CSCI 291: Data Mining and Pattern Recognition

Syllabus – Spring 2015

Nathan Sommer
sommern@moravian.edu
610-625-7786
PPHAC 213

Office Hours: MWF 9-10
Thursday 3-5
or by appointment

Course Description

This course is an introduction to algorithmic approaches to data science. We will cover a wide range of algorithms for classification, clustering, regression, and association rule learning. We will also discuss issues that must be taken into account when applying these algorithms such as generalization, overfitting, and the curse of dimensionality.

Course Goals

Upon completion of this course, a successful student will be able to:

- Select appropriate algorithms for different data mining and pattern recognition tasks.
- Use Python's scientific computing libraries to perform more efficient computations on numeric data sets.
- Reduce the dimensionality of a data set for more accurate and efficient learning.
- Divide data into training, test, and validation sets to measure generalization accuracy and prevent overfitting.

Required Text

The following text is required:

- *Machine Learning: An Algorithmic Perspective, 2nd Edition* by Stephen Marsland.

Reading will be assigned as we progress through the material. In addition to the required text, supplementary readings will be given periodically during the semester.

Course Outline

Here is a rough ordering of topics for the semester. Topics may be added or removed depending on how quickly we progress.

- Introduction

- Scientific Python
  - NumPy
  - matplotlib
  - Other useful packages
  - IPython
  - Python IDEs
- Math Background
  - Linear algebra
  - Probability theory
  - Statistics
- Supervised Learning
  - Accuracy, Generalization, and Overfitting
  - Linear Regression
  - Dimensionality Reduction
  - Neural Networks
  - Decision Trees
  - Support Vector Machines
  - Ensemble learning
- Unsupervised Learning
  - Similarity Measures
  - k-means
  - k Nearest Neighbors
- Itemsets and Association Rules
  - Apriori Algorithm
  - Eclat Algorithm

Assignments and Tests

Your grade will be calculated based on the following items:

- **Homework** – Homework problems will be assigned most class sessions and will be due the next class session. Each assignment will be some combination of small programming tasks, written exercises, and short answer questions. These assignments will help you explore topics more deeply and practice using Python's scientific computing libraries.

- **Programming Projects** – Larger programming projects will also be assigned throughout the semester. You will be given more time to complete these assignments than the homework assignments. Each programming project might involve implementing an algorithm or applying several techniques and algorithms to a specific problem.
- **Midterm Exam** – There will be a midterm exam in class. The exam is tentatively scheduled for Friday, March 6. You may only re-schedule a test for college approved absences or a documented illness. In either case you must contact me *before* the beginning of the test.
- **Final Exam** – There will be a cumulative final exam given in class on Wednesday, May 6 at 8:30 a.m.

Grading

Grades will be weighted as follows:

35% Homework
35% Programming Projects
15% Midterm Exam
15% Final Exam

I will use the standard 90, 80, 70, 60 grading scale with pluses and minuses. I may relax these standards as necessary but I will not raise them.

Course Policies

- **Late Policy** – Generally I expect assignments to be turned in on time. I understand that this is not always possible, so in most cases I will accept assignments one class session late without penalty. However, if this becomes a pattern it will start to affect your grade. To keep up the pace of the class, assignments will not be accepted after they are one session late unless there are special circumstances.
- **Extensions** – In certain circumstances granting an extension to an assignment's due date is perfectly reasonable. If you feel you need an extension, please contact me about it before the day it is due.
- **Absences** – You are expected to attend each class, but I understand that occasionally there will be exceptional circumstances. If you miss a class or know that you will miss an upcoming class, please contact me as soon as possible to explain the situation. You will still be expected to keep up with assignments and class content.
- **Academic Honesty** – You are encouraged to discuss and work with other students on homework assignments. However, the work you turn in should be your own. Help each other understand the concepts, but produce your own code and writing. See the Moravian College student handbook for more on the school-wide stance on academic honesty.
- **Disabilities** – Students who wish to request accommodations in this class for a disability should contact Ms. Elaine Mara, Assistant Director of Academic & Disability Support,

located on the first floor of Monocacy Hall (extension 1401). Accommodations cannot be provided until authorization is received from the Academic & Disability Support office.

Communication

Feel free to talk to me about any issues that may arise. Email is generally the fastest way to get in touch with me if you have a quick question. You are most likely to find me in my office during office hours, but I will be around other times as well so dont be shy about stopping by.

The details of this syllabus are subject to change based on our progress through the material.