Machine Learning in Practice/ Applied Machine Learning
11-344,11-663,05-834,05-434

Instructor: Dr. Carolyn P. Rosé, cprose@cs.cmu.edu
Office Hours: Gates-Hillman Center 5415, Time TBA

Teaching Assistants: TBA
TA Office Hours: TBA

Course Cross-listed in: HCII, LTI
Note: Blackboard link says **Applied Machine Learning**
Units: 12 (PhD/Master's/Undergrad level)

Course Meeting Time/Location: TR 3:00pm-4:20pm, Baker Hall A53

Books:
Witten, I. H., Frank, E., and Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*, third edition, Elsevier: San Francisco, ISBN 978-0-12-374856-0

Prerequisites: Some Java programming experience is desirable, but not necessary.

Course Description:

Machine Learning is concerned with computer programs that enable the behavior of a computer to be learned from examples or experience rather than dictated through rules written by hand. It has practical value in many application areas of computer science such as on-line communities and digital libraries. This class is meant to teach the practical side of machine learning for applications, such as mining newsgroup data or building adaptive user interfaces. There will be a heavy project focus, and when you have completed the course, you should be fully prepared to attack new problems using machine learning. Many students who have taken the course in previous semesters have reported that it was a key factor in getting a competitive internship or job and that it prepared them well for these important next steps in their career.

While it will be essential to learn conceptually how machine learning algorithms work and interact with data, the emphasis will be on effective methodology for using machine learning to solve practical problems. Note that this is not just learning to use a tool bench like Weka. It is about knowing how to conceptualize a problem, knowing how to represent your data, being able to interpret your results properly, doing an effective error analysis, and using the results of the error analysis

to make strategic decisions about how to adjust the way you have set up your data and selected and tuned your algorithms.

This course does not assume any prior exposure to machine learning theory or practice. And students from all over campus have been successful in completing this course. However, it should be known that this is a challenging course and will require the full 12 hours per week warranted by a 12 unit course, between readings, homework, lectures, and project work.

We will cover a wide range of learning algorithms that can be applied to a variety of problems. In particular, we will cover topics such as decision trees, rule based classification, support vector machines, Bayesian networks, and clustering. In addition to readings from the course textbook, we will have additional readings from research articles that will be announced ahead of time and distributed on Blackboard.

Grades will be based on assignments (which will be almost weekly) and quizzes, 2 midterms, and a course project. Note that the course project is 50% of your grade, and you cannot pass the course without completing it.

Course Procedures:

Below you'll see the material for the course divided into units of one to three weeks in duration. You can find the Assignments and Quizzes in the course documents folder on Blackboard as they are assigned. Below where the lectures are listed, you will see which reading assignments, quizzes, and assignments are associated with each unit. I suggest that for each unit, you do the readings prior to the lectures.

On most Tuesdays there will be a quiz in the first 5 minutes of class. The quizzes are meant to help you gauge your reading comprehension and to make certain ideas salient that will be key topics in the lectures. The lectures will make the material from the readings more clear, and will emphasize the topics from the readings that are most important from the standpoint of the course objectives. The assignments are meant to give you valuable practice at applying the principles covered in the readings and lectures. ***Note that quizzes will typically cover material that has been assigned for the week on which the quiz takes place but may not have been covered yet in lecture.***

Assignments will experiments and activities using the Weka toolkit (http://www.cs.waikato.ac.nz/ml/weka/) and the LightSIDE toolkit (http://www.cs.cmu.edu/~emayfile/side.html).

Assignments should be turned in to Blackboard. You should label your submission with your name and the name of the assignment (e.g., assignment1.doc). You will

receive feedback on your assignments, which is designed to help you learn from your mistakes, but you will not be graded down for your mistakes. Instead you will just get credit for having done the assignments. Similarly, you will receive feedback on the quizzes, but you won't get graded down for your mistakes. Both the assignments and the quizzes are meant more as learning activities than assessments.

Important Note: Assignments will usually be assigned in class on a Thursday and will be due the following Thursday before class, turned into to Turnitin Assignment on Blackboard (see http://www.youtube.com/watch?v=59iYdvx4Wyk for instructions on how to do that if you are not familiar with Blackboard 9!).

