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.
**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

**Queries:**

**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?

103 of 115 people found the following review helpful

[Review]
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

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

- **Web** datasets
  - [ratebeer](https://ratebeer.com), [reddit](https://www.reddit.com), [Epinions.com](https://www.epinions.com), [Netflix](https://www.netflix.com), [yelp](https://www.yelp.com), [Beeradvocate](https://www.beeradvocate.com), [Amazon](https://www.amazon.com), [Google+](https://plus.google.com), [Facebook](https://www.facebook.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.
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.)
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"
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
This course in 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
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
4: Recommender Systems

- Preference modeling
- Pricing
- Retrieval

User $\times$ Dress $\rightarrow$ Heart

User $\times$ Dress $\rightarrow$ Dollar

User $\rightarrow$ Existing items
• 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 model)

(review of “The Chronicles of Riddick”)

Sentiment analysis

Topic models
Tools and libraries

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

```python
In [2]: from urllib.request import urlopen


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

**Note:** acts like a file object once opened

**Note:** url of "The Great Gatsby" reviews
Tools and libraries

Matplotlib:

BeerAdvocate, ratings over time

Sliding window (K=10000)

long-term trends

seasonal effects
• 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

Bandit algorithms

AdWords

users
ads

.92
.75
.67
.24
.97
59
• 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
Python Data Products (Coursera)