MSCI446 – Data Mining and Warehousing - Fall 2018

Instructor: Lukasz Golab (lgolab@uwaterloo.ca)
Office hours: Tuesdays and Thursdays 4-4:30 in CPH4350

TA: Hannah Gautreau (hvgautreau@uwaterloo.ca)
Office hours: Fridays 3:30-4:30 in CPH4359 (data science lab)

Web: learn.uwaterloo.ca (announcements, handouts, discussion forum, etc.)

NOTE: We will be posting important announcements on Learn. You are responsible for monitoring the course website regularly.

Description: This course will present state-of-the-art practice and research in the storage, extraction, manipulation and analysis of data, with a view to using these processes for making better management decisions. Topics include: extracting, cleaning, and organizing data from transactional databases, discovering and validating patterns and relationships using statistical techniques, and using the extracted patterns for making improved management decisions.

Objectives: 1) Explain how data mining algorithms work [KB]*
2) Apply data mining algorithms to solve real(istic) problems [PA]*
** abbreviations in brackets are Graduate Attributes required for Accreditation; see appendix for explanation and descriptions*

Evaluation: **Midterm – 30%** (2:30-3:50 on Thursday, October 25, room TBA)
Final exam – 40% (scheduled by the registrar)
Project – 30%

In order to pass the course, the weighted average of your midterm and final exam grade must be at least 60%, regardless of your project grade.

There will be no make-up midterm. If you miss the midterm due to serious illness (and submit an acceptable doctor's note), your final exam will be worth 70% of your final grade.

Lecture notes: datascienceguide.github.io

Textbooks: **WF:** Data Mining: Practical Machine Learning Tools and Techniques, 3rd Ed, Witten & Frank
(slides at www.cs.waikato.ac.nz/ml/weka/book.html)

TSK: Introduction to Data Mining, Tan, Steinbach & Kumar
(slides at www-users.cs.umn.edu/~kumar/dmbook/index.php)

Other material: homes.cs.washington.edu/~pedrod/papers/cacm12.pdf
(a nice article on machine learning; we'll discuss it in class near the end of the course)

Prerequisites: 1) Computing (programming, algorithms, data structures)
2) Statistics/probability (probability functions, conditional probabilities, Bayes' theorem)
3) Optimization (objective functions, constraints, heuristics, local/global optima)
4) Database systems (SQL, relational model, keys/foreign keys)

Topics Covered + Relevant Slides/Links

1. Introduction to data mining

   WF: Chapter 1 slides 5-7, 9, 13-15, 17, 23-24, 41-42; Chapter 2 slides 3-8

   TSK: Chapter 1 slides 5-6, 8-11, 14, 17-19, 23-25, 29

   http://datascienceguide.github.io/data-science-framework

2. Data exploration and visualization

   WF: Chapter 2 slides 18-20, 23, 32-35

   TSK: Chapter 2 slides 2-3, 9; Chapter 3 slides 5-9, 15, 17-18

   http://datascienceguide.github.io/exploratory-data-analysis

3. Numeric prediction/Linear regression

   WF: Chapter 3 slides 5-6

   http://datascienceguide.github.io/regression

4. Bayesian inference/the Naive Bayes Classifier

   WF: Chapter 4 slides 13-17, 20-23, 27

   TSK: Chapter 5 slides 49-54, 56-57, 59

   http://datascienceguide.github.io/association-rule-mining

   Also see slides 1-4 at gicl.cs.drexel.edu/images/2/21/Regli-casino-example.pdf

5. Decision Trees

   WF: Chapter 4 slides 28-38

   TSK: Chapter 4 slides 6-15

   http://datascienceguide.github.io/information-entropy

   http://datascienceguide.github.io/decision-trees

6. Rule mining/the 1R algorithm/the Prism algorithm

   WF: Chapter 3 slides 17-21; Chapter 4 slides 4-6 and 50-58

   http://datascienceguide.github.io/rule-based-learning

7. Evaluating the effectiveness of classification algorithms/Cross-validation

   WF: Chapter 5 slides 5-8, 15-17, 19, 34-35, 38-40, 56-57

   TSK: Chapter 4 slides 74-77

   http://datascienceguide.github.io/model-evaluation

   http://datascienceguide.github.io/cross-validation

8. Association rule mining/the Apriori algorithm

   WF: Chapter 3 slides 25-26; Chapter 4 slides 61-68

   TSK: Chapter 6 slides 2-15

   http://datascienceguide.github.io/association-rule-mining

9. Distance metrics/Clustering/the k-Means algorithm

   WF: Chapter 4 slides 103-104

   TSK: Chapter 2 slides 49-53, 58-60; Chapter 8 slides 2-5, 20-22, 25

   http://datascienceguide.github.io/distance-measurements

   http://datascienceguide.github.io/clustering

10. Nearest neighbour classification

    WF: Chapter 4 slides 90, 92, 101

    TSK: Chapter 5 slides 37-39, 41-42, 45

    http://datascienceguide.github.io/k-nearest-neighbor

11. Logistic regression

    WF: Chapter 4 slides 76-79

    http://datascienceguide.github.io/logistic-regression

12. Linear models for classification/Support vector machines

    WF: Chapter 3 slides 7-8

    TSK: Chapter 5 slides 66-71

    http://datascienceguide.github.io/support-vector-machine

13. Outlier detection

    TSK: Chapter 10 slides 4-6, 14

14. Graph mining

    http://datascienceguide.github.io/graph-mining

15. Advanced topics

    TBA

Mental Health and Wellness

If at any time you are experiencing mental health concerns or personal circumstances that are impacting your ability to succeed in your studies, please reach out to any member of the teaching team.

