CS 422: Data Mining

Department of Computer Science
Illinois Institute of Technology
Vijay K. Gurbani, Ph.D.

Spring 2018

Course description and expectations

This course explores the implementation and application of essential data mining concepts and algorithms. It a survey-style course that introduces you to a smorgasbord of algorithms used in the field. During the semester the course will provide a survey of fundamental algorithms, including but not limited to, market basket analysis, nearest neighbor, decision trees, frequent itemsets, regression and classification and clustering. By the end of the semester, the student should be well versed in data mining architectures, vocabulary, techniques, and should be conversant in using appropriate tools to build data mining models and interpret the output of such models. They should be able to evaluate scalability properties of specific algorithms studied and should be able to build data mining systems using the tools, techniques and algorithms studied in the class.

The course is structured around lectures, programming assignments, 2-3 discussion topics and exams (midterm and final). The pre-requisites for the course are either CS 331 or CS 401 (Data Structures). It is expected that the student has a strong grounding in the role of data structures in computer science, including aspects of understanding computational complexity and evaluating runtime complexity of algorithms. While there is no pre-requisite of linear algebra, probability and statistics, it is nonetheless helpful to have the required background in these mathematical areas to obtain the most from the lectures and related learning material. Wherever appropriate, minimal background material related to these mathematical concepts will be covered, however, the implicit understanding remains that the more exposure you have to concepts in linear algebra and statistics, the better the grasp of the lecture.

To further help in the comprehension of the material, periodic homework will be assigned that will consist of programming problems and selected exercises from the book. An assignment will incur a **10% penalty** each day that it is late; so, an assignment submitted 1 day late will accrue 90% of the earned points, an assignment submitted 2 days late will accrue 80% of the earned points, and so on. In the event that the assignment is handed a week past the due date it will only accrue 30% of the earned points. The programming language used for code in the lectures and laboratory assignments will be R using the Rstudio development platform. Students should quickly gain familiarity in R and RStudio. Students who are not familiar with R will be provided a tutorial and related resources to learn the language; as a practical matter it is easy to pick up R if you are already familiar with any block structured procedural programming language like Java, C and its derivatives, or Python.

The homeworks will consist of two parts: the first part will contain questions and answers drawn from the book or other resource, and the second part will contain programming assignments pertinent to the data mining algorithm under study. All homework must be submitted by creating an archive that **must contain only the following files**:

1. A PDF file **(and only a PDF file, no other file formats will be accepted)** corresponding to the first part, either processed through LaTeX (preferred), a word processor of your choice, or neatly written in hand;
2. The programming assignments can be handed in using **one** of the following formats:
    i. A .Rmd file (R markdown file) that contains markdown with embedded R code chunks that can be loaded and executed. Corresponding to the .Rmd file should be a HTML notebook file that is produced by processing the .Rmd file. (If your homework file is saved as hw0.Rmd, processing it will produce hw0.nb.html) **Both files should be submitted and each question in the assignment should be clearly marked.**
    ii. If you use Jupyter notebooks, you can install an R kernel in the notebook by following instructions at https://www.datacamp.com/community/blog/jupyter-notebook-r. You will need to save the notebook as a .html file **(not a .ipynb file)** and the corresponding code in an .R file. Then, submit **both** the files. As before, **each question in the assignment should be clearly marked.**

The student is expected to create an archive (accepted formats for an archive: zip or tgz) and bundle all of the files related to a homework in the archive. The archive will then be uploaded to Blackboard.

There will be two Blackboard discussion groups held on appropriate topical areas. Student participation is mandatory in these discussion groups. You will be assigned papers to read, videos to watch on a specific topic related to data mining. You will be required to review the academic paper or video assigned during the discussion group, and furthermore, you will also be asked to critique the review of your peers in order to foster a discussion. **Discussions are due on the required date**, **there is no late submission policy**. Please upload your picture when you start your first discussion topic; that will allow me to associate names with students as the semester progresses.

There will be a midterm and a final exam; these will be closed books and closed notes. More information on the exams will be forthcoming as the semester progresses.

There will be no individual make-up homework assignments, projects or exams. Please do not ask me for individual efforts to better your grade at the end of the semester. Doing so is not fair to the remaining students, and as such, if a make-up resource is assigned, it will be assigned to the entire class instead of a few chosen students.

Students are expected to adhere to Illinois Tech's Code of Academic Honesty, please see https://web.iit.edu/student-affairs/handbook/fine-print/code-academic-honesty if you are not familiar with this policy. Inter-personal discussion is encouraged, however, this should be done to understand the material better instead of sharing solutions on individual laboratory assignments. Unless otherwise stated, students are expected to perform all programming assignments, discussions, and exams on an **individual basis**. Any variance from this policy, however minor, will be handled as outlined in the Academic Discipline section of the academic honesty policy, which at the very least awards a zero (0) grade to the affected students for the particular laboratory assignment or exam, but has the potential to take further detrimental punitive actions including expulsion from the course.

Class and TA Logistics
Instructor: Vijay K. Gurbani, Ph.D. <vgurbani@iit.edu>
Office hours: SB 105-C 5pm-6pm; outside of these hours, please contact me for an appointment.
TA: TBD
Lecture: Tue, 6:25pm – 9:05pm Wishnick 113.

