Web Mining and Recommender Systems

Introduction
What is this class about?

In this course we will build models that help us to **understand data** in order to gain **insights** and make **predictions**.
Examples – Recommender Systems

**Prediction:** what (star-) rating will a person give to a product? 
e.g. rating(julian, Pitch Black) = ?

**Application:** build a system to recommend products that 
people are interested in

---

103 of 115 people found the following review helpful

⭐⭐⭐⭐⭐ Excellent Sci-Fi
Pitch Black was arguably one of the most overlooked films of the early 
year. Although the setting of the film could seem routine to a casual 
viewer (space travelers stranded and bickering on a hostile planet infested 
with alien nasties), director David Twohy's wonderful use of color and 
stylistic flourishes more than makes up for any trivial complaints. 
For...
[Read the full review >](#)
Published on September 12, 2000 by Eric J. Pray

**Insights:** how are opinions influenced by factors like time, 
gender, age, and location?
Examples – Social Networks

**Prediction:** whether two users of a social network are likely to be friends

**Application:** “people you may know” and friend recommendation systems

**Insights:** what are the features around which friendships form?
**Examples** – Advertising

**Prediction:** will I click on an advertisement?

**Application:** recommend relevant (or likely to be clicked on) ads to maximize revenue

**Insights:** what products tend to be purchased together, and what do people purchase at different times of year?
Examples – Medical Informatics

**Prediction:** what symptom will a person exhibit on their next visit to the doctor?

**Application:** recommend preventative treatment

**Insights:** how do diseases progress, and how do different people progress through those stages?
1. Are the data associated with meaningful outcomes?
   • Are the data **labeled**?
   • Are the instances (relatively) independent?

   e.g. who likes this movie?

   Yes! “Labeled” with a rating

   e.g. which reviews are sarcastic?

   No! Not possible to objectively identify sarcastic reviews
What we need to do data mining

2. Is there a clear objective to be optimized?
   • How will we know if we’ve modeled the data well?
   • Can actions be taken based on our findings?

   e.g. who likes this movie?

   How wrong were our predictions on average?

   \[
   \frac{1}{N} \sum_{i=1}^{N} (r_{u,i} - \text{prediction}(u,i))^2
   \]
What we need to do data mining

3. Is there enough data?
   • Are our results statistically significant?
   • Can features be collected?
   • Are the features useful/relevant/predictive?
What is covered in this course?

This course aims to teach

• How to **model** data in order to make **predictions** like those above
• How to **test and validate** those predictions to ensure that they are meaningful
• How to **reason about** the findings of our models

(i.e., “data mining”)
What is covered in this course?

But, with a focus on applications from recommender systems and the web

- **Web** datasets
  - Netflix
  - Amazon
  - Reddit
  - Yelp
  - BeerAdvocate
  - RateBeer
  - Epinions.com

- Predictive tasks concerned with human activities, behavior, and opinions (i.e., recommender systems)
**Basic** data processing

- Text manipulation: count instances of a word in a string, remove punctuation, etc.
- Graph analysis: represent a graph as an adjacency matrix, edge list, node-adjacency list etc.
- Process formatted data, e.g. JSON, html, CSV files etc.
**Basic** mathematics

- Some linear algebra: $Ax = y \rightarrow x = (A^T A)^{-1} A^T y$
- Some optimization
- Some statistics (standard errors, p-values, normal/binomial distributions)
All coding exercises will be done in **Python** with the help of some libraries (numpy, scipy, NLTK etc.)
Expected knowledge

Idea with "expected knowledge" is not that you know all of these things, but rather than you learn those that you don't **on your own**

See e.g. some student comments on the course:

**Comment 1:** "I felt that the first four weeks of the course was slow... similar to all ML courses taught here, they review the same material on fundamentals of data science/machine learning"

**Comment 2:** "Difficult if you have not had any machine learning/data mining experience"
Compared to other UCSD courses:

The most related classes are
• CSE 150/151 ("Introduction to Artificial Intelligence")
• CSE 250A ("Principles of Artificial Intelligence: Probabilistic Reasoning and Decision-Making")
• CSE 250B ("Machine Learning")

None of these courses are prerequisites for each other!
• This course is more “hands-on” — the focus here is on applying techniques from ML to real data and predictive tasks, whereas others are focused on developing a more rigorous understanding of the underlying mathematical concepts
CSE 258 vs. CSE 158

Only one set of lectures this year, but different material. All videos will be online!

