COMP 543: Tools & Models for Data Science

Course Overview

Chris Jermaine & Risa Myers

Rice University

This Class is about *Data Science*

- What is THAT?
- Extraction of actionable knowledge from large volumes of data
    - Encompasses methods from:
        - Computer science
        - Statistics
        - Optimization/Applied Math
    - Also includes
        - Domain knowledge
        - Communication skills
        - Data management

Examples of Data Science Tasks

- Given a huge set of per-customer sales data, build a model to predict customer “churn”
- Given a large graph of Medicare payout data, find suspicious (potentially fraudulent) referral patterns
- Given a set of EMR data, find previously unknown side effects (ex: Vioxx and heart disease)
- Given data from an online learning tool find markers that are an early sign of later academic achievement problems
- Many, many more!

Both Tools and Models are Important

- Back in the day...
  - You had statisticians who dealt primarily with small data sets
  - You had computer scientists who were interested in advanced modeling
- But in the “Big Data” era, the two can’t live in isolation
  - You need advanced models to solve challenging prediction/analysis tasks
  - You need computer systems that can scale those models to the largest data sets
  - You need computer tools that make it easy to implement complicated models

Important Disclaimer

- 543 is a hard-core computer science class!
- This is not “tools and models” from a naive user’s perspective
  - No learning to be an end-user of classical analytics packages
  - This is not a “Get to know R” class
  - Nor is it a “Get to know SAS” class
  - No plugging data into a standard software package and writing a report on the results
  - A class covering such topics WOULD be useful
    - But that’s simply not this class
- Lots of focus on algorithms and engineering

Problem Domain

- Focus on data / problems in the biomedical domain
- Familiarity is helpful, but not required
- The models and tools are general, but we will explore applications in medicine and bioinformatics
  - Gene-gene interactions
  - Prescribing practices
  - Genetic sequence similarities
  - Medical publication abstract analysis

When We Say “Models”

- Strong focus on the math foundations of data science
- Lots of optimization theory, probability, statistics
- Even some continuous mathematics
- Here’s a slide from one of the later lectures:

Example Slide

- Nasty!! Or is it? Consider just one varible, z_1; try to separate out. Write as:

$$
\begin{aligned}
&= c_{1,1}\left(x_i \log(p_1) + (10 - x_i)\log(1 - p_1)\right) \sum_{\langle z_2, z_3, \ldots \rangle} a(\langle z_2, z_3, \ldots \rangle) \\
&+ \sum_{\langle z_2, z_3, \ldots \rangle} a(\langle z_2, z_3, \ldots \rangle) \sum_{i=2}^{n} b(\langle z_2, z_3, \ldots \rangle) \\
&+ c_{1,2}\left(x_i \log(p_2) + (10 - x_i)\log(1 - p_2)\right) \sum_{\langle z_2, z_3, \ldots \rangle} a(\langle z_2, z_3, \ldots \rangle) \\
&+ \sum_{\langle z_2, z_3, \ldots \rangle} a(\langle z_2, z_3, \ldots \rangle) \sum_{i=2}^{n} b(\langle z_2, z_3, \ldots \rangle) \\
&= c_{1,1}\left(x_i \log(p_1) + (10 - x_i)\log(1 - p_1)\right) + \text{other terms w/o } z_1 \\
&+ c_{1,2}\left(x_i \log(p_2) + (10 - x_i)\log(1 - p_2)\right) + \text{other terms w/o } z_1
\end{aligned}
$$

When We Say “Tools”

- We mean tools for manipulating large data sets
- Tools for scalable, distributed computation
- Focus is on “Big Data”!
- Specifically, we’ll learn about:
  - SQL databases
  - Python programming (NumPy, SciPy)
  - Hadoop (MapReduce software, Big Data file system)
  - Spark (distributed Big Data manipulation software)
  - TensorFlow (tool for building learning algorithms)

Example Use Case for Your 543 Skill Set

- Imagine...
  - You are working at a hospital
  - You collect 5TB of patient monitoring data each day...
  - And want a software to predict what will happen to a patient in the next hour
  - Such a software does not exist...
  - How to build it?
- Key questions to answer:
  - How will you process the raw data?
  - What model will you use to do prediction?
  - How will you train the model?
  - How will you scale to 5TB per day?
- After 543, you'll have the answers!

As Such, this Class...

- Will give an introduction to modern data management software...
    - First half of the class
    - Relational database systems and SQL
    - No-SQL systems such as Hadoop and Spark
- Will give an introduction to models for modern data analysis...
    - Second half of the class
    - Basic optimization theory
    - Supervised learning (linear models, support vector machines)
    - Unsupervised learning (clustering, matrix factorization)
    - Text mining
- Projects will focus on implementing the models using the tools

Skills You Need to Take this Class

- Should be a proficient programmer
  - Really good in a modern, general-purpose language
  - Python preferred
  - Two assignments use SQL (no knowledge assumed)
  - Four assignments use Python

More Skills You Need to Take this Class

- Should not be afraid of a bit of math
    - Some background in probability/statistics
        - Common distributions (e.g. Gaussian)
        - Expected value
        - Variance, covariance
        - Norms (e.g. L_1, L_2)
    - Some calculus (partial derivatives & the chain rule should not freak you out!)
    - Linear algebra
        - Vectors and scalars
        - Matrix inversion
        - Matrix transposition
        - Dot products
- Fluency in English to be able to read research papers, evaluate them critically, and find related papers

What About Overlap with Other Classes?

- COMP 533—biggest overlap
  - First three weeks of class are going to strictly be review
  - As will be the first two assignments (a lot like COMP 533 assignments)
- COMP 440/502/540/602
  - Many/all of the methods we'll cover will also be covered in those classes
- So, what's the point of taking this class?
  - The only place where you can get an overview of all of this in one place
  - Focus on big data and tools that operate on big data

Teaching Assistants

- Gau Pan
- Sean Wang
- Gabe Vacaliuc
- Office hours will be posted on Piazza

Class Syllabus

- Communication...
- Grading and Evaluation...
- Exams...
- Academic misconduct...
- Assignments... (more on the next slide)

Assignments

- Exercises
  - 4 short programming exercises designed to reinforce in-class concepts
- Labs
  - 7 one-hour activities to get initial hands-on experience with a practical concept
- Programming Assignments
  - 6 in-depth programming assignments
- Research
  - 7 research / writing assignments to increase understanding of a key topic and how it has been used in a domain of interest

Class Policies – Due Dates

- **Assignment & Exercise** Due Dates
  - Typically due at 11:55 PM
  - 1 second – 24 hours late = 10% penalty
  - 24 hours + 1 second – 48 hours late = 20% penalty
  - > 48 hours late: NOT ACCEPTED
  - Last assignment may **NOT** be submitted late
- Canvas is the time keeper – if Canvas says it's late, it's late
- Exceptions will only be made for EXTENDED Canvas outages
- Submit early!

Class Policies – Extensions

- Must be requested $\geq$ 1 week in advance
- Exceptions possible for very extenuating circumstances, with proper documentation

Class Policies – Regrades

- Must be requested within 1 week of assignment being returned
- Intended for errors in grading or MINOR errors
- Not a week-long extension to the assignment
- Process
  - Talk to Risa, after class or during office hours
  - **Type** up request
  - Submit in person or under the door to DH 2062

Questions?

- If there’s time: on to databases!!!
  - What’s a database system?