INTRODUCTION TO DATA ANALYSIS AND MINING

CSCI-B365

Fall 2017
Class meets:
Time: MW 11:15am – 12:30pm
Place: Woodburn Hall 004

Instructor:
Predrag Radivojac
Office: Lindley Hall 301F
Email: predrag@indiana.edu
Web: www.cs.indiana.edu/~predrag

Office Hours:
Time: MW 2pm-3pm
Place: Lindley Hall 301F

Course Web Site:
https://www.cs.indiana.edu/~predrag/classes/2017fallb365
Basic Information

Associate Instructors:

Moses Stamboulian
Email: mstambou@indiana.edu
Office: Lindley Hall 310

Office Hours:
Time: Tuesdays and Thursdays, 10am-11:30am
Place: Lindley Hall 325

Yuxiang Jiang
Email: yuxjiang@indiana.edu
Office: Lindley Hall 310

Office Hours:
Time: Tuesdays and Thursdays, 3pm-4pm
Place: Lindley Hall 325
Textbook

Introduction to Data Mining - by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar

Chapter 1: Introduction
Chapter 2: Data
Chapter 3: Exploring data
Chapter 4: Classification
Chapter 6: Association analysis
Chapter 8: Cluster analysis

Supplementary material will be provided in class!
ALSO GOOD READINGS...

- **Data Mining: Concepts and Techniques** - by Jiawei Han and Micheline Kamber

- **Data Mining: Practical Machine Learning Tools and Techniques** - by Ian Witten and Eibe Frank

- **Programming Collective Intelligence** - by Toby Segaran
Lecture Slides

• Introduction to Data Mining - by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar

• Data Mining: Concepts and Techniques - by Jiawei Han and Micheline Kamber

• Summary: Our own slides + some mix from the slides for the books above
MAIN DEFINITIONS AND GOALS

• Data mining is a generally well-founded practical discipline that aims to identify interesting new relationships and patterns from data (but it is broader than that).

• This course is designed to introduce basic and some advanced concepts of data mining and provide hands-on experience to data analysis, clustering, and prediction.

• The students are be expected to develop a working understanding of data mining and develop skills to solve practical problems.
How High Is Your XQ?
Your next job might depend on it

BY ELIZA GRAY

Is it true to say you have never hated anyone? Do you understand why stars twinkle? Have you used a display of emotion to get what you want? Would you rather read or watch TV? Do you usually notice when you are boring people? Do you hate opera singing? Would you consider yourself to be an ordinary person? Are you shy? Do you prefer problems that require a lot of thought? Do you enjoy giving parties? When you are free do you feel frequently rebellious? Do you prefer to work with people or by yourself? Do you receive help from others more than you give? Have you ever stressed at work? Do you think sometimes you are too hard on yourself? Do you like to have someone around at work? Do you have a lot of things about you that you are proud of? Do you have a lot of friends? Do you like to have people around you? Do you have a lot of friends? Would you like to change your work? Do you make new friends all the time? Do you ever make new friends? Have you ever pretended to know much? Do you need to form friendships at work? Would you see your colleagues if you were very confident? How much does
Find out if your personality fits your job

For a better sense of what these tests are like, TIME asked Hogan Assessments to devise a brief example for readers. The company’s co-founder, Robert Hogan, is a fellow at the Society for Industrial and Organizational Psychology. He says he believes testing can strengthen organizations and place people in the right jobs—but adds that he’s concerned about the lack of transparency and regulation in the growing test industry.

<table>
<thead>
<tr>
<th>Spontaneity</th>
<th>Independence</th>
<th>Competitiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>T F 1. I follow my instincts wherever they lead me.</td>
<td>T F 1. I prefer working alone.</td>
<td>T F 1. Life is a competition.</td>
</tr>
<tr>
<td>T F 2. Planning is one of my best abilities.</td>
<td>T F 2. I am a social person.</td>
<td>T F 2. I don’t care if others are more successful than I am.</td>
</tr>
<tr>
<td>T F 3. My friends would describe me as impulsive.</td>
<td>T F 3. I don’t like to rely on others to help me with my job.</td>
<td>T F 3. Some people think I am too competitive.</td>
</tr>
<tr>
<td>T F 5. I don’t do anything without having a plan in place.</td>
<td>T F 5. I enjoy working on teams.</td>
<td>T F 5. More harm than good is caused by competition.</td>
</tr>
<tr>
<td>T F 7. I go wherever the day takes me.</td>
<td>T F 7. Meeting with others is often a waste of time.</td>
<td>T F 7. I enjoy testing my skills against others.</td>
</tr>
<tr>
<td>T F 8. My friends say that I am unpredictable.</td>
<td>T F 8. I do not like to draw attention to myself.</td>
<td>T F 8. I am at my best when competing with others.</td>
</tr>
</tbody>
</table>

