Data mining is the science of discovering structure and making predictions in large, complex data sets. Nowadays, almost every organization collects data, which they hope to use to support improved decision making. Learning from data can enable us to better: detect fraud, make accurate medical diagnoses, monitor the reliability of a system, perform market segmentation, improve the success of marketing campaigns, and much, much more.

This course serves as an introduction to Data Mining for students in Business and Data Analytics. Students will learn about many commonly used methods for predictive and descriptive analytics tasks. They will also learn to assess the methods' predictive and practical utility.

By the end of the class, students will learn to:

- Use R to run many of the commonly used data mining methods
- Understand the advantages and disadvantages of various methods
- Compare the utility of different methods
- Reliably perform model/feature selection
- Use resampling-based approaches to assess model performance and reliability
- Perform analyses of real world data

License

All of the course materials on this page are licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Textbooks

Required textbook

There is one required textbook in this class. It is available for free at the link below. If you find the textbook to be useful, please show your appreciation.
by purchasing a copy for personal use.

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani
  An Introduction to Statistical Learning: with Applications in R

Recommended textbooks

In addition to the required text, the following references are highly recommended. Students may find it useful to own a personal copy of one or two of the texts below.

- Witten and Frank, Data Mining: Practical Machine Learning Tools and Techniques
- Hastie, Tibshirani, Friedman, Elements of Statistical Learning
- Provost and Fawcett, Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking
- Kuhn and Johnson, Applied Predictive Modeling

Helpful resources

There are many resources online that may help you with various parts of the class.

Learning R

Here are some resources to help you learn R if you don't know it already.

- RStudio
- R for Data Science
- swirl: Learn R, in R
- 94-842, My R Programming class
- Introduction to R Markdown
- R Style guide
- ggplot2 cheatsheet

Course Work

Your grade in this course will be determined by a series of 5 weekly homework assignments, lab participation, two exams, and a final project
Assignments (25%) 

Weekly assignments will take the form of a single R Markdown file: namely, code snippets integrated with captions and other narrative. Unless otherwise indicated, all assignments are due at the start of class (10:30AM) on the dates indicated on the Schedule below.

Your assignment score for the course will be calculated by averaging your four (4) highest homework scores. That is, your lowest homework score will not count toward your grade.

While the homework assignments may vary in length and/or difficulty, each will be graded out of a possible 20 points

Lab participation (10%) 

In addition to the two lectures, there is a weekly lab session that meets in HBH 1000 from 4:30 - 5:30PM each Friday. Lab attendance is mandatory and counts for 10% of your final grade. During the 1 hour lab section, students will get hands-on practice with the week's material by completing a set of structured data analytic exercises. Tasks may include but are not limited to: running or modifying code from the lecture, running methods, creating visualizations, writing short reports.

There is a Lab every Friday, with the exception of the last week of class. Thus there are a total of 6 Lab sessions. The 4th session is reserved for an in-class midterm, and therefore does not count toward your participation score. Your participation score for the course will be calculated based on the number of "regular" (non-midterm) lab sessions you attend and participate in as specified by the table below.

\[
\begin{array}{c|ccccc}
\text{Labs attended} & 0 & 1 & 2 & 3 & 4-5 \\
\text{Points (max = 10)} & 0 & 2.5 & 5 & 7.5 & 10 \\
\end{array}
\]

Midterm exam (15%) 

The Midterm exam will take place from 4:30 - 5:50PM on Friday, February 10, in HBH A301.

Only material covered during the first 3 weeks of class is eligible for the midterm exam.

The midterm exam will take the form an open book written test. The test will consist of several problems. Just about every problem will be TRUE/FALSE, Multiple choice, or a "and explain your answer" variant of such questions.

Sample question. Linear regression is only useful if you're certain that the true relationship between Y and your inputs X is linear. TRUE or FALSE? In a sentence or two, explain your answer.

General comment: The midterm is intended to assess your conceptual understanding of the material we covered in the first 3 weeks of class. Because the
test is open note, I will not be asking questions where the answer is explicitly written out in the notes. E.g., I will not ask you to write out a step-by-step description of Cross-validation.

However, I could ask you something like: Suppose that we have \( n = 2000 \) observations and we perform 20-fold Cross-validation. How many observations are used for Training at each step? (Answer: There will be \( 2000 / 20 = 100 \) observations in each Fold, so 1900 observations will be used for training and 100 for testing at each step).

**Final exam (20%)**

The time for the final exam is set by the University. Please check the official calendars for the latest time and date information.

The final exam will be a **closed book** written exam. This exam is intended to test your complete knowledge of the concepts and methods covered in the class.

**Final project (30%)**

This will be a data analysis project to be conducted in groups of 2-4 students. More details to follow.

Regardless of grading basis, students must receive a score of at least 50% on the final project in order to pass the class.

**Course Grading**

Your final course grade will be calculated according to the following breakdown.

