

# CISC 483/683: Data Mining

INSTRUCTOR: Sandra Carberry  
OFFICE: Room 448 Smith  
NETMAIL: carberry@cis.udel.edu  
CLASS WEB SITE: [www.cis.udel.edu/~carberry/CISC-483-683](http://www.cis.udel.edu/~carberry/CISC-483-683)

OFFICE HOURS: Mon. 8:30am-9:30am  
Thurs. 12:30pm-1:30pm

TA: Peng Wu  
OFFICE: 102 Smith

OFFICE HOURS: ???

## 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 predict what sequences are indicative of fraudulent credit card use, and then reject credit card transactions that 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 area of business, 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. Data mining is becoming increasingly important in many environments; a few of these include advertising, banking, bioinformatics, business, 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

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

Course project number:

## 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 developer version, not the book version, since the latter has not been updated in some time. 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 from the class web site.
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

### CISC-683:

ITEM	PERCENT OF GRADE
Homework assignments	35%
Midterm	20%
Final Exam	30%
Project	15%
Class Participation	described below

### CISC-483:

ITEM	PERCENT OF GRADE
Homework assignments	50%
Midterm	20%
Final Exam	30%
Project	Extra credit
Class Participation	described below

## Exams

The midterm and final exams 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 the midterm exam, please let me know right away.

## 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 put their late homework in the instructor's mailbox **at least 15 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:

<u>DUE-DATE</u>	<u>END-OF-GRACE-PERIOD</u>
3:30pm Tuesday	3:30pm on the following Thursday
3:30pm Thursday	3:30pm on the following Tuesday

After the end of the grace period, late homework sets will be penalized 25% of the total points that the assignment is worth, for each day that the assignment is late (not including Saturday and Sunday).

- **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 in the student's final grade in the course.

## 8 Lectures and Readings

The following is an initial reading list that outlines the topics that we will study in the course. This will be expanded and/or modified as the semester progresses.

<u>Topic</u>	<u>Reading: CIS-483-683</u>
<u>Introduction</u>	
What is data mining?	TSK: pp. 1-14
Preliminaries	TSK: pp. 19-44
Bias in Machine Learning	
Ethics in Data Mining	
Introduction to Weka	Handout
<u>Classification</u>	
1R and Naive Bayes	TSK: pp. 145-147, 227-240
Evaluation: error rate	TSK: pp. 148-149, 186-188, 179- 184, Handout
Evaluation: other issues	pp. 294-306
Decision Trees	pp. 150-179
Discretizing Numeric Attributes	pp. 57-62
More on Evaluation	TSK: pp. Pruning Decision Trees pp. 184-186
Evaluation: comparing classifiers	TSK: pp. 188-193
<u>Prediction</u>	
Regression	
<u>Rule-based classifiers</u>	
<u>Nearest neighbor classifiers</u>	
<u>Association rules</u>	
<u>Cluster analysis</u>	
<u>Ensembles</u>	
<u>Other topics</u>	