95-801 Data Mining Techniques  
Fall 2017

Announcements

- **10/24:** First day of class, welcome!
- **8/5:** Students are encouraged to study the Syllabus to have a general understanding of the course organization, as well as the Assignments to have an idea about the workload.

PEOPLE:

Instructor: Leman Akoglu, (lakoglu@andrew)  
- Office: HBH 2118C, office ph 412-268-30 four three  
- Office hours: Tue 1-2 PM; also, by appointment

Teaching Assistant: Abhinav Maurya, (amaurya@andrew)  
- Office: TBD  
- Office hours: TBD

CLASS MEETS:

When: Tue & Thu 4:30-5:50 PM  
Where: HBH 1202

COURSE DESCRIPTION:

Knowledge discovery from data is “the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data” --- Fayyad et al. (1996)

Motivation: Data generated by humans and machines is available everywhere and growing steadily. In today’s data-driven world it is crucial for students to acquire the fundamental skills for being able to analyze massive datasets and to develop data-driven techniques toward solving real-world problems.

This course will cover the fundamental concepts and techniques in data mining, and equip students with the basic skillset toward becoming good data scientists. Major topics include algorithms and tools for data exploration, fast similarity search, pattern mining, outlier detection, dimensionality reduction, ranking, and recommender systems. See the syllabus for details. The coursework involve mini-projects on various datasets to enable students to gain hands-on experience with data analytics.

Learning Objectives

By the end of this class, students will

- develop basic understanding of core data mining concepts
- learn algorithmic approaches to various data mining problems
- be able to analyze and assess data mining algorithms based on their accuracy, computational/storage complexity, and the tradeoffs thereof.
- gain hands-on experience using data analytics techniques on real-world datasets

TEXTBOOKS:

We expect students to have a copy of the following book, from which most readings will be assigned.
Below you can find a list of other recommended books to learn certain subjects in more depth. I will also post the lecture notes on the Blackboard.

- Anand Rajaraman, Jure Leskovec, and Jeffrey D. Ullman. "Mining of Massive Datasets" (FREE)

- Charu C. Aggarwal. "Data Mining, The Textbook"

- Jiawei Han, Micheline Kamber, and Jian Pei. "Data Mining: Concepts and Techniques," 3rd edition, Morgan Kaufmann, 2011

- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. "An Introduction to Statistical Learning " (FREE here)

- Foster Provost, Tom Fawcett. "Data Science for Business"


See resources for other pointers.

**BULLETIN BOARD and other info**

- We will use the Blackboard for course materials, homework deposits, announcements, and grades.
- We will use Piazza for questions and discussions.
- Carnegie Mellon 2017-2018 Official academic calendar

**MISC - FUN:**

joke1   joke2   joke3
Syllabus

You can find the list of topics by lecture below. Readings for each lecture will be posted here. Note that the syllabus is tentative and will be adjusted, if needed, as the semester proceeds.

<table>
<thead>
<tr>
<th>Date</th>
<th>Lectures and Readings</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/24</td>
<td><strong>Lecture 1: Introduction and Fast Similarity Search</strong></td>
</tr>
<tr>
<td>10/26</td>
<td>- kd-trees and Locality Sensitive Hashing</td>
</tr>
<tr>
<td></td>
<td>Reading:</td>
</tr>
<tr>
<td></td>
<td>- Multidimensional Binary Search Trees Used for Associative Searching, J.L. Bentley. CACM, 18,9, Sept. 1975</td>
</tr>
<tr>
<td></td>
<td>- Mining of Massive Datasets Chapter 3</td>
</tr>
<tr>
<td>10/26</td>
<td><strong>Lecture 2: Frequent Itemsets and Association Rules</strong></td>
</tr>
<tr>
<td>11/2</td>
<td>- Market-basket analysis and the Apriori algorithm</td>
</tr>
<tr>
<td></td>
<td>- Handling large datasets with limited-RAM and limited-pass algorithms</td>
</tr>
<tr>
<td></td>
<td>Reading:</td>
</tr>
<tr>
<td></td>
<td>- Mining of Massive Datasets Chapter 6</td>
</tr>
<tr>
<td>11/7</td>
<td><strong>Lecture 3: Dimensionality Reduction</strong></td>
</tr>
<tr>
<td>11/9</td>
<td>- Singular Value Decomposition (SVD)</td>
</tr>
<tr>
<td></td>
<td>- SVD applications, case studies</td>
</tr>
<tr>
<td></td>
<td>- CUR for sparse decomposition</td>
</tr>
<tr>
<td></td>
<td>Reading:</td>
</tr>
<tr>
<td></td>
<td>- Mining of Massive Datasets Chapter 11</td>
</tr>
<tr>
<td></td>
<td>(Optional) Tutorial on Principal Component Analysis (PCA), J. Schlens.</td>
</tr>
<tr>
<td></td>
<td>(Optional) Quantifiable data mining using ratio rules, F. Korn; A. Labrinidis; Y. Kotidis; C. Faloutsos. VLDB Journal 8(3-4): 254-266, 2000</td>
</tr>
<tr>
<td>11/14</td>
<td><strong>Lecture 4: Clustering</strong></td>
</tr>
<tr>
<td></td>
<td>- Distance measures</td>
</tr>
<tr>
<td></td>
<td>- Hierarchical clustering, k-means, BFR, CURE algorithms</td>
</tr>
<tr>
<td></td>
<td>Reading:</td>
</tr>
<tr>
<td></td>
<td>- Mining of Massive Datasets Chapter 7</td>
</tr>
<tr>
<td>11/16</td>
<td><strong>Lecture 5: Outlier Mining</strong></td>
</tr>
<tr>
<td></td>
<td>- Extreme-value analysis</td>
</tr>
<tr>
<td></td>
<td>- Density-based outlier detection</td>
</tr>
<tr>
<td></td>
<td>- Ensemble methods</td>
</tr>
<tr>
<td></td>
<td>Reading:</td>
</tr>
</tbody>
</table>
Lecture 6: Graphs: Link Analysis

