Outline

• Disclaimer
  – My qualifications to give and not give this talk

• Airline Revenue Management
  – Pricing and Seat Inventories in passenger air travel

• Machine Learning and Statistical Learning Theory
  – Data mining for classification and other tasks

• Recommender Systems for Internet Commerce
  – Using data mining to improve customer “productivity”
Pace of Innovation is Accelerating
Computing Power grows Exponentially
Growth of Machine-Generated Data

Machine-generated data is changing the game

Machine-generated data

Authored Data

Machine-generated versus authored data

- Storage online
- All medical imaging
- Medical data stored
- Surveillance bytes
- Surveillance for urban areas
- Personal multimedia
- In databases
- Static Web data
- Text data

Year

Giga-bytes/USCapita/year

“I think there is a world market for maybe five computers”

– Thomas Watson, Chairman of IBM, 1943
“Computers in the future may weigh no more than 1.5 tons”

–Popular Mechanics (magazine), 1949
“There is no reason anyone would want a computer in their home”

–Ken Olson, founder of DEC, 1977
“640K ought to be enough for anybody”

–Bill Gates, 1981
“Prediction is difficult, especially about the future”

–Yogi Berra (famous New York baseball player)
Some Challenges for 21st Century Operations Research

- How to deal with massive amounts of noisy data?
- How to deal with risk and uncertainty?
- How to solve operational real-time problems?
  - Internet OR problems are almost always real-time
Outline

• Disclaimer
  – My qualifications to give and not give this talk

• Airline Revenue Management
  – Pricing and Seat Inventories in passenger air travel

• Machine Learning and Statistical Learning Theory
  – Data mining for classification and other tasks

• Recommender Systems for Internet Commerce
  – Using data mining to improve customer “productivity”
Disclaimer

Not qualified to give this talk:

• My research is “narrow” and mostly “theoretical”
• I don’t participate in the full breadth of OR activities
• I do some applied work, but not much
• I do some consulting, but not much
• I’m too old
My qualifications to give this talk:

- I constantly interact with all sorts of OR people at MIT and elsewhere all the time
- After my PhD, I worked in consulting, often not using OR, and also often using OR
- I have always loved OR applications and theory
- I developed successful OR curriculum MIT Sloan MBA
- former Deputy Dean at MIT Sloan
Challenges for 21st Century Operations Research

• What challenges do you think OR faces?

• How should the profession address these challenges?

• How should the professionals (us) face these challenges?
Outline

• Disclaimer
  – My qualifications to give and not give this talk

• Airline Revenue Management
  – Pricing and Seat Inventories in passenger air travel

• Machine Learning and Statistical Learning Theory
  – Data mining for classification and other tasks

• Recommender Systems for Internet Commerce
  – Using data mining to improve customer “productivity”
Airline Revenue Management

One idea, many related names:

- “Pricing and Seat Inventory Management in Passenger Air Travel”
- Revenue Management
- Yield Management
- Demand Chain Management
- Demand Management
What is Revenue Management?

- “Selling the right seat to the right passenger at the right price and at the right time”
  - Rob Freund

- “Control and management of reservations inventory in a way that increases company profitability”
  - Barry Smith, AA
Special Features of the Economics of Airline Management

- Huge sunk and fixed costs
  - Airplanes, gate facilities, fuel costs
- Very low variable costs per passenger
- Near-perfect competition on many routes
  - Near-perfect information, zero cost of information, symmetric information
- No inventories of “product”
  - An empty seat is a lost revenue opportunity forever
  - Highly perishable inventory
Pricing Example

• Flight has 200 seats
  – Y-fare is $P_Y = $400, demand is $D_Y = 100$
  – Q-fare is $P_Q = $100, demand is $D_Q = 100$

• Pricing Options:
  – Set $P = $100, revenue is $20,000 = $100 \times 200$
  – Set $P = $400, revenue is $40,000 = $400 \times 100$
  – If we can figure out how to charge Q-fare customers $100 and Y-fare customers $400, revenue is $50,000 = $100 \times 100 + $400 \times 100$
Pricing Example, continued
Pricing and Product Differentiation

• In order to do differential pricing, one needs to have differential products
• Make “flexibility” part of the product:
  – Expensive ticket (Y-fare):
    • no restrictions on weekend stay
    • no restrictions on advanced purchase
    • no restrictions on changing ticket
  – Inexpensive ticket (Q-fare):
    • Requires a Saturday night stay-over
    • 14 day advanced purchase
    • restrictions/penalties for changing ticket after purchase
Problem: once you create inexpensive Q-fare prices, you must be careful not to sell off the entire seat inventory to low-fare demand.

