Principles of Knowledge Mining
CSI-777
Syllabus

Time: Thursdays, 4:30-7:10pm (on-line)
Classroom: Mason Global Center, room 1320C
Instructor: William G. Kennedy, PhD, Associate Professor
Office Phone: 703-993-9291
e-mail: wkennedy@gmu.edu
Office hours: Thursdays, 3-4pm, Research Hall, room 378

Course Description: Principles and methods for synthesizing task-oriented knowledge from computer data and prior knowledge and presenting it in human-oriented forms such as symbolic descriptions, natural language-like representations, and graphical forms. Topics include fundamental concepts of knowledge mining; methods for target data generation and optimization; statistical and symbolic approaches; knowledge representation and visualization; and new developments such as inductive databases, knowledge generation languages, and knowledge scouts.

Objectives:
1. Students can discuss the theory and tools of knowledge mining.
2. Students are able to perform knowledge mining of publicly available datasets.
3. Students understand issues associated with knowledge mining and presentation.

University Policies: The University Catalog, http://catalog.gmu.edu, is the primary resource for university policies affecting student and faculty conduct in university affairs.

Attendance Policy: Attendance is not graded, but most of the readings will be discussed in class each week and project will be presented to the class. Therefore, attendance is expected.

Office of Disability Services: If you are a student with disability and you need academic accommodations, please see me and contact the Disability Resource Center (DRC) at 709-993-2474. All academic accommodations must be arranged through that office.

Class communications: Mason uses e-mail to provide official information to students. Examples include communications from course instructors, notices form the library, notices about academic standing, financial aid information, class materials, assignments, questions, and instructor feedback. Students are responsible for the content of university communication sent to their Mason e-mail account, and are required to activate that account and check it regularly. I intend to respond to all student e-mails within a couple of hours of receipt and always within 24 hrs. I have official office hours during which I will be available for drop-in discussions. Other meetings outside class are possible but should be scheduled in advanced. I will also maintain a website with class materials throughout the course. Its address will be provided in the first class.

Academic Integrity: Mason is an Honor Code university; please see the University Catalog for a full description of the process. The principle of academic integrity is taken very seriously and violations are treated gravely. Academic integrity means when you are responsible for a task you perform that task. When you rely on someone else's work, text, or code, even if in the public domain, in any aspect of the performance of that task, you must cite the source in the proper, accepted form. Another aspect of academic integrity is the free play of ideas. Vigorous discussion and debate are encouraged in this course, with the firm expectation that all aspects of the class will be conducted with civility and respect for differing ideas, perspectives, and traditions. When in doubt (of any kind), please ask for guidance and clarification. As instructor for this course, I must reserve the right to enter a failing grade to any student found guilty of an honor code violation.

Late submission of class work: Homework is due at the beginning of next class. Lateness reduces the possible graded points at a rate of approximately 20% per day.

Evaluation:
Reviews of readings: 35%
Students are expected to write a short review (400-800 words) of selected readings identifying the contribution, strengths, and weaknesses of the reading. I have identified the seven (7) readings by * in class schedule. Each will be worth 5 pts.

Knowledge mining exercises: 40%
We will work through five (5) exercises of the reference text. Students are encouraged to consult each other on the exercises, but each student is expected to submit their own work, evidence of successful completion, and comments on the task (100-200 words) at the beginning of the next class. The knowledge mining exercises will be worth 8 points each.

Knowledge Mining project: 25%
The knowledge mining modeling project is intended to have students apply their knowledge of the subject by developing an analysis of a large dataset their choosing. Projects will be done individually or in teams of no more than 2 and presented to the class near the end of the classes. Students will propose a project (5 pts) and the instructor will provide feedback on scope and projected level of difficulty. Presented projects will be graded demonstrated knowledge mining operation (7 pts), usefulness of patterns reported (7 pts), and explanation of results, (6 pts).