(last year’s links)

CSE158:  
https://podcast.ucsd.edu/watch/fa19/cse158_a00/

CSE258:  
https://podcast.ucsd.edu/watch/fa19/cse258_a00/
In Lectures I try to cover:

- The basic material (obviously)
- **Motivation** for the models
- **Derivations** of the models
- Code examples
- Difficult homework problems / problems from past classes/homeworks/exams etc.
Web Mining and Recommender Systems

Course outline
The course webpage is available here:
http://cseweb.ucsd.edu/classes/fa20/cse158-a/
http://cseweb.ucsd.edu/classes/fa20/cse258-a/

This page will include data, code, slides, homework and assignments
This quarter’s content will be (roughly) similar
This course is in two parts:

1. **Methods:**
   - Regression
   - Classification
   - Unsupervised learning and dimensionality reduction

2. **Applications (including some of):**
   - Recommender systems
   - Text mining
   - Crawling data & other useful libraries
   - Social network analysis
   - Online Advertising
   - Mining temporal and sequence data
1: Regression

- Linear regression and least-squares
  - (a little bit of) feature design
  - Overfitting and regularization
    - Gradient descent
  - Training, validation, and testing
    - Model selection
1: Regression

How can we use **features** such as product properties and user demographics to make predictions about **real-valued** outcomes (e.g. star ratings)?

How can we prevent our models from **overfitting** by favouring simpler models over more complex ones?

How can we assess our decision to optimize a particular error measure, like the MSE?
2: Classification

• Logistic regression
• Support Vector Machines
• Multiclass and multilabel classification
• How to evaluate classifiers, especially in “non-standard” settings
2: Classification

Next we adapted these ideas to **binary** or **multiclass** outputs.

What animal is in this image?  Will I **purchase** this product?  Will I **click on** this ad?

Combining features using naïve Bayes models  Logistic regression  Support vector machines
3: Dimensionality Reduction

- Dimensionality reduction
- Principal component analysis
  - Matrix factorization
  - K-means
- Graph clustering and community detection
3: Dimensionality Reduction

Principal component analysis

Community detection
4: Recommender Systems

- Latent factor models and matrix factorization (e.g. to predict star-ratings)
- Collaborative filtering (e.g. predicting and ranking likely purchases)
4: Recommender Systems

Rating distributions and the missing-not-at-random assumption

Latent-factor models

my (user's) "preferences" • HP's (item) "properties"
4: Recommender Systems

- **Preference modeling**: Recommendation of items based on user preferences.
- **Pricing**: Recommendation based on pricing.
- **Retrieval**: Recommendation based on existing items in the database.
5: Text Mining

- Sentiment analysis
- Bag-of-words representations
  - TF-IDF
- Stopwords, stemming, and (maybe) topic models
yeast and minimal red body thick light a Flavor sugar strong quad. grape over is molasses lace the low and caramel fruit Minimal start and toffee. dark plum, dark brown Actually, alcohol Dark oak, nice vanilla, has brown of a with presence. light carbonation. bready from retention. with finish. with and this and plum and head, fruit, low a Excellent raisin aroma

Medium tan

Bags-of-Words

What we would like:

Document topics

Topic models

Sentiment analysis

(review of "The Chronicles of Riddick")
Week 6: Midterm (Nov 9)!

Take-home (probably 24hr)  
(More about grading etc. later)
Later: More "specialized" material

During the second half of the course I mostly cover specific topics in some detail – following are a few examples
Tools and libraries

• Crawling and parsing data from the Web
• Manipulating time and date data
  • Matplotlib
  • Tensorflow
Tools and libraries

Crawling and parsing data from the Web:

In [1]: from urllib import request

In [2]: from urllib.request import urllopen


In [4]: html = str(f.read())

Out[4]:

Note: acts like a file object once opened

Note: url of "The Great Gatsby" reviews

Oh Gatsby, you old sport, you poor semi-delusional hopeful dreamer with some heightened sensitivity to the promises of life, focusing your whole self and soul on that elusive money-colored green light—a dream that shatters just when you are "this" close to it.
Tools and libraries

Matplotlib:

