CSE 258
Web Mining and Recommender Systems

Introduction
What is CSE 258?

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

**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?
What we need to do data mining

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?

   $$\frac{1}{N} \sum_{u,i}^N \text{ratings}_{u,i} (r_{u,i} - \text{prediction}(u, i))^2$$
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?
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 CSE 258?

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

- **Web datasets**
  - Netflix
  - Amazon
  - Google
  - Reddit
  - Yelp
  - Beeradvocate
  - Epinions
  - Stackoverflow
  - Facebook

- Predictive tasks concerned with human **activities, behavior, and opinions** (i.e., recommender systems)
Expected knowledge

**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.
Expected knowledge

**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)
Expected knowledge

All coding exercises will be done in **Python** with the help of some libraries (numpy, scipy, NLTK etc.)
CSE 258 vs. CSE 250A/B

The two most related classes are
• CSE 250A (“Principles of Artificial Intelligence: Probabilistic Reasoning and Decision-Making”)
• CSE 250B (“Machine Learning”)

None of these courses are prerequisites for each other!
• CSE 258 is more “hands-on” – the focus here is on applying techniques from ML to real data and predictive tasks, whereas 250A/B are focused on developing a more rigorous understanding of the underlying mathematical concepts
CSE 258 vs. CSE 158

Both classes will be podcast in case you want to check out the more advanced material:

(last year’s links)

CSE158: http://podcasts.ucsd.edu/podcasts/default.aspx?PodcastId=3746&v=1

CSE258: http://podcasts.ucsd.edu/podcasts/default.aspx?PodcastId=3747&v=1
In Lectures I try to cover:

• The basic material (obviously)
• Motivation for the models
• Derivations of the models

• Code examples
• Difficult homework problems / exam prep etc.
• Anything else you want to discuss
CSE 258
Web Mining and Recommender Systems

Course outline
The course webpage is available here: http://cseweb.ucsd.edu/classes/fa17/cse258-a/

This page will include data, code, slides, homework and assignments
Course webpage

(winter’s course webpage is here): http://cseweb.ucsd.edu/classes/wi17/cse258-a/

This quarter’s content will be (roughly) similar (though the weighting of assignments/midterms etc. is different)
Course outline

This course is in two parts:

1. **Methods** (weeks 1-4):
   - Regression
   - Classification
   - Unsupervised learning and dimensionality reduction

2. **Applications** (weeks 4-):
   - Recommender systems
   - Text mining
   - Social network analysis
   - Mining temporal and sequence data
   - Something else... visualization/crawling/online advertising etc.
Week 1: Regression

• Linear regression and least-squares
• (a little bit of) feature design
• Overfitting and regularization
  • Gradient descent
• Training, validation, and testing
  • Model selection
Week 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?
Week 2: Classification

• Logistic regression
• Support Vector Machines
• Multiclass and multilabel classification
• How to evaluate classifiers, especially in “non-standard” settings
Week 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
Week 3: Dimensionality Reduction

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

Principal component analysis

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

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

Latent-factor models
Week 5: Guest lecture?

- Probably about deep learning / automatic optimization etc.
  (but TBD!)
Week 6: Midterm (Nov 8)!

(More about grading etc. later)
Week 7: Text Mining

- Sentiment analysis
- Bag-of-words representations
- TF-IDF
- Stopwords, stemming, and (maybe) topic models
Week 7: Text Mining

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:

87 of 102 people found the following review helpful

**Review:**

You know what you like. I don't. (2001)

**Author:**

By *militation* [Review Date: January 27, 2001]

The reviewer...

Point of view: I don't know if it is on my level of interest. I was pleased with the... and the movie.

**Summary:**

While changing the... to... and a little too much... It did not change the attempt or... aide. It left me... to just a little... and the movie. They did not change the attempt or... aide. It left me... to just a little... and the movie.

**Ratings:**

1 star out of 5 stars. I really enjoyed the movie. It did not change the attempt or... aide. It left me... to just a little... and the movie.

**Actions:**

Action: action, fast, explosion,

**Sci-fi:**

space, future, planet,

**Topic models**

Sentiment analysis

(review of "The Chronicles of Riddick")
Week 8: Social & Information Networks

• Power-laws & small-worlds
• Random graph models
• Triads and “weak ties”
• Measuring importance and influence of nodes (e.g. pagerank)
Week 8: Social & Information Networks

Hubs & authorities

Small-world phenomena

Power laws

Strong & weak ties
Week 9: Advertising

Matching problems

users

.92
.75
.67
.24
.97
.59

ads

Bandit algorithms

AdWords
Week 10: Temporal & Sequence Data

• Sliding windows & autoregression
• Hidden Markov Models
• Temporal dynamics in recommender systems
• Temporal dynamics in text & social networks
Week 10: Temporal & Sequence Data

- Memes over time
- Social networks 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/fa17/cse258-a/files/elkan_dm.pdf](http://cseweb.ucsd.edu/classes/fa17/cse258-a/files/elkan_dm.pdf))
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 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)
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
Previous assignments...
Assignment 1

• Prediction tasks on Amazon clothing data, run as a competition on Kaggle

Rating prediction

Purchase prediction

Helpfulness prediction
We’ll do something similar this year, but on Google Local data.
Assignment 2

Raw rating data  binned regression  dual regression

Andrew Prudhomme – “Finding the Optimal Age of Wine”