**Now add up your total score.**

<table>
<thead>
<tr>
<th>Score</th>
<th>Score</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>High (6–8): You would fit best in jobs that allow you to work primarily by yourself. Others will likely value your ability to take on and run with your own projects, but you might have difficulty asking for or lending help to others when needed. Moderate (3–5): You would fit best in jobs that contain a mix of independent work and working with others. Others will value your flexibility, but you might struggle during long periods of solitary work or work that requires constant and regular interactions with others. Low (0–2): You would fit best in an environment where you are constantly surrounded by and working with others. You likely derive energy from others, which in turn attracts people to you. You might struggle, however, when asked to focus on specific tasks or goals by yourself.</td>
<td>High (6–8): You would fit best in jobs where success is defined by excelling over others. Others would likely value your drive, determination and willingness to put in more effort than your competition. You might, however, struggle to recognize when you need to put your own personal goals aside for the greater good. Moderate (3–5): You would fit best in environments that contain a mix of collaboration and competition. You are likely able to work toward both individual and team goals simultaneously but must be careful to recognize when to focus on one over the other. Low (0–2): You would fit best in a team-based, collaborative environment. Others will frequently turn to you for help and, in turn, be willing to offer you help as needed. You might, however, struggle in situations where your personal performance depends on your ability to outshine others.</td>
<td></td>
</tr>
</tbody>
</table>
Motivating Example #2, from Reddit

Let Artificial Intelligence guess your attractiveness and age

#howhot
Sweeney published this “attack” in 2001

- anonymized health records of all 135,000 employees + families of the state of Massachusetts
- electoral list of Cambridge, MA – bought for $20 (54,805 people).
- Governor’s health records were identified (87% of people identifiable by b-day, ZIP, gender)
OLD FAITHFUL GEYSER DATA

Old Faithful, Wyoming

- Graph showing the relationship between Eruption time [min] and Waiting time [min].
- The data points are scattered, indicating variability in the eruption waiting times relative to the eruption time.
EVOLUTIONARY TREE

From Wikipedia
AND MORE...

What Can Be Automated?
What Cannot Be Automated?
EXPECTATIONS AND ASSUMPTIONS

• Basic mathematical skills
  – linear algebra, calculus, probabilities

• Some programming skills
  – programming languages: Matlab, Python, R, C/C++, Java

• You are patient and hardworking

• You are motivated to learn and succeed in class

• Your integrity is impeccable
OVERVIEW OF THE COURSE

See online syllabus…

- introduction to data mining
- introduction to Matlab and Matlab programming

- data representation and data preprocessing
- data visualization
- prediction methods (classification and regression)
- mining association rules
- clustering techniques
- privacy-preserving data mining
- case studies on various types of data

- and more... (how much? we’ll see!)
Hidden Goals

• To appreciate fundamental mathematical concepts

• To appreciate abstraction and to not be afraid of it

• To be able to understand how to transfer solutions from one set of problems to another set of problems (through abstraction)

• To learn to be a problem solver

• To recruit undergraduate researchers

• NSF Graduate Research Fellowship Program
# Grading

- Midterm exam: 20% (25%)
- Final exam or project: 30% (25%)
- Homework assignments (5-6): 30%
- Class participation and quizzes (4): 20%

---

- All assignments are individual
- Top performers in the class will earn A
- Distributions of scores will be generated after assignments; regularly
- You will know where you stand in the class, if you don’t - ask me
- Do not expect late I’s or W’s
LATE ASSIGNMENT POLICY

• The homework assignments are due on the specified due date; through Canvas or in class

• Late assignments will be accepted (unless there are legitimate circumstances) using the following rules

  – points (on time) } recommended!

  – points x 0.9 (1 day late)
  – points x 0.7 (2 days late)
  – points x 0.5 (3 days late)
  – points x 0.3 (4 days late)
  – points x 0.1 (5 days late)
  – 0 (after 5 days) } not recommended!
ANNOUNCEMENTS

• Midterm exam – wk 7 (in class); October 4
• Quizzes (4!) – wk 4, wk 9, wk 11 (all in class)
• Final exam time – December 13, 5-7pm
• Labor day – September 4 (no class)
• Thanksgiving break – November 18-26 (no classes)
ACADEMIC HONESTY

• *Code of Student Rights, Responsibilities, and Conduct* !!!

  – Many interesting things there, including that… Students are responsible to “facilitate the learning environment and the process of learning, including attending class regularly, completing class assignments, and coming to class prepared”.

• Academic honesty taken seriously!

  – I have to report every cheating incident to the university
  – Do the right thing
**MISCELLANEA**

- Do not record instructor(s) without explicit written permission
- Turn off cell phones and other similar devices during class
- Use laptops if you have to (unless it bothers someone)
- “will u be in ur office after class”; “I need a letter of recommendation.”

- BE NICE TO PEOPLE