STATISTICAL METHODS FOR DATA SCIENCE
DS 502/MA 543, Fall 2016

Course Information

DS 502/ MA 543, SL 411	Tuesdays 6:00PM-8:50PM

Instructor

Name: ***Prof. Fatemeh Emdad***	Email: femdad@wpi.edu
Office: Atwater Kent 129 (AK129)	Phone: 508-831-5303
Office Hours	Tuesdays 2:00PM-3:00PM or by Appointment

Description

This course surveys the statistical methods most useful in data science applications. Topics covered include predictive modeling methods, including multiple linear regression, and time series, data dimension reduction, discrimination and classification methods, clustering methods, and committee methods. Students will implement these methods using statistical software. Prerequisites: Statistics at the level of MA 2611 and MA 2612 and linear algebra at the level of MA 2071.

You will need to be able to get your hands dirty playing with, processing, and plotting data using the **R** computer language! The textbook uses **R,** the homework uses **R,** and that will be the officially supported language for the course and all lecture examples will be in **R**. Now, with that being said, this is not intended to be a programming course (i.e., your code will not be graded), but actually working with data will be extremely important (i.e., the results of the code will be graded)!

Teaching Assistant

Name: Binod Manandhar	Email: bmanandhar@wpi.edu
Office: SH 205	Phone: 508-395-4854
Office Hours	Fridays 9am-12noon (Tutoring center 2hrs possible)
Name: Chong Zhou	Email: czhou2@wpi.edu
Office: Now AK013 later AK123	Phone: TBA
Office Hours	2pm-3:30pm Thursdays and Fridays

High level course goals and learning objectives

By the end of the class you should be able to:

- *Use tools* such as Linear Regression, Logistic Regression, Trees, etc. for making predictions.
- *Explain* the pros and cons of various approaches.
- *Avoid* common pitfalls such as over fitting and data snooping.
- *Assess* the validity of the prediction, given a prediction generated from such a method.
- *Diagnose "*what can go wrong with a prediction?"

As you develop judgment and skill in the appropriate use of tools (including paper, pencil, and computers) never lose sight of the following fact:

The absolute most important tool for doing mathematics is your brain.

The numbers, graphs, and symbols produced by a computer, a software tool, or a website do not always provide useful and accurate information. Knowing how to use electronic devices is important, but it is more important to use care and judgment in entering and manipulating data and interpreting output.

Textbook
An Introduction to Statistical Learning, by G. James, D. Witten, T. Hastie, R. Tibshirani
If you have access to the WPI library then a PDF of the book can be downloaded for free from Springer. Just search for the title at the WPI library web page and then click on the ebook version.

Recommended texts
Other texts that would be useful for the course are:
● Linear Algebra and Its Applications, by D. Lay. This has been used as the textbook for MA2071 (one of the requirements for the course).
● Applied Statistics for Engineers and Scientists, by J. Petruccelli, B. Nandram, and M. Chen. This has been the textbook for MA2611 and MA2612 (the other requirement for the course).
● The Elements of Statistical Learning: Data Mining, Inference, and Prediction, by T. Hastie, R. Tibshirani, and J. Friedman. This is the "big brother" of our textbook, and a great resource that covers a lot of interesting material.
● Learning From Data, by Y. S. AbuMostafa, M. MagdonIsmail, and H. Tien Lin. This book is used in the Caltech "Learning from Data" course.
● Learning R: A StepbyStep Function Guide to Data Analysis, by Richard Cotton O'Reilly Media, September 2013.

Evaluation/Grades
Final grades will be determined based upon the following breakdown:

➢ Homework & Lab (2 person team)	40%
➢ Midterm exam	20%
➢ Final exam	20%
➢ Final project (3-5 person teams)	20%
---	---
Total	100%

The midterm exam and final exam will be in class, but **no collaboration will be allowed** and the exams will be graded based upon demonstrated understanding of key concepts.
The homework problems will be performed in **groups of at most two** and will be graded for demonstrated understanding of key concepts and quality of presentation. You can choose your own teammate. The final project will be performed in **groups of 3-5** and will be graded based upon the quality and completeness of a final presentation and final report.