You must protect some number of seats for late-arriving high-fare customers.

How many seats to protect for late-arriving high-fare customers?
How to Determine Seat Protection Level

Let $L$ denote the number of seats protected for high-fare demand.
Determining the Seat Protection Level, continued

Simple but careful analysis yields that the optimal seat protection level $L$ must satisfy:

$$\Pr(D_Y \leq L) = \frac{P_Y - P_Q}{P_Q}$$

**Example:**

If $D_Y \sim \text{Normal}(\mu = 28, \sigma = 5.5)$, then $L = 30$. 
In Practice Problem is More Complex

In practice many flights share the same aircraft
In practice each flight has multiple fare classes

Seattle

SF

NY

Reno

BOS

Albany
Historical Timeline of Revenue Management

- 1978: American Airlines develops SABRE reservations system for global airline industry and introduces differential fares
- 1980s: first Revenue Management (RM) software used at AA, developed by Bob Phillips at DFI, Inc.
- 1992: virtually every airline has implemented revenue management
- Late 1990s: RM starts to permeate other industries
Other Applications of Revenue Management

- Hotel industry
- Cruise lines
- Car rental industry
- Freight transportation (rail, ship, truck)
- Broadcasting, etc.

All of these industries have:
  - Advance reservations systems
  - Range of prices/services
  - Perishable inventory
  - Heterogeneous demand for flexible service
Impact at AA

Robert L. Crandall, then-Chairman, President, and CEO of AMR and American Airlines (1992):

“I believe that yield management is the single most important technical development in transportation management since we entered the era of airline deregulation in 1979…. The development of yield management was a key to American Airline's survival in the post-deregulation environment.”

“The development of the American Airline's yield-management system has been long and sometimes difficult, but this investment has paid off. We estimate that yield management has generated $1.4 billion in incremental revenue in the last three years alone. This is not a one-time benefit.”
Is Revenue Management Operation Research?

• RM uses optimization, statistics, simulation, computational science, stochastic systems, etc.

• The models are used on-line and in real-time
  – No “decision-maker” overseeing the outcomes
  – Perhaps these were the first suite of OR models to be used this way
My Inspiration
My Inspiration

“The final test of any theory is its capacity to solve the problems which originated it.”
“The final test of any theory is its capacity to solve the problems which originated it.”

-George Dantzig, opening sentence of 1963 book *Linear Programming and Extensions*
Outline

• Disclaimer
  – My qualifications to give and not give this talk

• Airline Revenue Management
  – Pricing and Seat Inventories in passenger air travel

• Machine Learning and Statistical Learning Theory
  – Data mining for classification and other tasks

• Recommender Systems for Internet Commerce
  – Using data mining to improve customer “productivity”
Statistical Learning: Basics

• We are given m points \( x^1, \ldots, x^m \) in \( \mathbb{R}^n \)

• each \( x^i \) has a label of \( y_i = +1 \) or -1

• each data record \([ (x^i)_1, \ldots, (x^i)_n ; y_i ]\) is an example

• the example points \( x^i \) with label \( y_i = +1 \) have property P

• the example points \( x^i \) with label \( y_i = -1 \) do not have property P

• We would like to use these m examples to develop a linear rule that can be used to predict whether or not other points \( x \) might or might not have property P
Illustration of Supervised Learning
Supervised Learning Linear Decision Rule
Supervised Learning, continued

• We seek a vector \( w \) in \( \mathbb{R}^n \) for which:
  - \( (x_i)^T w > 0 \) for all \( x_i \) for which \( y_i = +1 \)
  - \( (x_i)^T w < 0 \) for all \( x_i \) for which \( y_i = -1 \)

• We will then use \( w \) to predict whether or not other points \( v \) have property \( P \) or not, using the rule:
  - If \( v^T w > 0 \), then we declare that \( v \) has property \( P \)
  - If \( v^T w < 0 \), then we declare that \( v \) does not have property \( P \)
Vast Applications

• From medical image analysis to anti-terrorist security protocols to spam filters to credit-card fraud to ....

