CS373 Data Mining and Machine Learning
Lecture 1

Jean Honorio
Purdue University
Today’s topics

• First things first…
• Course goals
• Course introduction
First things first...

• Website: [http://www.cs.purdue.edu/homes/jhonorio/19fall-cs37300.html](http://www.cs.purdue.edu/homes/jhonorio/19fall-cs37300.html)
  - TAs
  - Textbooks
  - Assignments
  - Grading
  - Late policy
  - Schedule
Lectures

• Do not miss lectures
  - Ask questions when things are unclear

• Slides are available before lecture on the course website
  - “Whiteboard” notes are also available after lectures

• I will not repeat lectures or summarize lectures before an exam (midterm or final)

• Lectures, homeworks and project are complementary forms of learning

• Some lectures will be Case Studies

• I might cancel some lectures
  - Due to conferences, for instance, check e-mail and course website

• Never post lecture slides online, at any point in time
Homeworks

• Do not miss any of the 8 homeworks
  - Check your email and course website

• *Solve all homeworks yourself*

• *No extensions*
  - Not even one minute
  - Submit before the due date

• Some questions require implementing Machine Learning algorithms

• *Submission through Blackboard*

• *Never post the homework questions or your solutions online, at any point in time*
**Piazza**

- **Ask questions in Piazza**
  - Do not ask questions to me or TAs through email, unless your question relates to (possible) inaccuracies in your grades.
- **Wait up to one business day for an answer**
- Never ask one business day before a homework deadline, before an exam, etc.
- **Do not disclose homework solutions (total or partial) when asking**
- **Ask professionally**
  - For instance, do not say things such as “it does not make any sense”
Office hours

• *Each TA will hold office hours on different days of the week*
  - *Doodle soon*

• I do not hold regular office hours

• If after the TA answer, you still do not understand the material, please contact me asap
  - This has never happened so far
  - Send me an e-mail and we can arrange a meeting

• The TAs or me will devote at most approximately 10 minutes per student
  - Bring specific questions to the office hours

• Do not ask the TA or me to repeat a lecture that you missed
  - You should not miss lectures
Exams

• *Some midterm and final exams’ questions will be multiple choice*

• Questions are carefully created to avoid misunderstanding

• Exams will have several questions to avoid losing too many points for a wrong answer

• No points deducted when answering wrong

• *We define a general rule across the whole section for assigning partial points*
  - Objective and fair
Project

- *Exactly 2 members for each group*
- Project plan is due 1-2 weeks before the midterm
- Start thinking about your group members soon
  - If you did not find a group, I will assign you to a group randomly
- Every group member should make similar effort
  - We will talk about this later
- *More details in the course website*
Today’s topics

• First things first…
• Course goals
• Course introduction
Goals

• Identify key elements of data mining and machine learning algorithms
• Understand how algorithmic elements interact to impact performance
• Understand how to choose algorithms for different analysis tasks
• Analyze data in both an exploratory and targeted manner
• Implement and apply basic algorithms for supervised and unsupervised learning
• Accurately evaluate the performance of algorithms, as well as formulate and test hypotheses
• Elements of data mining algorithms
• Statistical basics and background
• Data preparation and exploration
• Predictive modeling
• Methodology, evaluation
• Descriptive modeling
• Pattern mining and anomaly detection
Today’s topics

• First things first…
• Course goals
• Course introduction
Data mining

The process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data

(Fayyad, Piatetsky-Shapiro & Smith 1996)

Machine learning: How can we build computer systems that automatically improve with experience? (Mitchell 2006)
The last 35 years of research in ML/DM has resulted in widespread adoption of predictive analytics to automate and improve decision making.

As “big data” efforts increase the collection of data… so will the need for new data science methodology. Data today have more volume, velocity, variety, etc.

**Machine learning** research develops statistical tools, models & algorithms that address these complexities.

**Data mining** research focuses on how to scale to massive data and how to incorporate feedback to improve accuracy while minimizing effort.
What is big data?

Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data.
The data mining process

Data

Selection

Target data

Preprocessing

Processed data

Knowledge

Interpretation evaluation

Patterns

Machine Learning

Mining
1. Application setup:
   • Acquire relevant domain knowledge
   • Assess user goals

2. Data selection
   • Choose data sources
   • Identify relevant attributes
   • Sample data

3. Data preprocessing
   • Remove noise or outliers
   • Handle missing values
   • Account for time or other changes

4. Data transformation
   • Find useful features
   • Reduce dimensionality
The data mining process

5. Data mining:
   • Choose task (e.g., classification, regression, clustering)
   • Choose algorithms for learning and inference
   • Set parameters
   • Apply algorithms to search for patterns of interest

6. Interpretation/evaluation
   • Assess accuracy of model/results
   • Interpret model for end-users
   • Consolidate knowledge

7. Repeat...
Example

These trains carry toxic chemicals

These trains do not carry toxic chemicals

Does this train carry toxic chemicals?
Example rule (1)

These trains carry toxic chemicals

These trains do not carry toxic chemicals

Does this train carry toxic chemicals?
Example rule (2)

These trains carry toxic chemicals

These trains do not carry toxic chemicals

Does this train carry toxic chemicals?
How did you devise rules?

- Look for characteristics of one set but not the other?
- Reject potential rules that didn’t cover enough examples?
- Examine several potential rules?
- Consider simple rules first?
This is data mining...

- Data representation: Describe the data
- Task specification: Outline the goal(s)
- Knowledge representation: Describe the rules
- Learning technique:
  - Search: Identify a rule
  - Evaluation function: Estimate confidence
- Prediction technique: Apply the rule
- Data mining system: Do above in combination
Complexities

- Data size: vastly larger or changing rapidly
- Data representation: can affect ability to learn and interpret models
- Knowledge representation: needs to capture more subtle forms of probabilistic dependence
- Search space: vastly larger