DATA SCIENCE II:
MACHINE LEARNING
MTH 9899
BARUCH COLLEGE

SPRING 2015

MIGUEL A. CASTRO

T.A.:

Liqun Zhu.

Class Meetings

There will be 7 class meetings from 6 to 8:30 PM on the following Wednesdays:

3/25, 4/01, [ ]*, 4/15, 4/22, 4/29, 5/06, 5/13. (*No class on 4/08 due to Spring Recess.)

Topics

We will explore a variety of topics in Machine Learning with emphasis on applications. Machine Learning is a vast field and we are a bit constrained in time with only 7 class meetings. Topics will include supervised and unsupervised learning, Artificial Neural Networks, the design of Machine Learning experiments, learning rules, measuring and comparing performance, neural net architecture, regularization approaches, classifiers, association rules, and other select topics, time permitting. We will consider one supervised learning and one unsupervised learning paradigm in some depth, and survey several other important techniques. The course assumes knowledge of calculus and probability, as well as some basic familiarity with data structures, optimization, and search algorithms. At the end of this course, the student should be able to apply Machine Learning algorithms or packages to extract actionable information from data, and evaluate the success of the process. Our focus will be on Machine Learning paradigms regardless of the domain of application, but we will emphasize applications related to finance whenever possible.

Reference Texts:

- *Introduction to Machine Learning*, 2nd Edition (Alpaydin, 2010)
- *The Elements of Statistical Learning, Data Mining, Inference, and Prediction,* 2nd Edition (Hastie, Tibshirani & Friedman, 2009)
- *Machine Learning, An Algorithmic Perspective* (Marsland, 2009)
- *Machine Learning* (Mitchell, 1997)

The reference texts will be placed on reserve in the library and you are not required to purchase them. We will not follow any textbook in class; the lecture notes and homework assignments are self-contained. The Alpaydin book is the closest to the class level. It has a fairly comprehensive and intuitive introduction to Machine Learning, covering a wide array of topics with a focus on clarity of exposition and understanding. The Tibshirani book offers a more rigorous treatment for those of you who want to dig deeper into a particular subject (*e.g.* for the Class Project, see below). The Marsland book emphasizes algorithms and applications and contains examples in Python. The Mitchell book offers a good blend of depth and breadth, and is considered a classic introduction. Although it's a bit dated now—given how fast the field has evolved in the last couple of decades—it's still worth a browse.

Homework Assignments (25%)

There will be 5 individual homework assignments due *before* the start of class. These should be emailed to both me and the T.A. in PDF format. The email must be time stamped *before* the start of class on the due date; no exceptions. The homework assignments will contribute 25% towards your final grade. You are encouraged to discuss assignments with your classmates, but what you turn in must be your own work. *Late homework assignments will receive a grade of "0".*

In-Class Exam (40%)

There will be cumulative Exam on May 8th covering all the material from the beginning of class until then. It will constitute 40% of your grade. Note that this is the week before the last day.

Class Project (25%)

The Class Project will account for 25% of your final grade in the course. You can choose to either do a literature reading (no programming required) or conduct research of your own (programming required). ***Note that if you choose a literature reading the grading will be much harsher.*** You will work in a group of 1 to 4 students.

For a literature reading, choose a published research paper related to a Machine Learning paradigm applied to a substantial problem of interest *applied to any area of finance*.

If you choose a research project of your own, your group will choose a Machine Learning paradigm and apply it to extract information or conclusions or make predictions from a corpus of data. It is not required that the problem be finance-related, but you are encouraged to choose a finance-related topic that interests you and for which you can find data. The report can be no less than 500 words long (this refers to the main body of the report, and it excludes bibliography, graphs, tables, source code, etc.).

You're encouraged to include an *Abstract*, an *Introduction*, and a *Conclusion* section along with a *Bibliography*. The report should contain graphs, tables, charts, equations, and source code, as needed. It should clearly state the Machine Learning paradigm or methodology being used, the goal of the project, the data used, a quantitative assessment and a discussion of the success (or lack thereof) of the Machine Learning paradigm used, and a discussion of why that paradigm was chosen in favor of others. The project could also consider a comparison of two or more paradigms applied to the same data corpus. It should also clearly state the findings of the study, along with any potential pitfalls, and suggestions for improvements, changes, or further work.