Textbooks:
Required: *Introduction to Data Mining*, 1st Edition, Pang-Ning Tan, Michael Steinbach, Vipin Kumar; ISBN-13: 978-0321321367
Optional: (but highly recommend): *R For Everyone*, 1st Edition, Jared P. Lander, Addison Wesley, ISBN-13: 978-0-321-88803-7
Supplementary: Some material may be taken from the following textbooks; there are free online versions of these books available (although, these books are worth having in your bookshelf as a data scientist).

1. *Mining of massive datasets*, 2e, Jure Leskovec, Anand Rajaraman and Jeffrey Ullman, Cambridge University Press, ISBN-13: 978-1-107-07723-2. (Online: http://infolab.stanford.edu/~ullman/mmds/book.pdf)
2. *Data mining and analysis: Fundamental concepts and algorithms*, 1e, Mohammed Zaki and Wagner Meira, Jr., Cambridge University Press, ISBN-13: 978-0521766333. (Online: http://www.dataminingbook.info/pmwiki.php)
3. *An introduction to statistical learning with applications in R*, 7e, Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, Springer, ISBN-13: 978-1-4614-7137-0. (Online: http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf)

Grade distribution (subject to change):

Homework assignments:	30%
Discussion topics:	05%
Midterm exam:	30%
Final exam:	35%

The canonical letter grading scale applies:

A	90 – 100
B	80 – 89
C	70 – 79
D	60 – 69*
E	0 – 59

* (Note well: this letter grade is not applicable to graduate students! For graduate students, 60 and below is an E as per the guidelines of the graduate school.)

The final grade is not curved; for a variety of reasons, I prefer not to curve. At time in the semester you should be able to know where you stand by applying the weights above to your points earned. If you end up on a cusp of a letter grade (defined as 2 percentage points below a letter grade, example, you are at a 0.78 or 0.79 raw score), I will evaluate your semester's worth of work to make a determination if you have demonstrated the work ethic and in-class performance to be bumped up a letter grade. This is done on my discretion given your work ethic and in-class performance, please understand that this is not an automatic upgrade for any student who is on a cusp.

Miscellaneous

If you feel you are falling behind in the class, the time to seek help is *immediately*. Please do not wait, as new concepts and algorithms are introduced on a regular basis and it is imperative that you have mastered the preceding material before moving forward.

ADA statement

Accommodations will be made for students with documented disabilities. In order to receive accommodations, students must intimate the Center for Disability Resources by filling in the form at the following link: https://sites.google.com/iit.edu/cdr-exam-scheduling/home. **It is up to the student to initiate this process with CDR.** The Center for Disability Resources (CDR) is located in Tech South, Room 1C3-2, telephone 312 567.5744 or disabilities@iit.edu.

Course outline

The course outline below is tentative and subject to change as we proceed through the semester. Please make sure you are able to keep up with the assigned reading material and class notes. Not all material in the lectures is drawn from the sources above, however, most of it is. For the material that is not, the class notes will serve as reference.

Week	Topics covered and related logistics (subject to change)	Remarks and notes
1. Jan-09	• Syllabus and expectations • Introduction to data mining (Tan, Ch. 1) • Introduction to R (Lander, Chs. 1, 2.1-2.2, 4, 5, 6)	- Homework 0 assigned (not graded)
2. Jan-16	• Exploring data (Tan, Ch. 3.1, 3.2, 3.3) • Measures of similarity and dissimilarity (Tan, Ch. 2)	- Jan-20: Last day to drop without penalty. - Homework 1 assigned
3. Jan-23	• Linear regression (James, Ch. 3)	
4. Jan-30	• Supervised learning: Classification ○ Decision tree models (Tan, Ch. 4) ○ Performance evaluation of decision tree models	
5. Feb-06	• Classification (continued) ○ Alternative techniques (Tan, Ch. 5.2, 5.3, 5.7.1, 5.7.2, 5.8) ○ Dimensionality reduction: PCA (Tan, Appendix B; Zaki, Ch. 7)	- Homework 1 due Wed, Feb-07 - Homework 2 assigned
6. Feb-13	• Association rules ○ Association analysis: Frequent itemsets and market-basket analysis (Tan, Ch. 6; Zaki, Ch. 8; Leskovec, Ch. 6)	
7. Feb-20	• Association rules (continued)	
8. Feb-27	• Unsupervised learning: Clustering (Tan, Ch. 8; James, Ch. 10; Zaki, Chs. 13-15); Leskovec, Ch. 7 ○ Representative-based (K-Means), hierarchical clustering and density-based (DBSCAN)	- Homework 2 due Wed, Feb-28 - Homework 3 assigned
9. Mar-06	**Midterm exam (upto and including Association rules)**	
10. Mar-13	**Spring break**	
11. Mar-20	• Clustering (continued) ○ Additional issues and algorithms (SOM) (Tan, Ch. 9.2.3)	
12. Mar-27	• Finding similar items (Leskovec, Ch. 3)	- Last day to withdraw is Mar-26
13. Apr-03	• Recommendation systems (Leskovec, Ch. 9) ○ Content-based ○ Collaborative filtering ○ Singular value decomposition (SVD)	- Homework 3 due Wed, Apr-04 - Homework 4 assigned
14. Apr-10	• Mining social network graphs (reading materials will be assigned)	
15. Apr-17	• Link analysis (Leskovec, Ch. 5) ○ PageRank	
16. Apr-24	• Hadoop, MapReduce and Spark Streaming (Leskovec, Ch. 2)	- Homework 4 due Wed, Apr-25
17. May-01	**Final exam (comprehensive)**	