“inflection” point
Assignment 2

ratings vs. time

ratings vs. review length

Ruogu Liu – “Wine Recommendation for CellarTracker”
Assignment 2

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

cellartracker:
- positive words in wine reviews
- negative words in wine reviews

RateBeer:
- positive words in beer reviews
- negative words in wine reviews
User age

- Rating vs. age
- Aroma vs. age
- Day of week vs. age
- Year vs. age
- Hour of day vs. age
- Category vs. age

Joseph Luttrell, Spenser Cornett
Assignment 2

Diego Cedillo & Idan Izhaki – “User Score for Restaurants Recommendation System”

Figure 3: Restaurant Ratings

Figure 6: Average rating per location

ratings per location

k-means of ratings per location

3.52

4.00
Assignment 2

Long Jin & Xinchi Gu – “Rating Prediction for Google Local Data”

\[ \hat{r}_{ui} = \mu + b_u + b_i + \left( q_i + \frac{1}{|M(i)|} \right) \sum_{n \in M(i)} |s_n| p_u \]

set of geographic neighbours  impact of neighbours
Topic model from Google Local business reviews

<table>
<thead>
<tr>
<th>Topic</th>
<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 francisco</td>
<td>computer</td>
<td>food</td>
<td></td>
</tr>
<tr>
<td>training</td>
<td>restaurant</td>
<td>car</td>
<td>store</td>
<td>mexican</td>
<td></td>
</tr>
<tr>
<td>fitness</td>
<td>wine</td>
<td>airport</td>
<td>phone</td>
<td>tacos</td>
<td></td>
</tr>
<tr>
<td>classes</td>
<td>menu</td>
<td>jose</td>
<td>buy</td>
<td>burrito</td>
<td></td>
</tr>
<tr>
<td>equipment</td>
<td>great</td>
<td>delicious</td>
<td>laptop</td>
<td>good</td>
<td></td>
</tr>
<tr>
<td>class</td>
<td>service</td>
<td>time</td>
<td>apple</td>
<td>salsa</td>
<td></td>
</tr>
<tr>
<td>life</td>
<td>dinner</td>
<td>rental</td>
<td>repair</td>
<td>taco</td>
<td></td>
</tr>
<tr>
<td>great</td>
<td>dishes</td>
<td>driver</td>
<td>problem</td>
<td>chips</td>
<td></td>
</tr>
<tr>
<td>workout</td>
<td>excellent</td>
<td>service</td>
<td>back</td>
<td>burritos</td>
<td></td>
</tr>
<tr>
<td>weight</td>
<td>dining</td>
<td>shuttle</td>
<td>fixed</td>
<td>fish</td>
<td></td>
</tr>
<tr>
<td>ve</td>
<td>meal</td>
<td>taxi</td>
<td>pc</td>
<td>chicken</td>
<td></td>
</tr>
<tr>
<td>work</td>
<td>italian</td>
<td>trip</td>
<td>drive</td>
<td>place</td>
<td></td>
</tr>
<tr>
<td>body</td>
<td>experience</td>
<td>city</td>
<td>price</td>
<td>delicious</td>
<td></td>
</tr>
<tr>
<td>yoga</td>
<td>amazing</td>
<td>cab</td>
<td>data</td>
<td>love</td>
<td></td>
</tr>
<tr>
<td>trainers</td>
<td>wonderful</td>
<td>lax</td>
<td>fix</td>
<td>fresh</td>
<td></td>
</tr>
<tr>
<td>people</td>
<td>atmosphere</td>
<td>area</td>
<td>iphone</td>
<td>great</td>
<td></td>
</tr>
<tr>
<td>years</td>
<td>small</td>
<td>experience</td>
<td>screen</td>
<td>beans</td>
<td></td>
</tr>
<tr>
<td>feel</td>
<td>decor</td>
<td>company</td>
<td>bought</td>
<td>restaurant</td>
<td></td>
</tr>
<tr>
<td>instructors</td>
<td></td>
<td></td>
<td></td>
<td>asada</td>
<td></td>
</tr>
</tbody>
</table>

Mohit Kothari & Sandy Wiraatmadja – “Reviews and Neighbors Influence on Performance of Business”
Wikispeedia navigation traces:

Figure 5: Graph of a complete path

<table>
<thead>
<tr>
<th>Path Type</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>

Shelby Thomas & Moein Khazraee – “Determining Topics in Link Traversals through Graph-Based Association Modeling”
Images from Chictopia

Power laws!

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

Over 15 years

Hour of the day

Over 7 years

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)

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

Sahil Jain, Alvin See, Anish Shandilya
TAs

- Karamchandani, Digvijay
- Kolasani, Sai Chaitanya
  - Misra, Rishabh
- Narayanan, Srinath
  - Pasricha, Rajiv
- Sharma, Saksham
- Yogendra Murali, Nikhil
  - Zhang, Hongyi

TAs will do most of the grading, and run office hours (in addition to my own)
• I will hold office hours on Tuesday mornings (9:00am-1:00pm, CSE 4102)
• TA office hours will be held on Mondays and Fridays from 10:00am-13:00pm in B250A (Monday) and B250B (Friday) (see course webpage for schedule)
Most announcements will be posted to Piazza

https://piazza.com/ucsd/fall2017/cse258/home

please participate!