- Ranking nodes in a graph
- Random walks (with restart), Pagerank, Topic-sensitive Pagerank, HITS

Reading:

- Mining of Massive Datasets Chapter 5

No Class: Thanksgiving

Lecture 7: Social Networks

- Centrality, communities, social influence and propagation

Reading:

- Mining of Massive Datasets Chapter 10.1, 10.2

Lecture 8: Time Series

- Similarity measures, linear and non-linear forecasting

Reading:

- Data Mining, The Textbook Chapter 14.1-14.3

Lecture 9: Data Streams

- Uniform-sampling: Reservoir sampling
- Filtering: the Bloom filter
- Counting distinct elements: Flajolet-Martin algorithm
- Counting frequencies: Count-min sketch

Reading:

- Mining of Massive Datasets Chapter 4
- (Optional) Data Mining, The Textbook Chapter 12

Lecture 10: Text Mining

- Topic modeling with LDA and visualization

Reading:

Assignments

COURSEWORK:
Coursework consist of 3 homework assignments and 1 take-home course project on data mining (grading in parentheses):

- **Homework** (60%, 20% each)
- **Course Project** (30%)
- In-class Participation (10%)

IMPORTANT DATES:

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Note</th>
<th>Out</th>
<th>Due</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homework 1</td>
<td>similarity search, freq. itemsets, SVD</td>
<td>Oct 31</td>
<td>Nov 14</td>
<td>20%</td>
</tr>
<tr>
<td>Homework 2</td>
<td>clustering, outliers, link analysis</td>
<td>Nov 14</td>
<td>Nov 28</td>
<td>20%</td>
</tr>
<tr>
<td>Homework 3</td>
<td>social nets, time series, streams, text</td>
<td>Nov 28</td>
<td>Dec 12</td>
<td>20%</td>
</tr>
<tr>
<td>Project Write-up</td>
<td>Open-ended problem (NO LATE DAYS!)</td>
<td>Dec 12</td>
<td>Dec 15</td>
<td>30%</td>
</tr>
<tr>
<td>Class participation</td>
<td>Pop-quizzes</td>
<td>--</td>
<td>--</td>
<td>10%</td>
</tr>
</tbody>
</table>

HOMEWORK:
The goal of the homework is to enable the students to practice the concepts learned in class using real-world datasets.

- ASSIGNMENTS ARE DUE AT THE BEGINNING OF LECTURE ON THE DUE DATE.
- All assignments are to be done **individually**. Please see the collaboration policy.
- To submit:
  - Submit your soft-copy in .pdf as well as all code in .zip on Blackboard.
  - Return a printed hard-copy on due date in class.
  - Make sure your answers are clear and writing is legible.
  - See course policies for assignment questions, late submissions, graded homework pick-up.

COURSE PROJECT
The goal of the project is to give the students the opportunity to tackle a large, interesting data mining problem using data analysis techniques and tools.

For the project, we will provide you with a list of large datasets as well as a list of data mining (DM) problems possible on the provided datasets. Your task will be to choose one of those DM problems, or define your own, on the provided dataset and address the problem of your choice with the big data analysis tools you learned during the course.

By design, the project is open-ended; you are free to decide how you want to approach the problem and what tools you want to employ. We want to see a best-effort solution that utilizes what you learned in class and also potentially trying new things beyond class.

Important things to note:

- This is a **take-home project**. A detailed handout about the project will be distributed at the end of
the course.

- The handout will specify a list of datasets, along with a list of suggested projects.
- You can define their own project/problem statement but they must use one of the provided datasets.
- Take-home projects will be carried out individually.
- You will have 72 hours to complete the project.
- You will submit a project write-up before the due date (NO LATE DAYS ALLOWED). This report should be 8 pages maximum, 4 pages minimum, including references and figures; this page limit is strict.
- We expect you to use a tool/skill-set acquired during the course.
- We expect you to address the problem sufficiently well to obtain interesting insights and/or solutions.
- Your project will be worth 30% of your final course grade.