If you're doing your own research project, a good approach to your project is to select a paper from the literature and attempt to replicate or extend its results. Below is a list of good candidate papers that you can choose from (thanks to former T.A. Alejandro Cañete for providing this list). You can also choose your own paper for your project.

If you're doing your own research project, you will be responsible for acquiring a data corpus to work with. There are many data sources available on the Internet (e.g.: kdnuggets, infochimps, datawrangling data sets, nber, worldbank, Rob J. Hyndman's time series data library, UCI Machine Learning Repository, Statlib, Delve, Amazon Web Services repository, along with research data sets from universities, …). For example, you can search for "public data sets" or "free data sets." The data should be interesting and plentiful. If you use artificially-generated data sets, make sure to state exactly how you generated the data, and how much noise you introduced into your data set, along with some plots.

You may write your own code (programming language should be Matlab; if you have other suggestions talk with the instructor first), or use pre-packaged software, be it public domain or commercial software. There are many such packages for neural nets, SVMs, classification trees, etc. that you can search for.

The project report is due in PDF format by the start of the last class. Be prepared to present a brief version of your report to the class via a short (10-minute) PowerPoint presentation. If there is more than one person in your team, you are also required to turn in by email a confidential assessment of how much each of your classmates, including yourself, contributed to the Class Project. The assessment will be a number between 1 and 10, with 1 meaning "Contributed very little or nothing," and 10 being "Contributed most significantly."

Project Proposal (one per group) consisting of one or two paragraphs describing the project and the members of each team is due *May 1st*.

Instructor Discretion (10%)

The instructor will have leeway to bump a grade up or down based on a student's participation, contribution to the class and project, etc.

Contact Info

Instructor: Miguel A. Castro
Email: Miguel.Castro98@gmail.com
Cell Phone: 212-203-2654

T.A.: Liqun Zhu
Email: gasquit@gmail.com

HERE'S LIST OF SAMPLE PROJECT PAPERS:

Tutorials: Neural Networks

Feed Forward Neural Networks
http://saba.kntu.ac.ir/eecd/Fatehi/Lectures/Intelligent%20Systems/NeuNet/Papers/NeuralNetworksTutorial.pdf

Support Vector Machines
http://www.mingzeng.net/wp-content/uploads/2009/11/support-vector-machines%E7%BB%8F%E5%85%B8.pdf

Seminal Papers: Neural Networks

Fuzzy Perceptron
http://raic.kunsan.ac.kr/paper/Ij/IJ-2005-33.pdf

Generalized Perceptron
http://www.cs.huji.ac.il/~shais/papers/ShalevSi05.pdf

Seminal Papers: SVM

Least Squares SVM
http://www.cs.mcgill.ca/~rshah3/least_squares_svm.pdf

Relevance SVM
http://jmlr.csail.mit.edu/papers/volume1/tipping01a/tipping01a.pdf

Seminal Papers: Kernel Methods

"Kernel LMS"
http://www.cnel.ufl.edu/~weifeng/filesfordownload/paper/kernel_least_mean_square_ieee_sp.pdf

"Kernel PCA"
http://www.tribesandclimatechange.org/docs/tribes_450.pdf

Machine learning in finance

"Kernel methods & Market Microstructure"
Multiple Kernel Learning on the Limit Order Book
http://jmlr.csail.mit.edu/proceedings/papers/v11/fletcher10a/fletcher10a.pdf

"Online universal portfolio tracking + mean reversion"
Confidence Weighted Mean Reversion Strategy for On-Line Portfolio Selection
http://jmlr.csail.mit.edu/proceedings/papers/v15/li11b/li11b.pdf

"Online optimized neural network for trade signal generation + trend following"
Learning to Trade via Direct Reinforcement
http://cseweb.ucsd.edu/~dboswell/PastWork/Moody01LearningToTradeViaDirectReinforcement.pdf

"Factor based conditional return forecasting/trading"
Financial Time Series Prediction Using Least Squares Support Vector Machines Within the Evidence Framework
http://cosmal.ucsd.edu/~gert/papers/tnn_01.pdf

"Option pricing using Neural Networks"
http://www.sfu.ca/~rgencay/jarticles/ieeenn-bagging.pdf