**CS 422: Data Mining**

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

**Fall 2022**

**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 prerequisites 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 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 homeworks will be assigned consisting 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. Of the homework assignments, you can submit two assignments --- but not the last assignment --- late without penalty, but the late assignments MUST be submitted within one week of the original due date, otherwise it will only accrue 30% of the earned points. (We need time to grade and submit the final grades, so the last assignment is exempt from this policy.) If you submit a homework late and want to avail yourself of the late penalty dismissal, you must document this by sending email to me and the TA to let us know **before the homework is due**. If you do not send email in time, the homework will be considered late and penalties outlined previously in this paragraph will apply.

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. Links to excellent primers on R are available in the lecture notes for the first lecture. (Aside: A battle also rages on whether data science should be taught in R or Python; for some thoughts, see https://github.com/matloff/R-vs.-Python-for-Data-Science.)

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** (if you submit any other file formats except the ones listed in (1) below, a penalty of 0.25 points will be levied; if you submit any other file formats except the ones listed in (2-i or 2-ii) below, a penalty of 0.25 points will be levied):

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 firstname-lastname.Rmd, processing it will produce firstname-lastname.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.** (If you only submit one of the files, a penalty of 0.25 points will be levied.)

iii. **Naming your files**: The .Rmd or .R files must be named using the *firstname-lastname*.Rmd or *firstname-lastname*.R format, where firstname and lastname are the first and last names of the student as registered in Illinois Tech.

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.

Note that multiple Homework submissions (2 homeworks for the most part) may be open at any given time, or often, students upload the wrong submission. **It is the responsibility of the student to upload the correct files that correspond to the outstanding homework. Please pay attention to which submission you are uploading the work to. If you upload the wrong submission, and after the due date, you want us to grade the correct submission, please understand that a blanket 30% penalty will apply, i.e., you will be awarded 70% of the accrued points.** (Note that this will be in addition to any late penalty, i.e., if a homework is submitted one day late, but the wrong file was uploaded, a request for re-grading will incur a 40% penalty.) Re-grading homeworks has a cost associated in time and effort, further Blackboard does not cooperate to make it easy to move the homeworks between submissions. So please ensure that you upload to the correct submission page. You can upload your homeworks multiple times before the submission due date, the last uploaded homework will be the one that is graded.

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.

Please be aware that backing up your work is your responsibility. Please do not leave yourself at the mercy of a crashed hard disk wiping away your efforts. Upload your work to the Illinois Tech Google cloud, or invest in a USB drive that can be used to backup important work.

**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**: Thu, 5:15pm-6:15pm, SB 105-C (in person); if you want to meet me outside of the office hours, please send me email and we will arrange to meet on Zoom. You can meet me in my office during regular office hours (if you are comfortable meeting in person during Covid). No appointments are needed for my regular office hours.

**TA 1**: Pinakin Nimavat <pnimavat@hawk.iit.edu>. Office hours: Tu, Th 3pm-4pm.

**TA 2**: Hua Xu <hxu40@hawk.iit.edu>. Office hours: Tu 2pm-3pm, Wed 1pm-2pm.

**Live Lecture**: Thu, 6:25pm – 9:05pm PS 111. For online sections, recorded lecture will be available 2-3 hours later on Blackboard (See link "Class Recordings | Panopto"). During the semester, recitations will be held as needed; the time for the recitations will be announced in advance, and the recitations will be recorded for asynchronous viewing. **Attendance is mandatory for all sessions!**

**Textbooks**:

**Required**: *Introduction to Data Mining*, 2nd Edition, Pang-Ning Tan, Michael Steinbach, Vipin Kumar; ISBN-13: 978-0133128901

**Optional**: (but highly recommend): *R For Everyone*, 2nd Edition, Jared P. Lander, Addison Wesley, ISBN-13: 978-0134546926

**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*, 2e, Mohammed Zaki and Wagner Meira, Jr., Cambridge University Press, ISBN-13: 978-0521766333. (Online: https://dataminingbook.info/book_html/)
3. *An introduction to statistical learning with applications in R*, 2e, Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, Springer, ISBN-13: 978-1-4614-7137-0. (Online: https://www.statlearning.com/)

**Grade distribution** (subject to change):

| | |
|---|---|
| Homework assignments: | 30% |
| Discussion topics: | 5% |
| Midterm exam: | 30% |
| Final exam: | 34% |
| Attendance: | 1% |

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! Graduate students scoring 69 and below will be awarded 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 1 percentage points below a letter grade, example, you are at a 0.79 raw score), I will evaluate other factors (class attendance, quality of discussions, class participation, submission timeliness, etc.) to bump you up to the next higher 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 such 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.

