

# CMPS 555 Fall 2017

## INTRODUCTION TO DATA MINING

<b>Instructor</b>	Osman Kandara (aka Dr. OK)
<b>Office</b>	142 T. T. Allain
<b>Office Hours</b>	TTh: 3:30pm - 5:00pm & 6:30pm - 7:30pm (as needed) MW: 2:00pm - 2:30pm Other time by appointment!
<b>Contact</b>	<a href="mailto:kandara@cmps.subr.edu">kandara@cmps.subr.edu</a> Office: 771 - 3556 / Front Office: 771-2060 Cell: 200 - 4072 (NO CALLS AFTER 7:00PM PLEASE!)
<b>Time/Location</b>	2:00pm - 3:20pm, TTh / Henry Thurman Jr. Hall, Room 209
<b>Text Book</b>	TBA

## Course Topics

In this course, the needs for data mining will be first illustrated. Then, fundamental concepts of data mining and various data mining techniques such as association rules, neural networks, genetic algorithms, decision trees, and various statistical methods will be covered. We will also discuss the real life applications of these techniques and utilize current data mining tools such as various software packages to have hands-on experience in the field.

- | <b>Week</b> | <b>Topic</b>   |
|-------------|--|
| 1           | Introductions to Data Mining <ul style="list-style-type: none"><li>• Definition of data mining</li><li>• Data Mining goals</li><li>• Stages of Data Mining process</li><li>• Related technologies such as Machine Learning, Database Management Systems (DBMS), On-Line Analytical Processing (OLAP), Data Warehousing, Statistics, etc.</li><li>• Overview of Data Mining techniques</li><li>• Examples</li></ul> |
| 2           | Introductions to Software Packages used in Data Mining <ul style="list-style-type: none"><li>• SPSS</li><li>• SAS</li><li>• MATLAB</li><li>• Etc.</li></ul>  |
| 3           | Data/Knowledge Representation and Processing <ul style="list-style-type: none"><li>• Task relevant data</li><li>• Data cleaning and transformation</li><li>• Representing input data and output knowledge</li><li>• Visualization techniques</li><li>• Examples</li></ul>  |
| 4           | Clustering <ul style="list-style-type: none"><li>• Basic issues in clustering</li><li>• Partitioning methods</li><li>• Hierarchical methods</li><li>• Conceptual clustering</li></ul>  |

- k-means
  - Scalable clustering algorithms
  - Examples/experiments with the software packages
- 5 Data Mining Algorithms: Association Rules
- Motivation and terminology
  - Applications
  - Basic idea: item sets
  - Generating item sets, frequent sets, and the association rules
  - Experiments with market basket data
- 6 Paper Review and Midterm Exam
- Reading recently published academic papers in Data Mining and discussing in class
  - Midterm exam
- 7 Data Mining Algorithms: Classification
- Motivation and terminology
  - Applications
  - Decision trees
  - Covering rules
  - Neural Network Classifiers
  - Examples/experiments with the software packages
- 8 Data Mining Algorithms: Prediction
- Motivation and terminology
  - Statistical based prediction
  - Bayesian networks
  - Linear models
  - Instance-based (nearest neighbor) methods
  - Examples/experiments with the software packages
- 9 Artificial Neural Networks (ANN)
- Theory of learning
  - Introduction to ANN
  - Issues in ANNs
  - Applications
  - Examples/experiments with the software packages
- 10 Genetic Algorithms (GA)
- Introduction to GA
  - Mutation
  - Cross-over
  - Issues in GAs
  - Applications
  - Examples/experiments with the software packages
- 11 Paper Review
- Reading recently published academic papers in Data Mining and discussing in class
- 12 Mining Real Data
- Preprocessing data from a real domain in various fields such as forensic, business, and so on
  - Applying various data mining techniques to extract useful knowledge

- 13 Text Data Mining
  - Issues in text data
  - Text mining techniques
- 14 Advanced Data Mining Techniques
  - Mining in spatial data
  - Web mining
- 15 Future Trend in Data Mining and Final Exam
  - Exploring other Data Mining techniques in non-conventional data such as audio, video, picture, etc.
  - Final Exam

Prerequisites: CMPS 201 (Data Structures), CMPS 270 (C Programming), and CMPS 307 (Numerical Methods) or their equivalents.