Project writeup

Course staff will use the following guidelines when grading your final project reports.

- **Introduction/Motivation/Problem Definition (10%)**
  - What is it that you are trying to solve/achieve and why does it matter?

- **Data Understanding and Preparation (15%)**
  - What preliminary analyses have you performed on the data? What observations have you made? How did those observations help shape your approach?
  - Provide the preliminary data analysis results and your observations.

- **Methodology (35%)**
  - This is where you give a detailed description of your primary contributions. It is especially important that this part be clear and well written so that we can fully understand what you did.
  - Specify your approach, in concrete steps.
  - Discuss choices for the data mining algorithm: what are other alternatives, and what are their pros and cons (in the context of the problem and as compared to your proposed solution)?
  - Discuss why and how this approach should “solve” the problem (i.e., improve along some dimension of interest).
  - Outline the big data analysis tools and libraries you have used.

It is not so important how well your method performs but rather, (a) how thorough and careful your methodology is, and (b) how interesting and clever the approaches your took and the tools you have used are.

- **Evaluation and Results (30%)**
  - We are interested in seeing a clear and conclusive set of experiments which successfully evaluate the problem you set out to solve. Make sure to interpret the results and talk about what we can conclude and learn from your approach.
  - How do you evaluate your solution to the specific question(s) you have addressed?
  - What do these results tell you about your solution?
  - Present and discuss your evaluation results and findings. You may use tables or figures to visualize and summarize your results.

- **Style and writing (10%)**
  - Overall writing, organization, figures and illustrations.
Course Policies

LECTURES

- All devices such as laptops, cell phones, noisy PDAs, etc. should be turned off for the duration of the lecture and the recitations, because they may distract our fellow students.
- Please come to all lectures on time and leave on time, again so that there are no distractions to the classmates.

PREREQUISITES

Students are expected to have the following background:

- Working knowledge of algorithms and data structures
- Understanding of the basics of probability and linear algebra
- Working knowledge of basic computer science principles at a level sufficient to write a reasonably non-trivial computer program in a language of preference

ASSIGNMENTS

- Assignments are due at the *beginning of lecture* on the due date.
- Due date of assignments are posted at the assignments page.
- Assignments will be posted on Blackboard.
- Students should submit the programming part of assignments electronically via Blackboard (no print outs).

Important Note: As we reuse problem set questions, covered by papers and webpages, we expect the students not to copy, refer to, or look at the solutions in preparing their answers. Since this is a graduate-level class, we expect students to want to learn and not google for answers. The purpose of problem sets in this class is to help you think about the material, not just give us the right answers. Therefore, please restrict attention to the books mentioned on the front page when solving problems on the problem set. If you do happen to use other material, it must be acknowledged clearly with a citation on the submitted solution.

Questions and requests

- You should use Piazza for all your questions about the assignments and the course material. Instructor and TA(s) will do their best to answer your questions timely.
- Rergrade requests should be done in writing/email,
  - within 2 days after graded assignments are distributed
  - to the TA that graded the question, and specifying
    - the question under dispute (e.g., 'HW1-Q.2.b')
    - the extra points requested (e.g., '2 points out of 5')
    - and the justification (e.g., 'I forgot to divide by variance, but the rest of my answer was correct')
  - In the remote case there is no satisfactory resolution, please contact the instructor.

Homework pick-up information

- You may pick up graded homeworks etc., from the course admin
  - Mrs. Adrienne McCorkle, HBH 2250
  - 9:00-11:30am, 1:30-4:30pm every weekday
  - with photo-id (for your privacy protection)

Late policy
• No delay penalties, for medical/family/etc. **emergencies** (bring written documentation, like doctor's note).
• Each student is granted '3 slip days' total for the whole course duration, to accommodate for coinciding deadlines/interviews/etc. That is, no questions asked, if the total delay is 3 days or less.
  ○ You can use the extension on any assignment during the course. For instance, you can hand in one assignment 3 days late, or 3 different assignments 1 day late each.
  ○ Late days are rounded up to the nearest integer. For example, a submission that is 4 hours late will count as 1 day late.
  ○ After you have used up your slip days, any assignment handed in late will be marked off **25% per day of delay**.
• To use slip days:
  ○ upload your homework on **Blackboard to mark the time of submission**
  ○ **make sure to return hard copy** next time in class
  ○ note on front page of hard copy submission: count of slip days you used, as well as the count of slip days left

**Collaboration**

You are encouraged to discuss homework problems with your fellow students. However, the work you submit must be your own. You must acknowledge in your submission any help received on your assignments. That is, you must include a comment in your homework submission that clearly states the name of the student, book, or online reference from which you received assistance.

Submissions that fail to properly acknowledge any help from other students or non-class sources will receive NO credit. Copied work will receive NO credit. Any and all violations will be reported to the Heinz College administration and may appear in the student's transcript.

**Academic integrity**

All students are expected to comply with **CMU's policy on academic integrity**. Please read the policy and make sure you have a complete understanding of it.