CSE 40647/60647 — Data Mining
M/W/F — 9:25–10:15am — DeBartolo Hall 131
http://www.nd.edu/~cse/2014sp/40647/

Everaldo Aguiar
eaguiar@nd.edu
www.nd.edu/~eaguiar
384L Nieuwland Science Hall
Office Hours: TBD

Reid Johnson
rjohns15@nd.edu
www.nd.edu/~rjohns15
384L Nieuwland Science Hall
Office Hours: TBD

Note: This course syllabus is subject to change without notice.

Course Description: Data mining can be viewed as the art of extracting knowledge from large bodies of data using methods from many disciplines, including (but not limited to) machine learning, pattern recognition, databases, probability and statistics, information theory, and data visualization. This course will largely place emphasis on the machine learning component, with relevant inclusions and references from other disciplines. The course will give students an opportunity to implement and experiment with some of the concepts, and to also apply them to real-world datasets. Additionally, we will touch upon some of the advances in related fields such as web mining and discuss the role of data mining in contemporary society.

Prerequisite(s): Demonstrable knowledge of Python programming.

Credit Hours: 3

Required Textbook:

- Introduction to Data Mining
  Authors: Pang-Ning Tan, Michael Steinbach, and Vipin Kumar
  Publication Date: 2006

Recommended Textbook:

- Python for Data Analysis
  Author: Wes McKinney
  Publication Date: 2013
Course Objectives:
Upon completion of this course, students will be able to:

1. Demonstrate knowledge of key principles and techniques of data mining.
2. Programmatically apply fundamental concepts of data mining.
3. Use statistical methods and visualization to explore and prepare data.
4. Describe the theoretical constructs and core processes of data mining.
5. Demonstrate knowledge of various predictive modeling techniques.
6. Use a data mining program to analyze data and develop predictive models.
7. Compare and evaluate the accuracy of predictive models.
8. Understand the ethical issues associated with data mining.
9. Understand and discuss the role of data mining in society.
10. Demonstrate knowledge of the emerging areas and applications of data mining.

Grading Procedure *(Subject to Change)*:

<table>
<thead>
<tr>
<th>Component</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participation</td>
<td>5%</td>
</tr>
<tr>
<td>Quizzes</td>
<td>10%</td>
</tr>
<tr>
<td>Assignments</td>
<td>20%</td>
</tr>
<tr>
<td>Course Project</td>
<td>30%</td>
</tr>
<tr>
<td>Midterm Exam</td>
<td>15%</td>
</tr>
<tr>
<td>Final Exam</td>
<td>20%</td>
</tr>
<tr>
<td>Extra Credit</td>
<td>+5%</td>
</tr>
</tbody>
</table>

Letter Grade Distribution:

<table>
<thead>
<tr>
<th>Grade</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>≥ 90.00 to 73.00</td>
</tr>
<tr>
<td>A−</td>
<td>87.00 to 89.99</td>
</tr>
<tr>
<td>B+</td>
<td>83.00 to 86.99</td>
</tr>
<tr>
<td>B</td>
<td>80.00 to 82.99</td>
</tr>
<tr>
<td>B−</td>
<td>77.00 to 79.99</td>
</tr>
<tr>
<td>C</td>
<td>73.00 to 75.99</td>
</tr>
<tr>
<td>C−</td>
<td>70.00 to 72.99</td>
</tr>
<tr>
<td>C+</td>
<td>67.00 to 69.99</td>
</tr>
<tr>
<td>D</td>
<td>60.00 to 66.99</td>
</tr>
<tr>
<td>F</td>
<td>≤ 59.99</td>
</tr>
</tbody>
</table>

Grade Assignment:
Grades in the C range represent performance that meets expectations. Grades in the B range represent performance that is substantially better than expectations. Grades in the A range represent work that is excellent. Graduate students will be held to a higher standard of grading, as they will be receiving graduate-level credit.

Grade Reference:
Grades will be maintained on Sakai. Students are responsible for tracking their progress by referring to the online gradebook.

Grading Discrepancies:
Any questions regarding how any assignment, project, exam, or other coursework is graded should be communicated to the instructors via email within seven days of receiving the grade. No regrade requests will be accepted orally, and no regrade requests will be considered after this deadline.
Course Project:

For the course project, students will be expected to collect a dataset (online or otherwise), formulate a question of interest, and perform the data analysis to address that question by using whatever tools they find appropriate. Students will be required to write a term paper and make a class presentation on their project. All project topics should be pre-approved by the instructors.

Undergraduate students are encouraged to work in teams of 2–4 (with a maximum of 4 members) for the course project. Graduate students are encouraged to work in teams of 2, unless there is a specific project related to their thesis/work, which should be discussed with the instructors.