Honor Code
By submitting any assignment for evaluation, you implicitly agree with the following statement:

I have neither given nor received unauthorized assistance on this assignment.

It is your responsibility as a member of this class to understand the nature and implications of the Honor Code.

Make-up Exam Policy
Makeup exams will only be allowed in the event of a documented emergency. The exam dates are listed on the syllabus and you are responsible for avoiding conflicts with the exams.

Late Assignment Policy
In general, late assignments will either not be accepted or, at best, be heavily penalized. If an emergency arises or you know in advance about a conflict, please let Prof. Emdad and your TA know as soon as possible.

Collaboration and Academic Honesty Policy
Collaboration is prohibited on the exams. Collaboration is encouraged on homework and the final project. Homework will be conducted in teams of one or two. You will also be allowed to select your own teams of 3-5 for the final project. On homework, you **may** discuss problems across teams, but each homework team is responsible for generating solutions and writing up results on their own **from scratch**. On the final project, each of the teams will be using their own data sets, but the same collaboration policy applies. All violations of the collaboration policy will be handled in accordance with the WPI Academic Honesty Policy.
As examples, each of the following would be a violation of the collaboration policy (this list is **not** exhaustive):
● Two different homework teams share a solution to any assigned problem.
● One homework or project team allows another homework or project team to copy any part of a solution to an assigned problem.
● Any code or plots are shared between homework or project teams.
As examples, each of the following would not be a violation of the collaboration policy:
● Students within a team sharing solutions and code for a problem.
● Students from different teams discussing an assignment at the level of goals, where ideas for solutions can be found in the book or notes, what parts are more challenging, or how one might approach the problem.

Accommodation for Special Needs or Disabilities
If you need course adaptations or accommodations because of a disability, or if you have medical information to share with me, please make an appointment with me as soon as possible. If you have not already done so, students with disabilities who believe that they may need accommodations in this class are encouraged to contact the Office of Disability Services as soon as possible to ensure that such accommodations are implemented in a timely fashion. This office is located in the West St. House (157 West St), (508) 8314908.

Accommodation for Religious Observance
Students requiring accommodation for religious observance must make alternate arrangements with Prof. Emdad at least one week before the date in question.

Personal Emergencies
In the event of a medical or family emergency, please contact Prof. Emdad to work out appropriate accommodations.

Tentative Course Outline

This course outline is to be used as a guide to where we plan to be at any given time during the semester. Occasionally, material may be modified, added, or deleted. You are responsible to follow up with any changes, on the other hand, I reserve the right to change the order and content of lectures to improve the learning experience for the course. I will ensure that the homework and exams match the material actually covered.

Week #	Day	Material Covered	HW
Week 1	Aug. 30	Course intro. Sec. 2.1, 2.2	
Week 2	Sep. 6	Linear Regression 1. Sec. 3.1, 3.2	HW1 assigned
Week 3	Sep. 13	Linear Regression 2. Sec. 3.3, 3.4, 3.5 Time series methods	
Week 4	Sep. 20	Classification Sec. 4.1, 4.2, 4.4, 4.5	HW1 Due HW2 Assigned
Week 5	Sep. 27	Resampling Sec. 5.1, 5.2	**Project proposal ideas Due**
Week 6	Oct. 4	Model Selection and Regularization Sec. 6.1, 6.2	HW2 Due HW3 Assigned Project definition assigned
Week 7	Oct. 11	***Review for the midterm, Midterm exam***	
Week 8	Oct. 25	Dimension Reduction Sec. 6.3, 6.4	HW3 Due HW4 Assigned
Week 9	Nov. 1	Nonlinear methods Sec. 7.1, 7.4	**Project proposals Due**
Week 10	Nov. 8	Nonlinear methods Sec. 7.5, 7.7	HW4 Due HW5 Assigned
Week 11	Nov. 15	Tree methods Sec. 8.1, 8.2	
Week 12	Nov. 29	SVM (Support Vector Machine) Sec. 9.1, 9.2, 9.3	HW5 Due
Week 13	Dec. 6	***Review for the final, Final exam***	
Week 14	Dec. 13	Project presentations/posters	Project report due

I wish a pleasant and successful semester for all of you.

Prof. Emdad