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?
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 Data Mining is NOT

Data mining is (hopefully) not

• Abusing and misusing private information, e.g. tracking people’s visits to a store by scanning the wifi signals from their phones
• Finding hypotheses from data
• Mistaking “random” occurrences as meaningful patterns

“Big Data Gone Wrong“:
• The Dangers and Blues of Data Mining (http://goo.gl/OiVZez)
• Ethics of Big Data: Balancing Risk and Innovation (http://www.amazon.com/dp/1449311792)
• Nordstrom tracking incident (http://goo.gl/uSnyMx)
• Lucia de Berk case (http://en.wikipedia.org/wiki/Lucia_de_Berk)
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?

$$\frac{1}{N} \sum_{ratings}^{N} (r_{u,i} - prediction(u,i))^2$$
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 CSE 255 is

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
**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
  \[ \frac{d}{dx} (Ax - y)^2 \]
- 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
The course webpage is available here: http://cseweb.ucsd.edu/~jmcauley/cse190/

This page will include data, code, slides, homework and assignments
Last quarter’s course webpage is here:
http://cseweb.ucsd.edu/~jmcauley/cse255/

190’s content will be (roughly) similar
Course outline

This course is in two parts:

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

2. **Applications** (weeks 5-):
   - Recommender systems
   - Visualization
   - Online advertising
   - Text mining
   - Social network analysis
   - Mining temporal and sequence data
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
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

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

Directed and undirected models

\[ p(a)p(b)p(c|a, b)p(d|c) \quad \psi(a, b, c)\psi(c, d) \]

Inference via graph cuts
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

my (user’s) "preferences"

HP’s (item) "properties"
Week 6: Midterm (May 4)!

(More about grading etc. later)

- & Data visualization

BeerAdvocate, ratings over time

Scatterplot

Sliding window (K=10000)

long-term trends

seasonal effects
Week 6: Midterm (May 4)!

(More about grading etc. later)

- & Data visualization
Time-series regression

Also useful to plot data:

BeerAdvocate, ratings over time

Scatterplot

Sliding window (K=10000)

long-term trends

seasonal effects

BeerAdvocate, ratings over time

timestamp

timestamp

rating

rating

Code on:

http://jmcauley.ucsd.edu/cse255/code/lecture8.py
Week 7: Guest lecture?

• & Models for Online Advertising
• Sentiment analysis
• Bag-of-words representations
  • TF-IDF
• Stopwords, stemming, and (maybe) topic models
Week 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 102 people found the following review helpful

Great film. You know what you kill

By nickticky “Mattie” (Washington state) - Dec 3, 2002

The movie is a classic. You can see it and then, while changing every things, left behind "back story" to be just Riddick. They did not change the attitude of films two up or bring out of the original character, which was very pleasing to "Rich Blais" fans. Highly recommended.

Point out, I hate that when playing the fight, the first selection to come up is Convict or fighting, and no explanation of the characters. This confused me at first, so I cut another, all of the fact that they are very different movie. From each other. Six different characters, eight different background movies. Select either one and continue with the movie.

(review of "The Chronicles of Riddick")

Topic models

Sentiment analysis
• 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
- Power laws
- Small-world phenomena
- 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

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 [here](http://cseweb.ucsd.edu/~jmcauley/cse190/files/elkan_dm.pdf)
Evaluation

• There will be **four** homework assignments worth 10% each. Your **lowest grade** will be dropped, so that 4 homework assignments = 30%
• There will be a midterm in week 6, worth 30%
• One assignment on recommender systems (after week 5), worth 20%
• A short open-ended assignment, worth 20%
• Homework should be handed in at the beginning of the Tuesday 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
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
Assignments from last quarter...
Assignment 1

- Three prediction tasks on Amazon electronics data, run as a competition on Kaggle

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”
Assignment 2

cellartracker:

positive words in wine reviews
negative words in wine reviews

RateBeer:

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”
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

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>life</td>
<td>service</td>
<td>rental</td>
<td>apple</td>
<td>taco</td>
</tr>
<tr>
<td>great</td>
<td>dinner</td>
<td>driver</td>
<td>repair</td>
<td>chips</td>
</tr>
<tr>
<td>workout</td>
<td>dishes</td>
<td>service</td>
<td>problem</td>
<td>burritos</td>
</tr>
<tr>
<td>weight</td>
<td>excellent</td>
<td>bus</td>
<td>back</td>
<td>fish</td>
</tr>
<tr>
<td>ve</td>
<td>dining</td>
<td>shuttle</td>
<td>fixed</td>
<td>chicken</td>
</tr>
<tr>
<td>work</td>
<td>meal</td>
<td>taxi</td>
<td>pc</td>
<td>place</td>
</tr>
<tr>
<td>body</td>
<td>italian</td>
<td>trip</td>
<td>drive</td>
<td>delicious</td>
</tr>
<tr>
<td>yoga</td>
<td>experience</td>
<td>city</td>
<td>price</td>
<td>love</td>
</tr>
<tr>
<td>trainers</td>
<td>amazing</td>
<td>cab</td>
<td>data</td>
<td>fresh</td>
</tr>
<tr>
<td>people</td>
<td>wonderful</td>
<td>lax</td>
<td>fix</td>
<td>great</td>
</tr>
<tr>
<td>years</td>
<td>atmosphere</td>
<td>area</td>
<td>iphone</td>
<td>beans</td>
</tr>
<tr>
<td>feel</td>
<td>small</td>
<td>experience</td>
<td>screen</td>
<td>restaurant</td>
</tr>
<tr>
<td>instructors</td>
<td>decor</td>
<td>company</td>
<td>bought</td>
<td>asada</td>
</tr>
</tbody>
</table>
Assignment 2

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>
Images from Chictopia

Power laws!

Wei-Tang Liao & Jong-Chyi Su – “Image Popularity Prediction on Social Networks”
TAs

• Pranay Kumar Myana
  (pmyana@eng.ucsd.edu)

• Long Jin
  (longjin@eng.ucsd.edu)
Office hours

- I will hold office hours on Wednesday afternoon (1pm-3pm, CSE 4102)
- Long, Pranay to add office hours t.b.d. later
Questions?