CS 6604: Data Mining/Fall 2007

Data mining has emerged as one of the most exciting and dynamic fields in computer science. The driving force for data mining is the presence of petabyte-scale online archives that potentially contain valuable bits of information hidden in them. Commercial enterprises have been quick to recognize the value of this concept; consequently, within the span of a few years, the software market for data mining has expanded to be in excess of tens of billions of dollars. CS 6604 is designed to provide graduate students with a broad background in the design and use of data mining algorithms, exposure to software tools, and specialized expertise in applying these ideas to a real-life situation. Case studies will be provided using practical examples of data mining systems.

Meeting Times	MW 4:00-5:15pm, McB 219
Instructor	Naren Ramakrishnan, 1-8451, 2160L Torg
	naren@cs.vt.edu, http://www.cs.vt.edu/∼ramakris
Office Hours	Mondays, Wednesdays 10am-12pm,
	or walk in any time.
Teaching Assistant	None assigned.
Listserv	CS6604_91720@listserv.vt.edu
	(yes, the name is rather long winded.)
Course Web Page	http://people.cs.vt.edu/∼ramakris/Courses/CS6604/

If you need adaptations or accommodations because of a disability (learning disability, attention deficit disorder, psychological, or physical), if you have emergency medical information to share with the instructor, or if you need special arrangements in case the building must be evacuated, please meet with the instructor ASAP.

Pre-requisites: There are really no formal pre-requisites beyond graduate student standing. You are expected to have basic knowledge of probability, statistics, must have taken undergraduate courses in CS theory (algorithms, NP-completeness), database systems, and must not be averse to math (e.g., numerical methods). Knowledge of propositional and predicate logic or experience with PROLOG will be beneficial. One or more of CS 5114 (algorithms), CS 5485 (numerical analysis), CS 5604 (information retrieval), CS 4604/5614 (database systems) would thus be beneficial. Background in basic statistics (one or more of STAT 5104, STAT 5114, STAT 5615-6) is recommended.

Evaluation: There will be 12 homeworks, which will involve a mix of theoretical problems, programming assignments, literature surveys, and questions that will focus on your reasoning and problem solving skills. The topical content of the programming assignments will usually be language-independent, so you are free to use your favorite platform/language. No late submissions will be accepted. Each homework will be due in 1-1.5 weeks. If you have a homework that you feel has been graded incorrectly, please contact the instructor, and we can discuss a re-grading if appropriate. There are no exams.

Keeping in Touch: Please use the listserv actively for discussions and exchanging ideas. Since it is created automatically by a central university system, any student registered in CS 6604 will be added to the mailing list. If you do not receive a test mail from the instructor by the end of the first week of classes, ensure that your email address is properly recorded in the university system.

Workload: The course moves at a very fast pace! I assume that this is your *only* graduate course this semester. You are expected to put in 2-3 sleepless nights per week (I did, when I was a graduate student). Most assignments involve a fair amount of design, so plan your schedule accordingly.

Book: There is really no textbook that covers all relevant themes in the required detail. Being graduate students, you are expected to be able to cull ideas from research papers and form your own conceptual model of what modern data mining is about. In short, no babysitting or spoonfeeding. If you are the type of person who cannot live without the planned layout and schematics of a textbook, you are advised to drop this course. Stay tuned to the course website for reading material, tutorials, and useful links.

Syllabus: The approximate sequence and topics will be as follows.

- **Introduction:** Models, methodologies, and processes. The KDD process. Generic tasks. Broad themes (search, induction, querying, approximation, and compression). Application areas. The good, bad, and ugly of data mining practice: data dredging, data fishing, and data scrubbing.
- **Discrete Structures:** Itemset mining. Concept lattices. Borders and levelwise theories. Condensed representations. Frequent pattern mining. Redescription mining. Graphs and other structures. Combinatorial tiles. Customized data structures for speeding up data mining algorithms.
- **Attribute-Value Learning Techniques:** Decision trees. Decision lists. Classification and regression trees. Association rules. Correlations. Rule-based mining. Sequential versus simultaneous paradigms.
- **Relational Mining Techniques:** Inductive logic programming. Main approaches to ILP. Rule induction, beam search, logical decision trees, clausal discovery. Inverse resolution, relative least general generalization. Propositionalization techniques. Operators for efficient search of relational spaces. Learning from interpretations. Comparative merits of attribute-value and relational mining techniques. Domain theories and incorporating prior background knowledge.
- **Probabilistic Techniques:** Conditional independence and its modeling. Inference and representational complexity. Bayesian networks. Connections between probabilistic (generative) and enumerative data mining paradigms. Probabilistic models for query approximation.
- **Sequences and order:** Total and partial orders. Episodes and event streams. Frequent episode mining. Order-theoretic methods. Modeling sequential data. Discovering sequence information from non-sequential data. Connections with HMMs.
- **Putting it all together:** Biclustering. Compositional data mining. Mining chains of relations. Integrated query/mining languages. Paradigms for interfacing with database systems.
- **Applications:** Data mining applications in bioinformatics, personalization, information retrieval, web modeling, filtering, and text processing.