• Business intelligence
• Fraud detection
• Tax audit and compliance
• Personalized medicine
• Etc.
Data often not Linearly Separable
Mathematical Model

• Create a matrix $A$ whose $i^{th}$ row is:

$$A_i = y_i x^i$$

• We seek $w$ for which:

$$Aw > 0,$$

or more realistically something like:

$$Aw \geq 0$$
Mathematical Model, cont.

• We seek \( w \) for which:
  \[ Aw > 0 , \]

  or more realistically:
  \[ Aw \geq 0 \]

Who here has worked or is working on this problem? Sebastian Maldonado, others?
Is Machine Learning Operation Research?

• Is this Operations Research? Why or why not?

• If not, is this statistics? Or engineering? Or computer science?

• I like to call this “information engineering”

• But to me, Information Engineering is part of Operations Research
Outline

• Disclaimer
  – My qualifications to give and not give this talk

• Airline Revenue Management
  – Pricing and Seat Inventories in passenger air travel

• Machine Learning and Statistical Learning Theory
  – Data mining for classification and other tasks

• Recommender Systems for Internet Commerce
  – Using data mining to improve customer “productivity”
Recommender Systems

Data on customer ratings of movies:

|       | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
| Juan  | 5 | 1 | 3 | 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Isabel|   |   |   |   | 4 | 4 | 1 | 2 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Carlos| 3 | 2 | 2 | 5 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Pablo |   | 1 | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Raul  |   |   |   |   | 2 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Jose  | 5 | 3 |   |   |   | 2 | 2 |   | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Carolina|   |   | 4 | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Sergio| 3 | 2 | 2 | 2 | 3 | 3 |   | 5 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

45
An Ordinary Fact about Matrices

Suppose M is an m by n matrix of rank 17

Then there are 17 n-vectors

\[ g_1, g_2, g_3, \ldots, g_{16}, g_{17} \]

for which every row of M is some linear combination of these n-vectors.
Recommender Systems

Data on customer ratings of movies:

|      | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
| Juan | 5 | 1 | 3 | 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Isabel|   |   |   |   | 4 | 4 | 1 | 2 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Carlos|   |   |   |   | 3 | 2 | 2 | 5 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Pablo |   |   |   |   |   | 1 | 1 |   |   | 4 | 5 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Raul  |   |   |   |   |   |   | 2 |   |   | 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Jose  |   |   |   |   | 5 | 3 |   |   | 2 | 2 |   |   |   |   |   | 1 |   |   |   |   |   |   |   |   |   |   |   |
| Carolina |   |   |   | 4 |   |   | 1 |   |   |   |   | 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Sergio| 3 | 2 | 2 | 2 | 3 |   |   | 5 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

47
Recommender Systems, cont.

How can we use this data to recommend a new movie for Carlos to watch?

<table>
<thead>
<tr>
<th></th>
<th>Movie Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 A B C D E F G H I J K L M N O P Q</td>
</tr>
<tr>
<td>Juan</td>
<td>5 1 3 4</td>
</tr>
<tr>
<td>Isabel</td>
<td>4 4 1 2</td>
</tr>
<tr>
<td>Carlos</td>
<td>3 2 2 5</td>
</tr>
<tr>
<td>Pablo</td>
<td>1 1</td>
</tr>
<tr>
<td>Raul</td>
<td>2</td>
</tr>
<tr>
<td>Jose</td>
<td>5 3 2 2 1</td>
</tr>
<tr>
<td>Carolina</td>
<td>4 1</td>
</tr>
<tr>
<td>Sergio</td>
<td>3 2 2 3 5</td>
</tr>
</tbody>
</table>
Matrix Completion

We would like to fill in all of the missing entries in the data rating matrix $D$. There are currently 156 missing entries in the matrix $D$.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Juan</td>
<td></td>
<td>1</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isabel</td>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carlos</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pablo</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raul</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Jose</td>
<td>5</td>
<td>3</td>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carolina</td>
<td></td>
<td></td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sergio</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Juan: Juan
Isabel: Isabel
Carlos: Carlos
Pablo: Pablo
Raul: Raul
Jose: Jose
Carolina: Carolina
Sergio: Sergio

---

Matrix Completion

We would like to fill in all of the missing entries in the data rating matrix $D$. There are currently 156 missing entries in the matrix $D$.  

