CISC 483/683: Data Mining

INSTRUCTOR: Sandra Carberry
OFFICE: Room 439 Smith
NETMAIL: carberry@udel.edu
CLASS WEB SITE: www.cis.udel.edu/∼carberry/CISC-483-683

OFFICE HOURS: Mon. 1:45pm-3:00pm
Wed. 8:00am-9:00am

TA: Haoran Wei
OFFICE: 203 Smith
NETMAIL: nancywhr@udel.edu

OFFICE HOURS Tues. 1:00pm-2:00pm
Fri. 2:00pm-3:00pm

1 Course Description

Data Mining attempts to identify interesting structural patterns in large data sets that can be used to make future predictions. For example, in the area of security, one might analyze a database of past credit card transactions to hypothesize what sequences are indicative of normal credit card use, and then reject credit card transactions that do not match this pattern. In the area of medical diagnosis, one might analyze patient histories to determine which patients are most likely to benefit from an expensive procedure. In the life science area, molecular biologists might analyze large sets of biological data to predict protein structure. In the area of consumer marketing, one might analyze supermarket data to determine what items are typically purchased with other items, and then display those items together to encourage more customers to purchase both items. And in the area of investment and finance, one might analyze economic data to identify stock market trends. Data mining is becoming increasingly important in many environments; a few of these include bioinformatics, advertising, banking, business, finance, security, medicine, and web page design, but there are many others.

This course will introduce fundamental strategies and methodologies for data mining along with the concepts underlying them, and will provide hands-on experience with a variety of different techniques. Students will learn to use the Weka workbench, a set of data mining tools.

2 Prerequisites

Prereq: CISC-220 (data structures) and at least one upper-level course in computer science, or permission of instructor.

3 Textbooks

TSK: *Introduction to Data Mining* **by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar**

WFH: *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)* **by Ian Witten, Eibe Frank, and Mark Hall. (On reserve in the library for your use.)**

4 Weka

We will be using the Weka toolkit. Please go to the Weka web site immediately at
http://www.cs.waikato.ac.nz/ml/weka
and download and install the Weka software on your laptop or PC; you should download the stable version for the 3rd edition of the Witten data mining textbook (weka-3-8-1). If you do not own a laptop or PC, please let the instructor know right away so that alternative arrangements can be made.

5 Handouts

There will be two kinds of handouts in the course:

1. Homework assignments: if you miss class, you can get an extra copy of a homework assignment from the instructor or Sakai.

2. Outlines of lectures, definitions, examples, algorithms: These handouts will often be used to help with lecture presentations and reduce the amount of note-taking effort for students. Since they are intended to serve as partial class notes and keeping track of them would be very difficult, they will **ONLY** be available in class. I will not save extra handouts from class — if you miss class, copies of these handouts will **NOT** be available.

6 Cell Phones, Laptops, and other Electronic Devices

Please turn off your cell phones before class begins, and please do not use laptops or other electronic devices except for group class activities.

7 Grading

ITEM	PERCENT OF GRADE
Homework assignments	40%
Midterms	30%
Final Exam	30%
Class Participation	described below

Exams

There will be two midterm exams and a final exam; they must be taken in class on the designated date. Makeups will not be given except in exceptionally extenuating circumstances such as hospitalization. If there is a date that you would like me to work around in scheduling an exam, **you must** let me know by email before Sept. 9 — this includes dates when you will miss class due to attending a conference. The second midterm is planned for Friday, Nov. 10. **Please make sure that you will be in class that day.**

Homework Sets

Homework assignments are intended to give you an opportunity to work with the concepts discussed in class. These will include both calculations by hand and projects using the Weka toolkit.

- Homework sets are due before class starts on the announced due date, and will be collected at that time. Once the homework has been collected and class begins, any homework sets turned in will be regarded as late.
- There is a grace period during which late homeworks will not be penalized. Students utilizing the grace period must slide their late homework under the instructor's office door (439 Smith) **at least 30 minutes prior** to the end of the grace period or put it in the instructor's mailbox (in the hallway along the 101 Smith corridor) **at least 30 minutes prior** to the end of the grace period, or give it directly to the instructor prior to the end of the grace period. The grace period for late homeworks is as follows:

DUE-DATE	**END-OF-GRACE-PERIOD**
12:20pm Monday	12:20pm on the following Wednesday
12:20pm Wednesday	12:20pm on the following Friday
12:20pm Friday	12:20pm on the following Monday