<table>
<thead>
<tr>
<th>Component</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assignments</td>
<td>25%</td>
</tr>
<tr>
<td>Lab participation</td>
<td>10%</td>
</tr>
<tr>
<td>Midterm exam</td>
<td>15%</td>
</tr>
<tr>
<td>Final exam</td>
<td>20%</td>
</tr>
<tr>
<td>Final project</td>
<td>30%</td>
</tr>
</tbody>
</table>

**Late submission**

Homework is to be submitted before the **start of class (1:20PM)** on the due date indicated.

Late homework will not be accepted for credit.
Note that your lowest homework score will not count toward your grade, so you can miss one homework without it counting toward your course grade.

**Collaboration**

You are encouraged to discuss homework problems with your fellow students. However, the work you submit must be your own. You must acknowledge in your submission any help received on your assignments. That is, you must include a comment in your homework submission that clearly states the name of the student, book, or online reference from which you received assistance.

Submissions that fail to properly acknowledge help from other students or non-class sources will receive no credit. Copied work will receive no credit. Any and all violations will be reported to Heinz College administration.

All student are expected to comply with the CMU policy on academic integrity. This policy can be found online at [http://www.cmu.edu/academic-integrity/](http://www.cmu.edu/academic-integrity/).

The course collaboration policy allows you to discuss the problems with other students, but requires that you complete the work on your own. Every line of text and line of code that you submit must be written by you personally. You may not refer to another student's code, or a "common set of code" while writing your own code. You may, of course, copy/modify lines of code that you saw in lecture or lab.

The following discussion of code copying is taken from the Computer Science and Engineering Department at the University of Washington. I discussed these issues early on in class, and they are also covered in some form in the academic guidelines for CMU and Heinz College.

"[It is] important to make sure that the assistance you receive consists of general advice that does not cross the boundary into using code or answers written by someone else. It is fine to discuss ideas and strategies, but you should be careful to write your programs on your own."

"You must not share actual program code with other students. In particular, you should not ask anyone to give you a copy of their code or, conversely, give your code to another student who asks you for it; nor should you post your solutions on the web, in public repositories, or any other publicly accessible place. [You may not work out a full communal solution on a whiteboard/blackboard/paper and then transcribe the communal code for your submission.] Similarly, you should not discuss your algorithmic strategies to such an extent that you and your collaborators end up turning in [essentially] the same code. Discuss ideas together, but do the coding on your own."

"Modifying code or other artifacts does not make it your own. In many cases, students take deliberate measures -- rewriting comments, changing variable names, and so forth -- to disguise the fact that their work is copied from someone else. It is still not your work. Despite such cosmetic changes, similarities between student solutions are easy to detect. Programming style is highly idiosyncratic, and the chance that two submissions would be the same except for changes of the sort made easy by a text editor is vanishingly small. In addition to solutions from previous years or from other students, you may come across helpful code on the Internet or from other sources outside the class. Modifying it does not make it yours."

"[I] allow exceptions in certain obvious instances. For example, you might be assigned to work with a project team. In that case, developing a solution as a
team is expected. The instructor might also give you starter code, or permit use of local libraries. Anything which the instructor explicitly gives you doesn't normally need to be cited. Likewise, help you receive from course staff doesn't need to be cited."

If you have any questions about any of the course policies, please don't hesitate to ask. You may post your questions on Piazza or ask me directly.

Policies

Computing:

The statistical computing package we will use in this course is R, which is available on many campus computers. You may download your own copy from http://www.r-project.org. We require that you use R Markdown to complete your assignments, which is enabled very nicely with RStudio.

Laptop Policy:

Students must bring their own laptops to the Friday lab sessions.

Communication:

Assignments and class information will be posted on Blackboard and the class website. Help with using Blackboard is available at www.cmu.edu/blackboard/help/.

Email:

The Piazza forum should be used for general course-related questions that may be of interest to others in the class. For other types of questions (e.g., to report illness, request various permissions) please contact Prof. Chouldechova via email.

Please include the course code 95791 in the subject line of your email.

Disability Services:

If you have a disability and need special accommodations in this class, please contact the instructor. You may also want to contact the Disability Resources office at 8-2013.

Tentative Schedule
Week 1: Introduction, Regression++

Part I

What is Data Mining?

Course logistics

What are predictive analytics (supervised learning)?

What are descriptive analytics (unsupervised learning)?