## Goals/Student Learning Outcome

### Goals:

We live in an information age. As a result of this, the amount of information/data we collect and process is becoming enormous. Data mining has emerged as one of the most exciting and dynamic fields to deal with the huge amount of data in computer science. Simply stated, data mining refers to a family of techniques used to detect 'interesting' nuggets of relationships/knowledge in the large amount of data. Hence, this course prepares graduate students by providing them with both the theoretical background and practical applications through real life projects for the competitive job market emerging in data mining related fields.

### Learning Outcome:

After successfully completing this course, students will be able to:

- understand the concept of data mining, the needs for it, and what is needed to get started;
- know various data mining techniques;
- differentiate these techniques based on the needs/problems to be addressed; and
- apply those techniques in various real life applications such as information systems, engineering, business, statistics, biology, forensic, and criminal analysis, etc. to extract patterns, trends, and other useful information from databases in the respected fields.

## Grading

<b>Grading Scale</b>	A – 90% to 100%
	B – 80% to 89%
	C – 70% to 79%
	D – 60% to 69%
	F – <60%
<b>Assignments/projects</b>	30%
<b>Midterm</b>	30%
<b>Attendance, participation, quizzes, etc.</b>	10%
<b>Final</b>	30%

## Exam Schedule (Tentative)

Midterm – During the midterm week as scheduled by the university.

Final – During the final week as scheduled by the university.

**All exams are comprehensive and/or they can be in a form of class presentation or take-home focusing on graduate level research activities. Home works can also be used in lieu of exams or as part of it.**

## Guidelines

1. Due for an assignment is the next class day unless otherwise said.
2. NO due extension.
3. NO sharing (*Punished severely!*)  
*Copying is considered scholastic dishonesty and is covered in University policy. Cheating will be prosecuted both in terms of grading and University sanctions.*
4. NO late submission (*A late submission is worth ZERO!*)
5. NO curving.
6. NO Phones of any kind during class time.
7. NO makeup exam. At the instructor's discretion, the student may be allowed to use the final exam or the other one to cover the missed one at a certain percentage.
8. Reading the relevant chapters in the text book and reviving the class notes for the current topic being discussed before and after the class is expected.
9. *At any time without any advance notice, a short pop-up quiz should be expected. Those quizzes may also be considered as assignments or bonus points to the exams.*
10. At any given time without any advance notice, a new homework/programming assignment/project might be given in the class or through e-mails. Even if you have to miss the class, you are still responsible for it. *Instructor may choose certain assignments or certain questions in a particular assignment to grade instead of all.*
11. For any given assignment/project, it is your responsibility to clarify ambiguous parts. Sometimes, certain things might be left questionable by intention to challenge you to think and to ask questions about them.
12. **IT IS YOUR RESPONSIBILITY TO KEEP ALL THE GRADED MATERIALS GIVEN BACK TO YOU IN THEIR ORIGINAL FORMS UNTIL THE SEMESTER IS OVER AND TO PROVIDE WHEN ASKED.**
13. All the graded materials will be discarded if they are not picked up within 3 days.
14. For all the questions regarding a particular assignment grade you got, you should contact to the instructor. It has to be resolved within 5 business days.
15. Graduate students might get more assignments and/or different ones
16. Any failure of the any equipment that you use to do/submit/print the assignment within the last 48 hours is NOT an acceptable excuse. You had better expect such a bad luck all the time and you should start right away.
17. The programming/assignment/project guidelines that will be given to you should be observed very seriously. Even if your work produces the right output, you will lose 30 to 50 percent of the total grade if the guidelines are not followed.
18. Students with documented disabilities who believe that they may need accommodations in this class are encouraged to contact the Disability Services Coordinator in the Office of Disability Services, 246 A.C. Blanks Hall, [225-771-3546](tel:225-771-3546) (Voice), [225-771-3949](tel:225-771-3949) (Fax), as soon as possible to ensure that such accommodations are implemented in a timely fashion. Students who need accommodations must be registered with the Office of Disability Services. Students are responsible for informing the instructor of any instructional accommodations and/or special learning needs at the beginning of the semester. All discussions will remain confidential.