DS 502/MA 543

STATISTICAL METHODS FOR DATA SCIENCE

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 MA2612 and linear algebra at the level of MA 2071.

Where and When

Tuesdays and Thursdays from 4:00pm-5:15pm - SL105

Instructor information

Prof. Randy Paffenroth
Office location: AK124
Office hours: 5:30pm-6:30pm on Tuesdays and Thursdays (right after class).
Other times are available by appointment, and walk-ins are always welcome if I am around and not otherwise indisposed.

Best ways to contact me:

- WPI email: rcpaffenroth@wpi.edu
- Office phone: (508) 831-6562

I should be able to turn around email questions relatively quickly 9am-5pm, Monday-Friday. My availability at night and on weekends is more limited and I certainly check my email far more infrequently, but you may feel free to try and contact me.

Teaching Assistant/Grader

TBD

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 from data.
- *Explain* the pros and cons of various approaches.
- *Avoid* common pitfalls such as overfitting and data snooping.
- Given a prediction generated from such a method, be able to *assess* the validity of the prediction.
- *Diagnose* what can go wrong with a prediction.

Recommended background for course

The recommended background for the course are statistics at the level of MA 2611 and MA2612 and linear algebra at the level of MA 2071.

In particular, you will need to know some linear algebra:

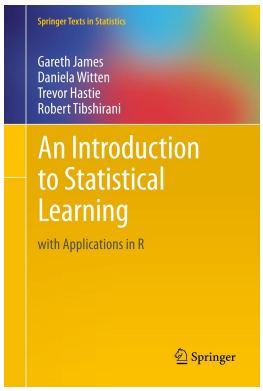
- Vectors (that they can represent points in space, column vs. row, etc.)
- Matrices (transposes, that they don't commute, etc.)
- Inner products
- Least squares
- How to solve linear systems
- etc.

You will also need to know some probability and statistics

- Random variables (what they represent, etc.)
- Descriptive statistics (mean, variance, etc.)
- Hypothesis testing
- Estimation and prediction
- etc.

You will need to be able 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)!

Textbook



An Introduction to Statistical Learning
Gareth James, Daniela Witten, Trevor Hastie, Robert 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 David Lay. This has been used as the textbook for MA2071 (one of the requirements for the course).
- Applied Statistics for Engineers and Scientists, by Joseph Petruccelli, Balgobin Nandram, and Minghui 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 Trevor Hastie, Robert Tibshirani, and Jerome Friedman. This is the “big brother” of our textbook, and a great resource that covers a lot of interesting material.
- Learning From Data, by Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. This book is used in the Caltech “Learning from Data” course and does a great job covering things like cross validation and VC dimension.
- Learning R: A Step-by-Step 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:

Homeworks (5 assignments, 2 person teams)	20%
Midterm exam	20%
Final project (3-5 person teams)	30%
Final exam	30%

The midterm exam and final exam will be in class, cumulative, and open note, but **no collaboration will be allowed** and the exams be graded based upon demonstrated understanding of key concepts. For each exam, you are allowed to bring in up to four (4) 8 ½ by 11 sheets of paper (either printed or handwritten) with whatever notes you want for the exam.
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, but team changes will need to be approved by Prof. Paffenroth. 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.
I reserve the right to curve the final grades (either up or down) based upon the aggregate performance of the class.

Make-up Exam Policy

Make-up exams will only be allowed in the event of a documented emergency or religious observance. 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 (50% of possible points). If an emergency arises or you know in advance about a conflict please let Prof. Paffenroth know as soon as possible.

Collaboration and Academic Honesty Policy

Collaboration is prohibited on the exams. Collaboration is encouraged on homeworks and the final project. Homeworks 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 homeworks 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.
- Of course, you can ask Prof. Paffenroth any questions you like, show him code, etc.

If there is any doubt as to what is allowed and what is not allowed, please just ask!

Schedule

On this schedule the homework, exam, and final project dates are fixed. 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 homeworks and exams match the material actually covered.

	Tuesday	
Class 1&2	January 17 & 19	Course introduction Section 2.1 Section 2.2
Class 3&4	January 24 & 26	Linear Regression 1 Section 3.1 Section 3.2 HW 1 assigned
Class 4&5	January 31 & February 2	Linear Regression 2 Section 3.3 Section 3.4 Section 3.5 Time series methods
Class 6&7	February 7 & 9	HW 1 due Classification Section 4.1 Section 4.2 Section 4.4 Section 4.5 HW 2 assigned
Class 7&8	February 14 & 16	Resampling Section 5.1 Section 5.2
Class 9&10	February 21 & 23	HW 2 due Model Selection and Regularization Section 6.1 Section 6.2 HW 3 assigned Project definition assigned
Class 11&12	February 28 & March 2	Review for the midterm Midterm exam
	March 7 & 9	Term break

Class 13&14	March 14 & 16	HW 3 due Dimension Reduction Section 6.3 Section 6.4 Johnson-Lindenstrauss/concentration of measure HW 4 assigned
Class 15&16	March 21 & 23	Project proposals due Nonlinear methods Section 7.1 Section 7.4 Section 7.5 Section 7.7
Class 17&18	March 28 & 30	HW 4 due Tree methods Section 8.1 Section 8.2 HW 5 assigned
Class 19&20	April 4 & 6	SVM Section 9.1 Section 9.2 Section 9.3
Class 21&22	April 11 & 13	HW 5 due Unsupervised Learning Section 10.2 Section 10.3 Non-linear dimension reduction
Class 23&24	April 18, 20, & 25	Special topics Project presentations/posters Project report due
Class 14	April 7 & May 2	Review for the final Final exam

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) 831-4908.

Accommodation for Religious Observance

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

Personal Emergencies

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