**Course Description**

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](https://creativecommons.org/licenses/by-nc-sa/4.0/).

**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](https://www.stat.berkeley.edu/~hastie/ElemStatLearn/)

**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.

- Hastie, Tibshirani, Friedman, [Elements of Statistical Learning](https://web.stanford.edu/~hastie/Papers/ESLII.pdf)
- Provost and Fawcett, [Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking](https://www.amazon.com/Data-Science-Business-You-Know/dp/1119095061)

**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 (20%)

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 **before the start of the Thursday class session (2:50PM)** 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 1002 from **4:50 - 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.

<table>
<thead>
<tr>
<th>Labs attended</th>
<th>Points (max = 10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>2.5</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>7.5</td>
</tr>
<tr>
<td>4-5</td>
<td>10</td>
</tr>
</tbody>
</table>

Midterm exam (15%)

The Midterm exam will take place from **4:30 - 5:50PM on Friday, February 9, in HBH 1002**.

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 (30%)**

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.

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

**Final project (25%)**

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

**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>20%</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>30%</td>
</tr>
<tr>
<td>Final project</td>
<td>25%</td>
</tr>
</tbody>
</table>

**Late submission**

Homework is to be submitted by **2:50PM** 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 students 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 Canvas and the class website.

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

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Due</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1: Introduction, Regression++</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

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]

Friday

- 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

03/25 - 03/29

Bias-Variance trade-off

Testing-training

Cross-validation

Suggested reading

- ISLR §7.4, 7.5.1, 7.7.1
Week 3: Model Selection, Classification

Part I

Model selection in regression

Subset selection

Regularized regression

AIC/BIC

Part II

Introduction to classification

04/01 - 04/05

Bayes classifier

Logistic regression

Links: Suggested reading:

- ISLR §6.1, 6.2
- ISLR §5.3.4
- ISLR §2.2.3
- ISLR §4.1, 4.2, 4.3

Links:

Week 4: Classification

Part I

04/08 - 04/12

Logistic regression decision boundary

k-Nearest Neighbours

HW 3
Linear Discriminant Analysis

Part II

Quadratic Discriminant Analysis

Naive Bayes

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 ]

Friday, April 12

Midterm exam

Week 5: Tree-based methods, Advanced methods

Part I

Decision trees

Part II

Decision Trees

Bagging

Random forests

Final project assigned.

04/15 - 04/19 HW 4

Suggested reading:

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

Links:

[ Lecture 5 notes ] [ Final project ] [ Project descriptions ]

Lab 4: [ Rmd ] [ html ]

Lab 4: Solutions

- 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
Hierarchical clustering
Association rule mining

Suggested reading:

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

Links:

[Lecture 6 notes]
Lab 5: [Rmd] [html]
Lab 5: Solutions

Friday

- 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

04/29 - 05/03

Gaussian mixture models
Dimensionality reduction
Principal components regression

Suggested reading:
Course Info

Instructor: [Prof. Alexandra Chouldechova](mailto:yyyy@cmu.edu)  
yyy@cmu.edu, where yyy=achould  
HBH 2224

Office Hours: See Piazza.

Teaching Assistants:
- **Alton Lu**  
  yyyy@andrew.cmu.edu, yyyy=altonl
- **Andrew Olson**  
  yyyy@andrew.cmu.edu, yyyy=atolson
- **David Wang**  
  yyyy@andrew.cmu.edu, yyyy=minhengw
- **Jun Zhang**  
  yyyy@andrew.cmu.edu, yyyy=junz3

Class Meetings:
- TR 3:00 - 4:20PM, HBH 1202 (B4)  
- F 4:30 - 5:30PM, HBH 1002 (B4)

This Website:  
[http://www.andrew.cmu.edu/~achoulde/95791/](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 1 [TBA]

Due 2:50PM, Thursday, March 28

Homework 2 [TBA]

Due 2:50PM, Thursday, April 4

Homework 3 [TBA]

Due 2:50PM, Thursday, April 11
Homework 4 [TBA]

Due 2:50PM, Thursday, April 18

Homework 5 [TBA]

Due 2:50PM, Thursday, April 25

Final Project [TBA]

Due TBD

Design by Free CSS Templates.