BeerAdvocate, ratings over time

- Sliding window (K=10000)
- Seasonal effects
- Long-term trends
• Power-laws & small-worlds
  • Random graph models
  • Triads and “weak ties”
• Measuring importance and influence of nodes (e.g. pagerank)
Social & Information Networks

- Hubs & authorities
- Power laws
- Small-world phenomena
- Strong & weak ties
Advertising

Matching problems

AdWords

Bandit algorithms
• Sliding windows & autoregression
  • Temporal dynamics in recommender systems
• Temporal dynamics in text & social networks
Temporal & Sequence Data

Topics over time

Social networks over time

Memes over time
There is **no textbook** for this class

- I will give chapter references from *Bishop: Pattern Recognition and Machine Learning*
- I will also give references from Charles Elkan’s notes ([http://cseweb.ucsd.edu/classes/fa18/cse258-a/files/elkan_dm.pdf](http://cseweb.ucsd.edu/classes/fa18/cse258-a/files/elkan_dm.pdf))
Python Data Products (Coursera)
Evaluation

- There will be four homework assignments worth 8% each. Your lowest grade will be dropped, so that 4 homework assignments = 24%
- There will be a (take-home) midterm in week 6, worth 26%
- One assignment on recommender systems (after week 5), worth 25%
- A short open-ended assignment, worth 25%
Evaluation

HW = 24%
Midterm = 26%
Assignment 1 = 25%
Assignment 2 = 25%

Actual goals:
• Understand the basics and get comfortable working with data and tools (HW)
• Comprehend the **foundational** material and the motivation behind different techniques (Midterm)
• Build something that **actually works** (Assignment 1)
• Apply your knowledge creatively (Assignment 2)
Evaluation

• Homework should be delivered by the beginning of the Monday lecture in the week that it’s due
• All submissions will be made electronically (instructions will be in the homework spec, on the class webpage)
Evaluation

Schedule (subject to change but hopefully not):

Week 1: Hw 1 out
Week 3: Hw 1 due, Hw2 out
Week 5: Hw 2 due, Hw3 out, Assign. 1 out
Week 6: midterm
Week 7: Hw 3 due, Hw4 out, Assign. 2 out
Week 8: Assignment 1 due
Week 9: Hw4 due
Week 10: Assignment 2 due
Assignment 1

- Prediction tasks on (e.g.) Amazon clothing data, run as a competition on Kaggle

  - Rating prediction
  - Purchase prediction
  - Helpfulness prediction
Assignment 1

• We’ll do something similar this year, but on Goodreads (probably?) data
Assignment 2

Raw rating data  binned regression  dual regression

"inflection" point

Andrew Prudhomme – “Finding the Optimal Age of Wine”
positive words in wine reviews
negative words in wine reviews

positive words in beer reviews
negative words in wine reviews

Ben Braun & Robert Timpe – “Text-based rating predictions from been and wine reviews”
Assignment 2

Diego Cedillo & Idan Izhaki – “User Score for Restaurants Recommendation System”
\[ \hat{r}_{ui} = \mu + b_u + b_i + (q_i + \frac{1}{|M(i)|} \sum_{n \in M(i)} |s_n|) p_u \]

set of geographic neighbours
impact of neighbours

Long Jin & Xinchi Gu – “Rating Prediction for Google Local Data”
### Topic model from Google Local business reviews