  After the end of the grace period, late homework sets will be penalized 10% of the total points that the assignment is worth for the first day that the assignment is late (not including Saturday and Sunday) and 25% of the total points for each additional day late.
- When turning in an assignment late, you should mark the current date and time on the first page. If the date or time is falsified, the penalty will be doubled.
- **All work must be done independently.** You may consult with others about conceptual problems with assignments (unless it is explicitly forbidden for a particular assignment) and for help with Weka. But collaboration beyond this is not allowed. Please keep in mind that homework solutions downloaded from the web are **NOT** your own. Downloading answers to homework or assignments is plagiarism and is strictly forbidden according to the University's Code of Conduct.
- In the case of questions regarding the grading of homework assignments, you should first contact the teaching assistant. If you still have questions after meeting with the teaching assistant, contact the instructor.

Class Participation: Class participation is strongly encouraged and leads to a much more enjoyable and productive class. So please actively contribute to the class discussions and feel free to ask questions — I very much want to help you master data mining techniques and get as much from the course as possible. Particularly good class contributions will positively affect borderline decisions on final grades in the course. Disruptive or distracting behavior hurts the whole class; such behavior will result in a reduction of up to two letter grades (for example, from "A-" to "C-") in the student's final grade in the course.

Lectures and Readings

The following reading list outlines the topics that we will study in the course. This may be modified as the semester progresses. Readings preceded by a * relate to Weka; you only need to read what is pertinent to the methods we are studying.

Topic	Reading: CIS-483-683
Introduction	
What is data mining?	TSK: pp.1-14
Preliminaries	
Data and Attributes	TSK: pp.19-44
Data Preprocessing	TSK: pp.44-57
Aggregation, Sampling	
Dimensionality reduction, Feature selection	
Decision Tree Classification	
Constructing Decision Trees	TSK: pp. 145-172; *WFH: pp.445-451, pp.454-457
Evaluation	TSK: pp. 186-188
Discretizing Numeric Attributes	TSK: pp.57-62; *WFH: pp.432-445
Using Weka	*WFH: pp.51-60, 407-427
Evaluation: generalization error, confidence interval	TSK: pp.179-184, pp.189-190
Model Overfitting	TSK: pp.172-179
Pruning Decision Trees	TSK: pp. 184-186
Naive Bayes Classification	
Introduction	TSK: pp.227-231
Naive Bayes Classifier	TSK: pp.231-240
Evaluation: error estimates and class imbalance problem	TSK: pp.294-301
Accounting for cost	TSK: pp.302-304; *WFH: pp.477
Evaluation: comparing classifiers	TSK: pp. 188-193
Numeric Prediction	
Linear regression	WFH, pp.124-125; *WFH: pp.459-469
Logistic regression	WFH: pp.125-127
Regression and model trees	WFH: pp.251-261
Evaluation: error rate	WFH: pp.180-182
Rule-based methods	
Classification rules:	TSK: pp.207-223, *457-459
Association rules: construction	TSK: pp.327-353, *WFH: pp.485-487
Association rules: handling numeric attributes	TSK: pp.415-426
Association rules: evaluation	TSK: pp.370-386
Lazy Learning	
Nearest neighbor algorithms	TSK: pp.223-227, *WFH: pp.472
Efficiency: KD Trees	WFH: pp.132-138
Clustering	
Cluster analysis:	TSK: pp.487-496, *WFH, pp.480-485
K-Means algorithm	TSK: pp.496-513, pp.69-76, pp.83-84
Density based clustering	TSK: pp.526-532
Hierarchical clustering	TSK: pp.515-526
Cluster evaluation	TSK: pp.532-546, pp.548-555
Fuzzy Clustering	TSK: pp.577-582
Statistical clustering:	TSK: pp. 577, pp.583-593
Detecting Anomalies	
Approaches	TSK: pp.651-675
Ensemble Methods (?)	
Introduction	TSK: pp.276-283
Bagging	TSK: pp.283-285, *WFH: pp.474-476
Boosting	TSK: pp.285-290, *WFH: pp.476-477
Option trees	WFH: pp.365-368
Other topics (?)	
Association rules: sequential patterns	TSK: pp.429-441
Association rules: subgraph patterns	TSK: pp.442-457