Information Infrastructure II
(Data Mining)

I211

Spring 2010
Basic Information

Class meets:
  Time: MW 9:30am – 10:45am
  Place: I2 130

Instructor:
  Predrag Radivojac
  Office: Informatics 219
  Email: predrag@indiana.edu
  Web: www.informatics.indiana.edu/predrag

Office Hours:
  Time: MW 2:00pm-3:30pm
  Place: Informatics 219

Course Web Site:
  http://www.informatics.indiana.edu/predrag/classes/2010springi211/2010springi211.htm
Basic Information

Associate Instructors:

Michael Conover
Lab Section: 8180/8181 (Office hours TBA)
Email: midconov@indiana.edu
Office: TBA
Time: Check his profile at Informatics web site

Rajeswari Swaminathan
Lab Section: 8181/8180 (Office hours TBA)
Email: rajswami@indiana.edu
Office: TBA
Time: Check her profile at Informatics web site
Textbook Information

Good readings for programming:
• MATLAB primer - by Timothy Davis and Kermit Sigmon (freely accessible at IUCAT)

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

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

- **Principles of Data Mining** - by David Hand, Heikki Manilla, and Padhraic Smyth ([freely accessible at IUCAT](#))
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
Overview of the Course

See online syllabus...

- introduction to Matlab and Matlab programming
- introduction to data mining
- data representation and data preprocessing
- data visualization
- concepts from linear algebra
- concepts from probability and information theory
- mining association rules
- classification and regression methods
- clustering techniques
- case studies on various types of data
- and more... (how much? we’ll see!)
Goal of the Course!!!

• This course is designed to introduce basic concepts of Data Mining and provide hands-on experience to data analysis, clustering, and prediction by using Matlab.

• The students will be expected to develop a basic understanding of Data Mining and develop skills to solve practical problems.

• Data Mining is a practical discipline that aims to identify interesting new relationships and patterns hidden in numerous databases and real life.
Hidden Goals of the Course

• 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
What do I assume?

• You have taken I210 and you can program in C, Java, or Python…
• You have basic mathematical skills (e.g. calculus)
• You are patient

What would I like to see?

• You are motivated to learn
• You are motivated to succeed in this class
Grading Policy and Announcements

• Midterm exam (1): 20%
• Final exam (1): 20%
• Homework assignments (6-7): 40%
• Class participation and quizzes (4): 20%

Midterm exam – week 7 (in class), maybe week 8
Final exam – May 5 (8:00am)
Spring break – March 15-19
MLK Jr. day – January 18 (no classes!)
Late Assignment Policy and Academic Honesty

• The homework assignments are due on the specified due date through Oncourse.

• 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!

• All assignments are individual!!!

• All the sources used for problem solution must be acknowledged (people, web sites, books, etc.)
Late Assignment Policy and Academic Honesty

- Grading: top performers in the class will earn A, class average will be about B
- Distributions of scores will be generated after assignments; regularly
- You will know where you stand in the class, if you don’t - ask instructor
- Do not expect late I’s or W’s

- Read *Code of Student Rights, Responsibilities, and Conduct* !!!
  - [http://www.indiana.edu/~code/](http://www.indiana.edu/~code/)
  - 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”.
  - Students ≠ Customers (*not* for instructors, *yes* for administration)
Some Specifics for I211

• Do not record instructor(s)

• Turn off cell phones, smart phones, and other similar devices during class

• Use laptops only if you have to

• Be considerate
OK, Let’s Start!

- **Lots of data is being collected**
  - web data, e-commerce
  - purchases at department/grocery stores
  - bank/credit card transactions
  - medical data, biological data
  - streaming data
  - agricultural data

- **Computers have become cheaper and more powerful**

- **Competitive pressure is strong**
  - provide better, customized services for an *edge*
Why mine data? Scientific viewpoint.

• Data collected and stored at enormous speeds (GB/hour)
  – remote sensors on a satellite
  – telescopes scanning the skies
  – microarrays generating gene expression data
  – scientific simulations generating terabytes of data

• Traditional techniques infeasible for raw data

• Data mining may help scientists
  – in classifying and segmenting data
  – in hypothesis generation
Mining large data sets - Motivation

• There is often information “hidden” in the data that is not readily evident
• Human analysts may take weeks to discover useful information
• Much of the data is never analyzed

The Data Gap

Total new disk space (TB) since 1995

Number of analysts

From: R. Grossman, C. Kamath, V. Kumar, “Data Mining for Scientific and Engineering Applications”
More on the data explosion...

• We’re drowning in information and starving for knowledge.
  
  Rutherford D Rogers, librarian, Yale
  
  (NY Times 25 Feb 85)

In Data Mining we typically say

• We are drowning in data, but starving for knowledge!

We need

• Automated analysis of massive data sets
What is Data Mining?

• Alternative name: Knowledge Discovery from Data (KDD)
• Many definitions
  – Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  – Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

• Why Data Mining name?
  – Gold mining is looking for gold, correct?
  – A few interpretations
Some (not so useful) patterns...

• “rules” for American presidents (before 2004 elections)
Some (not so useful) patterns...

• “rules” for American presidents (before 2004 elections)
  – if the Washington Redskins win their last home game before the election, the incumbent’s party will be re-elected
Some (not so useful) patterns...

• “rules” for American presidents (before 2004 elections)
  – if the Washington Redskins win their last home game before the election, the incumbent’s party will be re-elected
  – no Republican has ever won a presidential election without carrying Ohio
Some (not so useful) patterns...

• “rules” for American presidents (before 2004 elections)
  – if the Washington Redskins win their last home game before the election, the incumbent’s party will be re-elected
  – no Republican has ever won a presidential election without carrying Ohio
  – no incumbent with a four-letter last name has ever been re-elected (Polk, Taft, Ford, Bush Sr.)
Some (not so useful) patterns...

• “rules” for American presidents (before 2004 elections)
  – if the Washington Redskins win their last home game before the election, the incumbent’s party will be re-elected
  – no Republican has ever won a presidential election without carrying Ohio
  – no incumbent with a four-letter last name has ever been re-elected (Polk, Taft, Ford, Bush Sr.)
  – Americans won’t unseat a wartime President
What is NOT Data Mining?

**NOT Data Mining**

- look up phone number in phone directory
- query a Web search engine for information about “Amazon”
- SQL query processing
- statistical tools
- data visualization?
- data warehousing?

**YES Data Mining**

- Certain names are more prevalent in certain US locations (O’Brien, O’Rurke, O’Reilly… in Boston area)
- Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com, Amazons – all female warriors in Greek mythology)
Data Mining: confluence of disciplines

- Database Technology
- Statistics
- Machine Learning
- Pattern Recognition
- Algorithms
- Visualization
- Other Disciplines
Data Mining and Business Intelligence

Increasing potential to support business decisions

Decision Making

Data Presentation
Visualization Techniques

Data Mining
Information Discovery

Data Exploration
Statistical Summary, Querying, and Reporting

Data Preprocessing/Integration, Data Warehouses

Data Sources
Paper, Files, Web documents, Scientific experiments, Database Systems

End User
Business Analyst
Data Analyst
DBA
Data Mining tasks

• Predictive methods
  – Use some variables to predict unknown or future values of other variables

• Descriptive methods
  – Find human-interpretable patterns that describe the data

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996