project, an in-class midterm, and an in-class final exam, using the weights below.

Homework	10%

Computer Assignments	10%
Final Project	10%
Midterm	30%
Final	40%

Other sources: Look here for updates as the course proceeds.

Honor Code: Students are expected to adhere to the UNC honor code at all times.

Prerequisites

1. Calculus: Basic properties of integration and differentiation. Integration and differentiation of simple functions (e.g. exponential functions, trigonometric functions and polynomials). Integration and partial differentiation of functions of several variables. Taylor series, minima and maxima of functions.

2. Probability: Joint and conditional densities and probability mass functions. Cumulative distribution functions. Random variables. Definition and basic properties expectation, variance, covariance, and correlation. Key discrete and continuous distributions and their basic properties: Bernoulli, binomial, Poisson, geometric; uniform, normal, exponential, gamma. Finding the distribution of a function of a random variable: the CDF method and the general change of variables theorem.

3. Statistics: Sample vs. population quantities. One- and two-sample z- and t-statistics. Basics of point estimation, hypothesis testing and p-values.

4. Linear algebra: Vector spaces, dimension, subspaces. Matrix addition and multiplication, rank. Some knowledge of determinants, inverses, eigenvectors, eigenvalues, symmetric matrices, and non-negative definite matrices.

Tentative Syllabus: The course will begin with some introductory material, basic exploratory data analysis, and a review of inequalities and matrix analysis. The bulk of the remaining course material will be divided into the study of unsupervised methods and supervised methods. The following is a tentative syllabus for the course.

I. Introduction and Preliminaries

Overview of machine learning

- Supervised analysis: Fitting functions to labeled observations
- Unsupervised analysis: Finding patterns and groups in unlabeled data

Review of numerical inequalities

- Maxima, minima, absolute values
- Cauchy-Schwartz

Convexity, machine learning and optimization

- Definition and basic properties of convex sets and functions
- Local and global minima
- The canonical convex program

Introduction to exploratory data analysis

- Univariate sample statistics: mean, median, mode, standard deviation; histograms and density plots; z- and t-statistics
- Bivariate sample statistics: correlation, scatter-plots, r-squared values

II. Unsupervised Methods

Review of matrix algebra

- Symmetric matrices
- Eigenvectors and eigenvalues
- Matrix inverse, determinant, and trace
- Inner products, orthonormal vectors
- Non-negative definite matrices
- Rank
- Outer products

Principal Component Analysis (PCA) and dimension reduction

- Finding good summary directions in high-dimensional data
- Derivation of PCs from eigenvectors of sample variance matrix

The Singular Value Decomposition (SVD)

Clustering

- Overview of the problem, finding group structure in data
- K-means clustering
- Hierarchical clustering, trees and dendrograms
- Applications
- Extensions: co-clustering, biclustering

III. Supervised Methods

Classification

- Introduction. Classification rules, decision regions, decision boundaries.

- Basic marginal and conditional distributions
- Bayes risk and Bayes rule
- Nearest neighbor rules
- Logistic regression
- Linear discriminant analysis, quadratic discriminant analysis
- Maximum margin classifiers
- Support vector machines, separable and non-separable cases

Cross-Validation

- Training and test sets
- Training error vs. test error
- k-fold cross-validation

Linear Regression

- Basic problem
- Ordinary least squares: derivation and some basic properties
- Ridge regression: derivation; shrinkage and regularization

L-1 Penalized Estimation

- The LASSO

IV Selected Topics

The EM Algorithm

Boosting and bagging

Analysis of networks, including community detection and the friendship paradox

Decision and regression trees

Online learning and individual sequences

Multiple testing and the false discovery rate

Disclaimer: The instructor reserves the right to make changes to the syllabus, and to the due dates of assignments in response to unforeseen circumstances. The latter will be announced as early as possible.

Office Hours: If you have questions about the homework assignments or lecture material, please speak with the instructor after class, or during his office hours. If you have questions about the computer assignments, please speak with the TA during his office hours. The instructor and the TA may not be able to respond to emails (including those received shortly before assignments are due), so please begin assignments well before they are due.

Study tips

1. Keep up with the reading and homework assignments. If the reading assignment is long, break it up into smaller pieces (perhaps one section or subsection at a time).

2. Always look over the notes from lecture k before attending lecture k+1. This will help keep you on top of the course material. Ideas from one lecture often carry over to the next: you will get much more out of the material if you can maintain a sense of continuity and keep the “big picture” in mind.

3. Complete the reading *before* doing the homework. Trying to find the right formula or paragraph for a particular problem often takes as much time, and it tends to create more confusion than it resolves.

4. When looking over your notes or the reading assignment, keep a pencil and scratch paper on hand, and try to work out the details of any argument or idea that is not completely clear to you. Even if the argument or idea is clear, it can be helpful to write it down again in a different way in order to test and strengthen your understanding.

5. It is important to know what you know, but it's especially important to know what you don't know. As you look over the reading material and your notes, ask yourself if you (really) understand it. Keep careful track of any concepts and ideas that are not clear to you, and make efforts to master these in a timely fashion.

6. One good way of seeing if you understand an idea or concept is to write down (or state out loud) the associated definitions and basic facts, without the aid of your notes and in complete, grammatical sentences. Translating mathematics into English, and back again, is an important research skill, and a good way to build and assess your understanding.

7. The homework and computer assignments play two important roles in the course. First, they provide an opportunity to actively think about, engage with, and learn the course material. In addition, they provide feedback on your understanding of the material. Carefully look over your corrected assignments. Most students do well on the assignments: even if you received a good score, make sure to note and understand and correct any mistakes you have made.

8. Begin studying for exams at least one week before they are given. Look over your notes, homework, and the text. Write up a study guide containing the main concepts and definitions being covered, and use this to get a clear picture of the overall landscape of the material. For every topic on the study guide, you should know the relevant definitions, motivating ideas, and at least one or two examples.