There are a number of ways to seek help on campus free of charge:

- **Engineering Counselling:** https://uwaterloo.ca/engineering/current-undergraduate-students/engineering-counselling
- **Counselling Services:** https://uwaterloo.ca/campus-wellness/counselling-services
- **UW MATES**: (Mentor Assistance Through Education and Support) is a peer to peer counseling program run through counselling services: https://uwaterloo.ca/campus-wellness/counselling-services/uw-mates-peer-support

There are also a number of community resources that can be accessed free of charge:

- **Self Help Alliance:** (519) 570-4595
- **Good to Talk:** 1-866-925-5454 (available 24/7)
- **Here 247:** 1 844 437 3247 (available 24/7)

If you are experiencing a mental health emergency, you should go to Counselling Services or your nearest Emergency Room if it is after hours. The closest Emergency Department to the University of Waterloo is at Grand River Hospital. If you are unsafe and you cannot get to the Emergency Department, call 911.

Academic Policies

Academic integrity: In order to maintain a culture of academic integrity, members of the University of Waterloo community are expected to promote honesty, trust, fairness, respect and responsibility. [Check the Office of Academic Integrity for more information.]

Grievance: A student who believes that a decision affecting some aspect of his/her university life has been unfair or unreasonable may have grounds for initiating a grievance. Read Policy 70, Student Petitions and Grievances, Section 4. When in doubt, please be certain to contact the department's administrative assistant who will provide further assistance.

Discipline: A student is expected to know what constitutes academic integrity to avoid committing an academic offence, and to take responsibility for his/her actions. [Check the Office of Academic Integrity for more information.] A student who is unsure whether an action constitutes an offence, or who needs help in learning how to avoid offences (e.g., plagiarism, cheating) or about "rules" for group work/collaboration should seek guidance from the course instructor, academic advisor, or the undergraduate associate dean. For information on categories of offences and types of penalties, students should refer to Policy 71, Student Discipline. For typical penalties, check Guidelines for the Assessment of Penalties.

Appeals: A decision made or penalty imposed under Policy 70, Student Petitions and Grievances (other than a petition) or Policy 71, Student Discipline may be appealed if there is a ground. A student who believes he/she has a ground for an appeal should refer to Policy 72, Student Appeals.

Note for students with disabilities: AccessAbility Services, located in Needles Hall, Room 1132, collaborates with all academic departments to arrange appropriate accommodations for students with disabilities without compromising the academic integrity of the curriculum. If you require academic accommodations to lessen the impact of your disability, please register with AccessAbility Services at the beginning of each academic term.

Turnitin.com: Text matching software (Turnitin®) may be used to screen assignments in this course. Turnitin® is used to verify that all materials and sources in assignments are documented. Students' submissions are stored on a U.S. server, therefore students must be given an alternative (e.g., scaffolded assignment or annotated bibliography), if they are concerned about their privacy and/or security. Students will be given due notice, in the first week of the term and/or at the time assignment details are provided, about arrangements and alternatives for the use of Turnitin in this course.
It is the responsibility of the student to notify the instructor if they, in the first week of term or at the time assignment details are provided, wish to submit alternate assignment.

Appendix

All engineering programs are reviewed by the Canadian Engineering Accreditation Board (CEAB). One of the required accreditation criteria is that institutions ensure students have sufficient knowledge and proficiency with respect to the 12 Graduate Attributes (GAs) listed below. These attributes are mapped to the learning objectives in each course for assessment, as shown in the brackets. This allows the program to both comply with CEAB requirements and continuously improve.

#	Acro-nym	Attribute Name	Attribute Definition
1	KB	Knowledge Base	Demonstrated competence in university level mathematics, natural sciences, engineering fundamentals, and specialized engineering knowledge appropriate to the program.
2	PA	Problem analysis	An ability to use appropriate knowledge and skills to identify, formulate, analyze, and solve complex engineering problems in order to reach substantiated conclusions.
3	Inv	Investigation	An ability to conduct investigations of complex problems by methods that include appropriate experiments, analysis and interpretation of data, and synthesis of information in order to reach valid conclusions.
4	Des	Design	An ability to design solutions for complex, open-ended engineering problems and to design systems, components or processes that meet specified needs with appropriate attention to health and safety risks, applicable standards, and economic, environmental, cultural and societal considerations.
5	Tools	Use of Engineering Tools	An ability to create, select, apply, adapt, and extend appropriate techniques, resources, and modern engineering tools to a range of engineering activities, from simple to complex, with an understanding of the associated limitations.
6	Team	Individual and team work	An ability to work effectively as a member and leader in teams, preferably in a multi-disciplinary setting.
7	Comm	Communication skills	An ability to communicate complex engineering concepts within the profession and with society at large. Such ability includes reading, writing, speaking and listening, and the ability to comprehend and write effective reports and design documentation, and to give and effectively respond to clear instructions.
8	Prof	Professionalism	An understanding of the roles and responsibilities of the professional engineer in society, especially the primary role of protection of the public and the public interest.
9	Impact	Impact of engineering	An ability to analyze social and environmental aspects of engineering activities. Such ability includes an understanding of the interactions that engineering has with the economic, social, health, safety, legal, and cultural aspects of society, the uncertainties in the prediction of such interactions; and the concepts of sustainable design and development and environmental stewardship.
10	Ethics	Ethics and equity	An ability to apply professional ethics, accountability, and equity.
11	Econ	Economics and project management	An ability to appropriately incorporate economics and business practices including project, risk, and change management into the practice of engineering and to understand their limitations.
12	LL	Life-long learning	An ability to identify and to address their own educational needs in a changing world in ways sufficient to maintain their competence and to allow them to contribute to the advancement of knowledge.