Course Policies:

- Collaboration Policy
  - Students are encouraged to discuss the courses’ ideas, materials, and assignments with others in order to better understand them. However, unless instructed otherwise, students must turn in work that is their own. Students must write their own code, run their own data analyses, and write up their own results and answers to assignment questions.

- Labs and Assignments
  - There will be several assignments. Unless announced otherwise, assignments will be due at 11:59pm ET on the provided submission date.
  - Graduate students may be assigned additional work.

- Quizzes and Exams
  - There will be several quizzes, one midterm exam, and one final exam.
  - Make-up quizzes and exams will be allowed as per the du Lac Class Absence Policy. Whenever possible, students are expected to provide advance notice if they will be unable to take a quiz or exam. Make-up quizzes and exams for travel to academic conferences will be provided at the discretion of the instructors.

- Late Policy
  - Assignments submitted after the submission deadline but within the next day are counted as one day late. The next 24 hours will be counted as two days late, and so on. Each day late contributes a 10% penalty to the original assignment value. Assignments more than three days late will ordinarily receive a 0.

- Academic Honesty
  - A commitment to honesty is expected of all students. The du Lac Academic Codes, which relate to academic integrity, will be strictly followed. All references and sources, both to text and code, should be properly cited in all submitted work.
Course Contents:
Details may change depending on time and class interests.

- **General Topics:**
  
  *Data Understanding*: types of data; information and uncertainty; classes and attributes; interactions among attributes; relative distributions; summary statistics; data visualization

  *Data Quality*: inaccurate data; sparse data; missing data; insufficient data; imbalanced data

  *Social Challenges*: data ownership; data security; ethics and privacy

- **Unsupervised Learning Topics:**
  
  *Data Reduction and Feature Enhancement*: standardizing data; sampling data; using principal components to eliminate attributes; limitations and pitfalls of principal component analysis (PCA); curse of dimensionality

  *Clustering*: dissimilarity and scatter; categorization; $k$-means clustering; hierarchical clustering; distance measures; shape of clusters; determining the number of clusters; evaluating clusters

  *Association Analysis*: association rule learning; the Apriori algorithm; FP-Growth; market basket analysis

- **Supervised Learning Topics:**
  
  *Regression*: review of linear regression; assumptions underlying linear regression

  *Classification*: supervised categorization; linear classifiers; logistic regression; regression trees; classification trees; Bayes’ Theorem; naïve Bayes; support vector machines (SVMs); confusion matrices; receiver operating characteristic (ROC) curves; precision and recall; lift curves; cost curves

  *Model Selection and Validation*: training error and optimism; the Bayes error rate; inductive bias; the bias-variance tradeoff; overfitting; Occam’s Razor; minimum description length (MDL); sampling bias; the validation set approach; leave-one-out cross-validation; $k$-fold cross-validation; bootstrapping; jackknifing; data snooping

  *Ensemble Learning*: bootstrap aggregating (bagging); boosting; stacking/blending; random subspaces; random forests

- **Miscellaneous Topics (as time permits):**
  
  *Recommender Systems; Reinforcement Learning; Active Learning; Semi-supervised Learning; Transfer Learning; Deep Learning; Data Stream Mining*
**Tentative Course Outline:**

The weekly content coverage is subject to change.

<table>
<thead>
<tr>
<th>Week</th>
<th>Content</th>
</tr>
</thead>
</table>
| Week 1   | Introduction to Data Mining  
Collaborative Filtering (Part I)                                      |
| Week 2   | Classes and Attributes  
Types of Data  
Summary Statistics  
Data Visualization                                             |
| Week 3   | Data Reduction  
Feature Enhancement                                                      |
| Week 4   | Association Rules                                                        |
| Week 5   | Clustering: \( k \)-Means  
Clustering: Expectation Maximization  
Clustering: Hierarchical Methods  
Clustering: Evaluation                                                  |
| Week 6   | Regression: Linear Regression  
Regression Trees  
Regression: Evaluation                                                  |
| Week 7   | Classification: Introduction and Nearest Neighbor  
Classification: Bayesian Learners  
Classification: Naïve Bayes                                           |
| Week 8   | Review for Midterm Exam  
Midterm Exam                                                                |
| Week 9   | Winter Break                                                            |
| Week 10  | Classification: Rule-Based Learners  
Classification: Classification Trees                                      |
| Week 11  | Classification: Overfitting  
Classification: Evaluation                                                  |
| Week 12  | Classification: Comparisons and Tradeoffs                                  |
| Week 13  | Ensembles: Resampling Methods  
Ensembles: Hybrid Methods                                                 |
| Week 14  | Advanced Topics  
Collaborative Filtering (Part II)                                         |
| Week 15  | Project Presentations                                                     |
| Week 16  | Concluding Discussion  
Review for Final Exam                                                       |