Grading scale: (points = percentage)

94-100 = A	85-89 = B+	76-79 = B-	<69 = F
90-93 = A-	80-84 = B	70-75 = C	

Required Text: none.

Recommended Text: Witten, Frank, Hall, and Pal, ***Data Mining: Practical Machine Learning Tools and Techniques*****. 4th Ed.** Morgan Kaufmann.

Class Plan (subject to adjustment)

Class 1: (27Aug)	<u>In class</u>:	Introduction: models, architectures, cognitive plausibility, AI, Cognitive Science, Computational Social Science
	<u>After class</u>:	(due by the start of the next class): Read: (Rowley, 2006)*, text chap. 1 Do: install Anaconda-Navigator, R/R-Studio, =or= WEKA system
Class 2: (3Sep)	<u>In class</u>:	Data Mining Sources & Concepts
	<u>After class</u>:	Read: text chap. 2* **Ex1**: Load the specified weather dataset and remove all instance with high humidity. Submit resulting dataset.
Class 3: (10Sep)	<u>In class</u>:	Representation of Knowledge Output
	<u>After class</u>:	Read: text chap. 3*
Class 4: (17Sep)	<u>In class</u>:	Basic Algorithms I: Rules & Trees
	<u>After class</u>:	Read: text chap. 4.1-4.5 **Ex2**: For provided glass dataset, plot histograms for numeric attributes. For provided iris data, create discretized intervals with consistent distributions.
Class 5: (24Sep)	<u>In class</u>:	Basic Algorithms II: Linear Models, Instance-Based Learning, and Clustering
	<u>After class</u>:	Read: text chap. 4.6-4.9
Class 6: (1Oct)	<u>In class</u>:	Evaluating Mined Knowledge
	<u>After class</u>:	Read: text chap. 5* **Ex3**: Using the glass dataset, run a k-nearest-neighbors classifier for k=1..5 and evaluate each with a 10 fold cross-validation. Discuss the accuracy of each.
Class 7: (8Oct)	<u>In class</u>:	Adv. Approaches I: Trees & Rules
	<u>After class</u>:	Read: text chap. 6* **Ex4:** Do one of the following (A or B): A: Using the iris data, evaluate C4.5 using the full training set and 10 fold cross-validation. Which is more realistic and why? B: Using the weather data, generate and count all the rules for combinations of minimum confidence 0.7-0.9 and support 0.1-0.3.
Class 8:(15Oct)	<u>In class</u>:	Adv. Approaches II: Instance-based & Linear Model
	<u>After class</u>:	Read: text chap. 7*

Class 9:(22Oct) In class: Data Transformations & Research Project Management
After class: Read: text chap. 8
Submit project proposal

Class 10:(29Oct) In class: Probabilistic Approaches
After class: Read: text chap. 9
Do: knowledge mining project
Ex5: For provided diabetes dataset, measure the performance of the standard naïve Bayesian classifier using cross-validation. What do the results indicate?

Class 11:(5Nov) In class: Ensemble, Semisupervised, & Multi-instance Learning
After class: Read: text chap. 10
Johnstone and Titterington, 2009 &
Wegman, 2003*
Do: knowledge mining project

Class 12:(12Nov) In class: Visualizing High Dimensional Data
After class: Read: (Crooks, et al., 2013)
Do: knowledge mining project

Class 13:(19Nov) In class: Social Data Mining (Dr. Crooks)
After class: Do: knowledge mining project

THANKSGIVING BREAK (25-27 Nov)

Class 14:(3Dec) In class: Project presentations
After class: Do: knowledge mining project

Exam: (10Dec) Exam time (if needed): Project presentations

References:

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, *17*(1), 124-147.

Johnstone, I. M., & Titterington, D. M. (2009). Statistical challenges of high-dimensional data.

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of information science*, *33*(2), 163-180.

Wegman, E. J. (2003). Visual data mining. *Statistics in medicine*, *22*(9), 1383-1397.

Witten, Frank, Hall, and Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. 4th Ed. Morgan Kaufmann.