**Policy on double-registrations**

Students **MUST NOT** register for another online class that meets at the same time as CS422, nor should they register for another live class in the same period and hope to keep up with CS422 through the online videos. The reason is that double-registration of this form create final examination conflicts as both CS422 and the conflicting class will hold finals on the same date and time. This simply create an undue burden on all the involved parties. As such, requests for holding final examination out of the normally allocated time for CS422 will not be entertained. If despite warnings students insist on

such double-registration, it is up to the students to work it out with professor teaching the conflicting class for alternate accommodations to take the conflicting final exam.

**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 is tentative and subject to change as we proceed through the semester. The due dates of assignments may change to accommodate the subject coverage in the class. So PLEASE BE AWARE.**

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. Aug-25 | • Syllabus and expectations<br>• Introduction to data mining (Tan, Ch. 1)<br>• Introduction to R (Lander, Chs. 1, 2.1-2.2, 4, 5, 6) | |
| 2. Sep-01 | • Exploring Data (Tan, Ch. 2.1, 2.2, 2.3) | - Sep-03 day to add/ drop without fee. |
| 3. Sep-08 | • Linear regression (James, Ch. 3) | - Sep-09 last day for late registration |
| 4. Sep-15 | • Components of learning<br>• Supervised learning: Classification<br>◦ Decision tree models (Tan, Chs. 3.1 – 3.3, 3.5, 3.6) | |
| 5. Sep-22 | • Classification (continued)<br>◦ Performance evaluation of decision tree models (Tan, Ch. 3.4, Ch. 3.6, Ch. 5.7.2)<br>◦ Alternative techniques: (Tan, Chs. 4.10, 4.11, 4.12) | |
| 6. Sep-29 | • Classification (continued) | |
| 7. Oct-06 | • Classification (continued)<br>• Association rules (Zaki, Chs. 8.1, 8.2, 8.2.1, 8.2.3, 8.3, 12.1, and Tan, Chs. 4.10, 4.11, 4.12) | |
| 8. Oct-13 | **Midterm exam (upto and including Association rules)** | |
| 9. Oct-20 | • Association rules (continued)<br>• Neural networks: The perceptron model (Tan, Ch. 4.7.1) | |
| 10. Oct-27 | • Neural networks: The perceptron model (continued)<br>• Neural networks (continued) (Tan, Ch. 4.7.2-4.7.3)<br>◦ Multilayer perceptrons<br>◦ Forward propagation | - Oct-31: Last day to withdraw from classes. |
| 11. Nov-03 | • Neural networks (continued)<br>◦ Back propagation<br>◦ Gradient Descent<br>• Bias Variance Tradeoff (Tan, Ch. 4.10.3, Zaki, Ch. 22.3, pg. 570-574)<br>• Scaling and Standardizing variables (Class notes) | |
| 12. Nov-10 | • Unsupervised learning:<br>◦ Clustering I (Tan, Ch. 7; James, Ch. 10; Zaki, (Chs. 13-15); Leskovec, Ch. 7<br>◦ Representative-based (K-Means) | |
| 13. Nov-17 | • Clustering II (continued).<br>◦ Hierarchical clustering, density-based clustering and cluster evaluation<br>◦ Principal component analysis (PCA. James, Chs. 6.3.1, 12.2, 12.2.1, 12.2.2; for a more theoretical approach, see Zaki, Ch.7) | |
| 14. Nov-24 | Thanksgiving break – No class | |
| 15. Dec-01 | • Recommendation systems (Leskovec, Ch. 9)<br>◦ Content-based<br>◦ Collaborative filtering (Time permitting)<br>• Mining social network graphs (reading materials will be assigned) | - Dec-04: Last day to request an incomplete grade. |
| 16. Dec-08 | **Final exam (comprehensive)** | |