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Juan</td>
<td></td>
<td>1</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isabel</td>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carlos</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pablo</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raul</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Jose</td>
<td>5</td>
<td>3</td>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carolina</td>
<td></td>
<td></td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sergio</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Matrix Completion

We would like to fill in all of the missing entries in the data rating matrix $D$. There are currently 156 missing entries in the matrix $D$.  

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Juan</td>
<td></td>
<td>1</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isabel</td>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carlos</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pablo</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raul</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Jose</td>
<td>5</td>
<td>3</td>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carolina</td>
<td></td>
<td></td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sergio</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Matrix Completion

We would like to fill in all of the missing entries in the data rating matrix $D$. There are currently 156 missing entries in the matrix $D$.  

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Juan</td>
<td></td>
<td>1</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isabel</td>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carlos</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pablo</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raul</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Jose</td>
<td>5</td>
<td>3</td>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carolina</td>
<td></td>
<td></td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sergio</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---
Data on Movie Ratings

• We don’t have enough information

or

• We do have enough information
Assume a “Group” Structure Model

• Assume there are 17 stereotypical groups:
  – Group 1: high-school males
  – Group 2: arts-oriented adults
  – …
  – Group 17: married women ages 35-45

• Each of the 17 Groups has some rating of all movies, for example Group 2 might have ratings as follows:

|   | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
| Group 2 | 35 | 12 | 49 | 07 | 39 | 22 | 50 | 01 | 38 | 24 | 44 | 50 | 00 | 13 | 19 | 26 | 29 | 42 | 31 | 25 | 41 | 07 | 02 | 25 | 31 |
Assumptions for Group Model, cont.

Every current customer’s ratings (profile) can be represented as some weighted sum of group ratings:

Carlos’ rating = \( w_{3,1} \text{ Group}_1 + w_{3,2} \text{ Group}_2 \)
\[ + \ldots + w_{3,17} \text{ Group}_{17} \]

For example, Carlos’ rating of movies might be:

-0.3(high-school male)
+ 2.4(arts-oriented adult) + …
+ … +1.6(married women ages 35-45)
Implications of Group Model

Let $M$ be the unknown complete data matrix on customer ratings

Then under the model assumptions, we have:

$\text{rank}(M) \leq 17$
Optimization Problem to “Solve”

Try to solve:

\[
\text{minimize } \sum_{(i,j) \text{ in } \Omega} (M_{ij} - D_{ij})^2
\]
\[
\text{s.t. } \text{Rank}(M) \leq 17
\]

where \( \Omega \) is the collection of known entries in the matrix \( D \)

A solution \( M \) to this problem will satisfy the rank condition, and \( M \) will be close to our data \( D \).
Low-rank nearby matrix problem:

\[
\text{minimize } M \sum_{(i,j) \in \Omega} (M_{ij} - D_{ij})^2 \\
\text{s.t. } \text{Rank}(M) \leq 17
\]

Suppose \(M\) is a reasonably good solution

Then \(\text{rank}(M) \leq 17\) and \(M \sim D\) (original data)

We will use the entries of \(M\) as the customer ratings for recommending movies.
“Solving” the Optimization Problem

minimize \( M \sum_{(i,j) \in \Omega} (M_{ij} - D_{ij})^2 \)

s.t. \( \text{Rank}(M) \leq 17 \)

- This problem is huge-scale
- And it is NP-hard
- But we don’t need/want to solve it accurately
- There are excellent practical (and theoretically efficient) methods that are used to approximately solve this problem in huge-scale
Recommender System Models

• Is this Operations Research? Why or why not?

• If not, is this applied mathematics? Or engineering? Or computational science?

• This too is “information engineering”

• In my view, this is the future of OR
Recommended Personal Choices

• Work on research that is fun and that you love

• Also try and work on research that is “important”
  – To business or the economy or the military or the environment or the world or….

• Develop and teach curricula that you are passionate about

• Seek impact, somehow, some way….
Backup/Extra Materials
Material on Education here