CSE 190
Data Mining and Predictive Analytics

Introduction
What is CSE 190?

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?
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?

   \[
   \frac{1}{N} \sum_{i}^{N} \sum_{u}^{r_{u,i}} (r_{u,i} - \text{prediction}(u, i))^2
   \]
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?
What is CSE 190?

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
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.
Basic mathematics

- Some linear algebra \[ Ax = y \Rightarrow x = (A^T A)^{-1} A^T y \]
- Some optimization \[ \frac{d}{dx}(Ax - y)^2 \]
- 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 190 vs. CSE 150/151

The two most related classes are
• CSE 150 (“Introduction to Artificial Intelligence: Search and Reasoning”)
• CSE 151 (“Introduction to Artificial Intelligence: Statistical Approaches”)

None of these courses are prerequisites for each other!
• CSE 190 is more “hands-on” – the focus here is on applying techniques from ML to real data and predictive tasks, whereas 150/151 are focused on developing a more rigorous understanding of the underlying mathematical concepts
CSE 255 is the **graduate** version of this class. It is roughly the same, though there are some differences:

- CSE 255 will have more on graphical models (we’ll cover it a little bit in 190, but not much)
- CSE 255 will have a little bit more on optimization (e.g. gradient based methods). We’ll cover these too, but not really with complex derivations – in this class some of the more complex linear algebra / calculus will be treated in more of a “black box” way
- CSE 255 will cover more academic papers

- As long as you do the CSE 190 assessments, you’re welcome to attend either class (but not this week!)
Both classes will be podcast in case you want to check out the more advanced material:

CSE190:  
http://podcasts.ucsd.edu/podcasts/default.aspx?PodcastId=3004&v=1

CSE255:  
http://podcasts.ucsd.edu/podcasts/default.aspx?PodcastId=3003&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 190
Data Mining and Predictive Analytics
Course outline
The course webpage is available here: http://cseweb.ucsd.edu/classes/fa15/cse190-a/

This page will include data, code, slides, homework and assignments
(last quarter’s course webpage is here): http://cseweb.ucsd.edu/~jmcauley/cse190/

This quarter’s content will be (roughly) similar (though the weighting of assignments/midterms etc. is different)
This course is in two parts:

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

2. **Applications** (weeks 4/5-):
   - Recommender systems
   - Text mining
   - Social network analysis
   - Mining temporal and sequence data
   - Something else if there’s time: (there probably won’t be) 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
Week 4: Graphical Models (maybe!)

- Dealing with interdependent variables
- Labeling problems on graphs
- Hidden Markov Models and sequential data
Week 4: Graphical Models

Directed and undirected models

Inference via graph cuts

$p(a)p(b)p(c|a, b)p(d|c)$

$\psi(a, b, c)\psi(c, d)$
Week 4: Graphical Models

Maybe not though...

- Not many people used material from this lecture in their assignments, so I want to keep it to a minimum
- I plan to cover only the simplest cases, and possibly return to this material at the end of the quarter

\[ p(Sun=-6 \mid Sat=-7)p(Mon=-8 \mid Sun=-6)p(Tue=-6 \mid Mon=-8) \ldots \]
Week 5: Recommender Systems

• Latent factor models and matrix factorization (e.g. to predict star-ratings)
• Collaborative filtering (e.g. predicting and ranking likely purchases)
Week 5: Recommender Systems

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

Latent-factor models
Week 6: Midterm (Nov 2)!

(More about grading etc. later)
Week 7/8: Text Mining

- Sentiment analysis
- Bag-of-words representations
  - TF-IDF
- Stopwords, stemming, and (maybe) topic models
Week 7/8: 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 152 people found the following review helpful

By *Quickly* "Jitlog" (Washington State) - see all my reviews

Not my style. The Chronicles of Middle Earth are intended to be enjoyed and savored, not devoured. Many fans and friends of mine and me, and my family, I have to say, have had a hard time getting into this book. That being said, I have enjoyed it, and I plan to continue with it. However, after reading the first few chapters, I was a bit surprised to find that the story was not as engaging as I had expected. The pace is a bit slow, and the writing style is not as polished as I would have liked. Overall, I would recommend this book to fans of the series, but others may want to consider looking elsewhere for their next read.

Point off, let me say that when playing the trial, the first selection to come up is Connors or Grit. And Earl's boys. The choice is as final, so I'll leave it at that. The more I play, the more I understand the importance of this world. The deeper the meaning behind the choices, the more I am left with a sense of pride. The beauty of the world, the characters, the story, all come together to create a truly unforgettable experience. I cannot wait to see what the future holds for the Chronicles of Middle Earth.

(review of "The Chronicles of Riddick")

Topic models

Sentiment analysis
Week 9: Social & Information Networks

• Power-laws & small-worlds
  • Random graph models
  • Triads and “weak ties”
• Measuring importance and influence of nodes (e.g. pagerank)
Week 9: Social & Information Networks

Hubs & authorities

Small-world phenomena

Power laws

Strong & weak ties
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

Topics over time

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/~jmcauley/cse190/files/elkan_dm.pdf](http://cseweb.ucsd.edu/~jmcauley/cse190/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 25%
- One assignment on recommender systems (after week 5), worth 25%
- A short open-ended assignment, worth 25%
- We’ll find that extra 1% somewhere
Evaluation

HW = 24%
Midterm = 25%
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)
• Homework should be handed in at the beginning of the Monday lecture in the week that it’s due
• If you can’t attend the lecture drop off homework outside my office (CSE 4102) before the lecture
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...
• Prediction tasks on Amazon electronics data, run as a competition on Kaggle

Rating prediction

Purchase prediction

Helpfulness prediction
Assignment 1

- We’ll definitely do this again, but with different data and possibly different tasks

Rating prediction

Purchase prediction

Helpfulness prediction
Assignment 2

- Raw rating data
- Binned regression
- Dual regression

“Inflection” point

Andrew Prudhomme – “Finding the Optimal Age of Wine”
Assignment 2

ratings vs. time

ratings vs. review length

Ruogu Liu – “Wine Recommendation for CellarTracker”
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

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

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

set of geographic neighbours
impact of neighbours

Long Jin & Xinchi Gu – “Rating Prediction for Google Local Data”
Assignment 2

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

Topic model from Google Local business reviews
 Wikispeedia navigation traces:

Figure 5: Graph of a complete path

<table>
<thead>
<tr>
<th>Path Description</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>
Images from Chictopia

Power laws!

Assignments 2

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
• Long Jin is the TA for this course
• Homework will be mostly marked by tutors (Archit Khosla, Mingshan Wang)
Office hours

• I will hold office hours on Tuesday mornings (9:30am-11:30am, CSE 4102)
• CSE 255 office hours will be held afterwards (11:30-1:30)
• TA office hours t.b.d. later
Questions?