<table>
<thead>
<tr>
<th>“Fitness”</th>
<th>“Italian Restaurants”</th>
<th>“Airport &amp; Rentals”</th>
<th>“Computer Repairs”</th>
<th>“Mexican”</th>
</tr>
</thead>
<tbody>
<tr>
<td>gym</td>
<td>food</td>
<td>san</td>
<td>computer</td>
<td>food</td>
</tr>
<tr>
<td>training</td>
<td>restaurant</td>
<td>francisco</td>
<td>store</td>
<td>mexican</td>
</tr>
<tr>
<td>fitness</td>
<td>wine</td>
<td>car</td>
<td>phone</td>
<td>tacos</td>
</tr>
<tr>
<td>classes</td>
<td>menu</td>
<td>airport</td>
<td>system</td>
<td>burrito</td>
</tr>
<tr>
<td>equipment</td>
<td>great</td>
<td>jose</td>
<td>buy</td>
<td>good</td>
</tr>
<tr>
<td>class</td>
<td>delicious</td>
<td>time</td>
<td>laptop</td>
<td>salsa</td>
</tr>
<tr>
<td>great</td>
<td>dinner</td>
<td>driver</td>
<td>apple</td>
<td>taco</td>
</tr>
<tr>
<td>workout</td>
<td>dishes</td>
<td>service</td>
<td>repair</td>
<td>chips</td>
</tr>
<tr>
<td>weight</td>
<td>excellent</td>
<td>back</td>
<td>problem</td>
<td>burritos</td>
</tr>
<tr>
<td>ve</td>
<td>dining</td>
<td>fixed</td>
<td>fish</td>
<td>fish</td>
</tr>
<tr>
<td>work</td>
<td>meal</td>
<td>pc</td>
<td>chicken</td>
<td>chicken</td>
</tr>
<tr>
<td>body</td>
<td>italian</td>
<td>drive</td>
<td>place</td>
<td>place</td>
</tr>
<tr>
<td>yoga</td>
<td>experience</td>
<td>price</td>
<td>delicious</td>
<td>delicious</td>
</tr>
<tr>
<td>trainers</td>
<td>amazing</td>
<td>data</td>
<td>love</td>
<td>love</td>
</tr>
<tr>
<td>people</td>
<td>wonderful</td>
<td>fix</td>
<td>fresh</td>
<td>fresh</td>
</tr>
<tr>
<td>years</td>
<td>atmosphere</td>
<td>iphone</td>
<td>great</td>
<td>great</td>
</tr>
<tr>
<td>feel</td>
<td>small</td>
<td>screen</td>
<td>beans</td>
<td>beans</td>
</tr>
<tr>
<td>instructors</td>
<td>decor</td>
<td>company</td>
<td>restaurant</td>
<td>restaurant</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>bought</td>
<td>asada</td>
</tr>
</tbody>
</table>
Wikispeedia navigation traces:

Figure 5: Graph of a complete path

<table>
<thead>
<tr>
<th></th>
<th>Average Click</th>
<th>Average Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finish Path</td>
<td>4.72</td>
<td>158.27</td>
</tr>
<tr>
<td>Finished Path Back</td>
<td>6.75</td>
<td>158.31</td>
</tr>
<tr>
<td>Unfinished Path</td>
<td>2.97</td>
<td>835.29</td>
</tr>
<tr>
<td>Unfinished Path Back</td>
<td>5.2</td>
<td>836.00</td>
</tr>
</tbody>
</table>
Power laws!

Images from Chictopia

Wei-Tang Liao & Jong-Chyi Su – “Image Popularity Prediction on Social Networks”
Crime (Chicago)

Goal: to predict the number of incidents of crime on a given day

Joshua Wheeler, Nathan Moreno, Anjali Kanak
Predicting Taxi Tip-Rates in NYC

(data from archive.org)

(pickup and dropoff)

Distance, time taken, speed, and time of day (also on geo)

Sahil Jain, Alvin See, Anish Shandilya
TAs: CSE 158

- Li, Shuyang, shl008@ucsd.edu
- Lin, Ting-Chou, til002@ucsd.edu
- Lin, Xiaokang, xil671@ucsd.edu
- Liu, Zijie, zil022@ucsd.edu
- Mukhopadhyay, Ahan, amukhopa@ucsd.edu
- Nguyen, Viet Vincent, vvn012@ucsd.edu
- Zhang, Yueheng, yuz087@ucsd.edu
- Zhao, Mayu, m5zhao@ucsd.edu

TAs will do most of the grading, and run office hours (in addition to my own)
• He, Zexue, zehe@ucsd.edu
• Majumder, Bodhisattwa, bmajumde@ucsd.edu
• Suh, Dong Jun, dsuh@ucsd.edu
• Wu, Yen-Yi, yew001@ucsd.edu
• Yadav, Abhishek Kumar, akyadav@ucsd.edu

TAs will do most of the grading, and run office hours (in addition to my own)
Office hours

• I will hold office hours on Monday evenings immediately after class (zoom link on Piazza)
• TA office hours will be held throughout the week (exact times and links will be posted to the class webpage & Piazza)
Thanks!

Most announcements will be posted to Piazza

https://piazza.com/ucsd/fall2020/cse158/home
https://piazza.com/ucsd/fall2020/cse258/home

please participate!