Introduction to the central themes of the class

Part II

Linear regression as a predictive tool

Wed 01/18 Polynomial regression

Step functions

Suggested reading

- ISLR §2.1
- ISLR §3.1, 3.2, 3.3, 3.4
- ISLR §7.1, 7.2
- 94-842 Lecture 9: Linear regression in R
- 94-842 Lecture 10: Factors and interactions in linear regression

Links

[Lecture 1 notes] [Rmd code] [html]
Lab 1: [Rmd] [html]
Lab 1: Solutions [Rmd] [html]

Fri 01/20
- Introduction to R, RStudio, R Markdown
- Linear regression in R

Week 2: Model selection and validation in regression

Part I

Splines
Additive models
Local regression

Part II

Bias-Variance trade-off
Testing-training

Wed 01/25
Cross-validation

Suggested reading

- ISLR §7.4, 7.5.1, 7.7.1
- ISLR §2.2.1, 2.2.2
- ISLR §5.1, 5.2
- GAMs R tutorial

Links

[Lecture 2 notes]
Lab 2: [Rmd] [html]
Lab 2: Solutions [Rmd] [html]

Fri 01/27

- Validation, Cross-validation in R
- Splines, additive models

Week 3: Model Selection, Classification

Part I

Model selection in regression

Subset selection

Regularized regression

AIC/BIC

Part II

Introduction to classification

Wed 02/01

Bayes classifier

Logistic regression

HW 2

Links: Suggested reading:

- ISLR §6.1, 6.2
- ISLR §5.3.4
- ISLR §2.2.3
- ISLR §4.1, 4.2, 4.3
Fri 02/03
- Best subset, Forward, and Backward variable selection
- AIC, BIC
- Validation and Cross-validation for variable selection
- Lasso

**Week 4: Classification**

**Part I**

- Logistic regression decision boundary
- k-Nearest Neighbours
- Linear Discriminant Analysis

**Part II**

- Quadratic Discriminant Analysis
- Naive Bayes

Wed 02/08

**Assessing performance of classifiers**

- Calibration plots
- Confusion matrices
- Cost-based assessment
- ROC, AUC

**Suggested reading:**
• ISLR §2.2.3
• ISLR §4.4, 4.5
• ISLR §5.1.5
• APM Chapter 11: Measuring Performance in Classification Models

Links:

[Lecture 4 notes]
[pROC package examples]

Fri 02/10  Midterm exam

Week 5: Tree-based methods, Advanced methods

Part I

Assessing performance of classifiers
Decision trees

Part II

Decision Trees
Bagging

Wed 02/15

Random forests

Final project assigned.

Suggested reading:

• APM Chapter 11: Measuring Performance in Classification Models
• ISLR §8.1, 8.2

Links:
Fri 02/17

- Classification and Regression trees

**Week 6: Unsupervised learning**

**Part I**

Random Forests

Boosting

Bootstrap SE estimates, CI's

**Part II**

What is Unsupervised learning?

K-means clustering

Wed 02/22

Hierarchical clustering

Association rule mining

**Suggested reading:**

- ISLR §8.1, 8.2
- ISLR §5.3.4
- ISLR §10.1, 10.3

**Links:**
Fri 02/24

- Random forests
- Boosting
- K-means, Hierarchical Clustering

**Week 7: Unsupervised learning**

**Part I**

What is Unsupervised learning?

K-means clustering

Hierarchical clustering

Association rule mining

**Part II**

Gaussian mixture models

Dimensionality reduction

Principal components regression

**Suggested reading:**

- ISLR §10.2

**Links:**
Lecture 7 notes
Review session
Fri 03/03
[Wrap-up lecture] [Review slides]

Course Info

Instructor: Prof. Alexandra Chouldechova
yyy@cmu.edu, where yyy=achould
HBH 2105D

Office Hours:
Prof. Chouldechova T 10AM - 12PM, (HBH 2224)
Zalak Ghetia W 12:30 - 2:30PM (HBH 2108)
Xinyao Li T 4:30 - 6:30PM (Heinz lower lounge)
Abhinav Maurya M 7-9PM (HBH 3034)

Teaching Assistants:
Zalak Ghetia
yyy@andrew.cmu.edu, yyy=zghetia
Xinyao Li
yyy@andrew.cmu.edu, yyy=xinyaol
Abhinav Maurya
yyy@cmu.edu, yyy=ahmaurya

Class Meetings:
W 6:00 - 8:50, A301 A301 (B3)
F 4:30 - 5:30, HBH A301 (B3)

This Website:
http://www.andrew.cmu.edu/~achoulde/95791/

All course materials will be posted on this site.
Homework submission:
Assignments to be submitted via Blackboard.

Prerequisites:
Students must be enrolled in a graduate program in Heinz College. Special permission can be granted by the College.

"Homework Zero"

Due: Friday, January 20, 11:59pm Please fill out the following survey.

Homework 1 [Rmd] [html]
Due 5:50PM, Wednesday, January 25

Homework 2 [Rmd] [html]
Due 5:50PM, Wednesday, February 1

Homework 3 [Rmd] [html]
Due 5:50PM, Wednesday, February 8

Homework 4 [Rmd] [html]
Due 4:20PM, Friday, February 17

Homework 5 [Rmd] [html]
Due 5:50pm, Wednesday, March 1
Final Project [Description]

Due 6:00PM, Friday, March 10

Copyright (c) 2017 CMU. All rights reserved. Design by Free CSS Templates.