

CSE 255

Data Mining and Predictive Analytics

Introduction

What is CSE 255?

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?

People You May Know



Jure Leskovec

🏠 Professor at Stanford University
9 mutual friends



Stéphane Ross

🏠 Software Engineer at Google self-driving car
3 mutual friends



Jim Tink

8 mutual friends



Cristian Danescu

🏠 Stanford
2 mutual friends

Examples – Advertising

Prediction: will I click on an advertisement?

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

query → engagement rings

Web Shopping Images Maps News More Search tools









About 39,300,000 results (0.33 seconds)

Tiffany & Co.® Engagement - Tiffany.com
Ad www.tiffany.com/Engagement
Your Perfect Match Deserves The Best. Consult An Expert At Tiffany®
Tiffany & Co. has 122,932 followers on Google+
Rings Diamond Consultation Speak to an Expert A Tiffany Diamond

Unique Engagement Rings - HaroldStevens.com
Ad haroldstevens.com/Engagement-Rings
Shop Harold Stevens - 2014 LUXURY Retailer of the Year Finalist
Wedding Bands - Antique & Vintage Rings - Beverley K - Single Stone Rings
525 B Street, Suite 150, San Diego, CA (619) 231-0520

Engagement Rings - ritani.com
Ad www.ritani.com/Engagement-Rings (888) 978-0802
4.8 ★★★★★ rating for ritani.com
1000s of unique engagement ring designs. Free FedEx & easy returns.
Free In-Store Previews · Handcrafted in New York · Lifetime Care Package
Create Your Own Ring - The Ritani® Difference - Ritani® Diamonds

Shop for engagement rings on Google Sponsored

 French-Set Halo Diamond... \$1,990.00 Ritani	 18K White Gold Delicate... \$950.00 Brilliant Earth ★★★★★ (57)	 18K White Gold Fancy D... \$1,825.00 Brilliant Earth ★★★★★ (13)	 Chamise Diamond Eng... \$975.00 Brilliant Earth ★★★★★ (7)
 Vintage Cushion Halo... \$4,140.00	 Princess Cut Diamond Eng... \$1,906.82	 18K White Gold Hudson... \$975.00	 18K White Gold Harmon... \$1,675.00

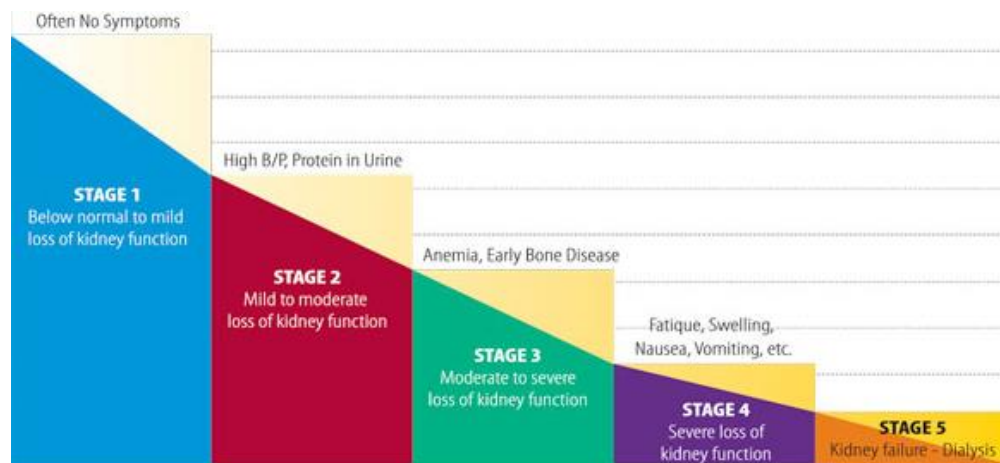
ads →

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?

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

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

★★★★★ **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?

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

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 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

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 $\frac{d}{dx} (Ax - y)^2$
- 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.)