Mid-terms will serve as formal assessments and will serve to measure your level of competence in connection with the course objectives.

The term project will involve applying machine learning to a substantial problem of the student's choice. Several options are found in the Projects subfolder of the Course Documents folder on blackboard. Students may select one of these projects or may propose one of their own design. Students who wish to design their own project should check in about their plans with the instructor as early as possible in the semester.

Grading Criteria

Quizzes (10%)
Assignments (20% total)
Mid-terms (10% each)
Course project (50%)

Videos of lectures from a previous semester are available as a resource to you, and you should be able to locate relevant video lectures by matching the unit names and then following the links in the syllabus from the online course you will find in blackboard in the Course Documents folder. Note that these online lectures only work properly in Windows. While the content of the video lectures will substantially overlap with what we will cover this semester, you should note that the content of the lectures has been revised.

Course Schedule

Adjustments will be made if necessary.

Week 1 *Course Intro/ Weka Intro* (Witten & Frank, CH 1, 10-11)
Week 1 Lecture 1
Week 1 Lecture 2 Assignment 1 assigned [Getting to know Weka](short!!)

Week 2 *Input and Output* (Witten & Frank, CH 2, 3, supplementary paper in Readings folder)
Week 2 Lecture 1
Week 2 Lecture 2 Assignment 2 assigned [Intro to Error analysis and Data Representation]

Week 3-4 *Basic Statistical Models and Linear Models* (Witten & Frank, Ch 4.2, 4.6)
Week 3 Lecture 1
Week 3 Lecture 2 Assignment 3 assigned [Understanding Naïve Bayes]

Week 4 Lecture 1
Week 4 Lecture 2 Project proposal is due!!

Week 5 *Applied Machine Learning Process and Evaluation* (Witten & Frank, CH 5, 13)
Week 5 Lecture 1
Week 5 Lecture 2 Assignment 4 assigned [Understanding proper evaluation methodology]

Week 6-7 *Working with Text Part 1*(Witten & Frank 9.5, 9.6, Application papers in Readings folder)
Week 6 Lecture 1 Finish Evaluation and Start Text
Week 6 Lecture 2 Assignment 5 assigned [Extracting meaningful features from text]

Week 7 *Working with Text Part 2*(SIDE user's manual (from the SIDE webpage))
Week 7 Lecture 1
Week 7 Lecture 2 Assignment 6 assigned [End to end machine learning process with text data], due 11:59pm on Friday, March 9. Feedback will be posted before Monday, March 19.

Week 8 *Advanced Text Mining – Working towards generalizable models* (Readings in Week 8 folder)
Week 8 Lecture 1 More on Text
Week 8 Lecture 2 Guest Lecture/SIDE Review

Week 9 *Advanced Tree and Rule Based Learning* (Witten & Frank, CH CH 3.3-4, 4.1, 4.3, 4.4, 6.1, 6.2, 6.5)
Week 9 Lecture 1 Midterm 1 assigned [due 24 hours later]
Week 9 Lecture 2

Week 10 More advanced *Linear and Statistical Models, Instance Based Models, and Clustering* (Witten & Frank, CH 6.4, 6.5, 6.6, 6.7, 6.8)
Week 10 Lecture 1 Assignment 7 assigned [Investigating the impact of assumptions about the form of solutions: Contrasting Tree and Rule based learning]
Week 10 Lecture 2 Guest Lecture

Week 11 *Feature Selection and Optimization* (Witten & Frank, CH 7)
Week 11 Lecture 1 Assignment 8 assigned [Doing a simple optimization]
Week 11 Lecture 2

Week 12 *Semi-Supervised Learning, Machine Learning Extensions* (Witten & Frank, CH 8)
Week 12 Lecture 1 Assignment 9 assigned [Doing a more complex optimization]
Week 12 Lecture 2

Week 13-15 *More Machine Learning Applications* (Chapter 9 and Application papers to be announced)
Week 13 Lecture 1
Week 13 Lecture 2

Week 14 Lecture 1
Week 14 Lecture 2

Week 15 Lecture 1 Midterm 2 Assigned (due 24 hours later)
Week 15 Lecture 2

No final exam!
Projects due by Dec 14