ML4Trading

From Quantwiki

ML4TradingOverview

Contents

Machine Learning for Trading 2012

keywords: algorithmic, algorithms, trading, machine learning, ML, finance, equities, markets, quantitative finance, georgia tech, gt, Tucker Balch

Important Announcements

New:

- Project 1 is posted. 2012Fall7646_Project_1
- Homework 5 is posted. 2012Fall7646_Homework_5

Course Overview

This course introduces students to the real world challenges of implementing machine learning based trading

strategies including the algorithmic steps from information gathering to market orders. The focus is on how to apply probabilistic machine learning approaches to trading decisions. We consider statistical approaches like linear regression, KNN and regression trees and how to apply them to actual stock trading situations.

- For more detail on topics to be covered see ML4TradingOverview

Who the Course is For

The course is open to and intended for graduate and upper level undergraduate students in Computing, ISYE, Math & Management.

Prerequisites: Students should have *strong coding skills* and some familiarity with equity markets. No finance or machine learning experience is assumed. Here's a short test to check if you have strong programming skills: quiz (http://www.intelligentappliances.org/multirobot/index.php?title=Strong_programming_skills) . If you don't do well on that quiz, you should either drop the course, or be sure to plan so that you can devote extra time to the course.

Note that this course serves CS major students with machine learning experience, as well as students in other majors such as ISYE, MGMT, or MATH who have different experiences. All types of students are welcome! The ML topics might be "review" for CS students, while finance parts will be review for finance students. However, even if you have experience in these topics, you will find that we consider them in a different way than you might have seen before, in particular with an eye towards implementation for trading.

Student Responsibilities

- Read the emails sent to the course email list. Check at least daily.
- Participate in class and via the piazza site.
- Don't plagiarize.

Course Logistics

- Instructor: Associate Professor Tucker Balch (http://www.cc.gatech.edu/~tucker)
  - Office hours: Tu/Th 4:30-5:30 (after class) or by appointment
  - firstname at cc.gatech.edu
  - phone 678-523-8685
- TA Fang Li
  - Office Hours: Tuesdays 1PM - 3PM
- TA Jia Zhao
  - Office Hours: Thursdays 10AM - 12 Noon
  - CCB 308
- TA Sourabh Bajaj
- Course Website: http://wiki.quantsoftware.org/wiki/ML4Trading. (this webpage)
- Some parts of the course will be based on readings from this book: Active Portfolio Management by Grinold & Kahn (http://www.amazon.com/Active-Portfolio-Management-Quantitative-Controlling/dp/0070248826/ref=sr_1_1?ie=UTF8&s=books&qid=1263182044&sr=1-1) . You will have to complete the readings, but you don't have to buy the book if you are a GT faculty or student, you may be able to read the book online [here (http://portal.library.gatech.edu/vufind/Record/651256)] at no cost.
- Prerequisites: Machine learning and portfolio management experience is *not* assumed; the course is

designed to provide students with the necessary background they will need on these topics. Programming will be in the Python language. Students are expected to be strong programmers (or willing to invest significant effort in learning to program in Python).
- Computing_environment_for_ML4Trading
- GT T-Square site for the class: t-square is here (https://t-square.gatech.edu/portal/site/XLS0823152204201208.201208) .
- discussion forum on piazza.com (https://piazza.com/class#fall2012/cs7646) .
- Resources

Goals For the Course

By the end of this course, students should be able to:

- Understand data structures used for algorithmic trading.
- Know how to construct software to access live equity data, assess it, and make trading decisions.
- Understand 3 popular machine learning algorithms and how to apply them to trading problems.
- Understand how to assess a machine learning algorithm's performance for time series data (stock price data).
- Know how and why data mining (machine learning) techniques fail.
- Construct a stock trading software system that uses current daily data.

Some limitations/constraints:

- We use daily data. This is not an HFT course, but many of the concepts here are relevant.
- We don't interact (trade) directly with the market, but we will generate equity allocations that you could trade if you wanted to.

Grading

- Projects: 60%
- Exam 1: 5%
- Exam 2: 5%
- Homework: 20%
- Final: 10%

Plagiarism

In most cases I expect all code that you submit was written by you. I will present some libraries in class that you are allowed to use (such as pandas and numpy). Otherwise, all source code, images and write ups you provide should have been created by you alone.

What is allowed:

- Meeting with other students to discuss implementations. You should talk about solutions at the pseudo code level.
- Sharing snippets of code to solve specific (small) problems such as examples of how to address sections of arrays in Python. In this case the shared code should not be more than 5 lines.
- Searching the web for other solution outlines that you may draw on (but not copy directly). If you are inspired by a solution on the web, you MUST cite that code with comments in your code.

What is not allowed:

- Copying sections of code longer than 5 lines. Note that merely changing variable names does not suffice.
- Copying code from the web.
- Use of ideas from the web that are not cited in comments.

August

DATE		TOPICS	NOTES
Tuesday	August 21	Class overview	
Thursday	August 23	The $1T Bet Movie	
Tuesday	August 28	Mechanics of Stock Prices: the order book What is HFT? Tour of a hold-overnight hedge fund	
Thursday	August 30	Approach: I assume you want to run a hedge fund Why run a hedge fund? How to assess a fund?	

September

DATE		TOPICS	NOTES
Tuesday	September 4	Company Value	
Thursday	September 6	Video: [1] (http://podcast.cnbc.com.edgesuite.net/MadMoney-083011.mp4) Cramer's Guide to the Stock Market	
Tuesday	September 11	Capital Assets Pricing Model Read: Grinold & Kahn Chapter 2	
Thursday	September 13	Recap of CAPM Definition of Beta and example with scatter plots Discussion of scatter plots (fatness) Mechanics of shorting Magic of long/short investing: removing market risk Relationship of Long/Short to CAPM	
Tuesday	September 18	Biased coin betting (HW 2 overview)	

		Workshop on setting up computing environment
Thursday	September 20	Find correlated returns (HW 3 overview) Workshop on reading real stock data (tutorial 1) Balch away
Tuesday	September 25	HW 3 overview Project 1 overview
Thursday	September 27	Big picture for next group of homeworks: Goal: Software to develop and test your strategies Hypothesis testing: Does event X affect price? (Event Studies) What is the shape of the return curve? Generate the trades Test with market simulator

October

DATE	----------	TOPICS----------	NOTES----------
Tuesday	October 2	Efficient Markets Hypothesis (3 versions) Technical analysis Fundamental analysis A bit about data Create & plot your own technical indicator (HW 5 overview)	HW 3, Find correlated stocks Due HW 4, Optional, Due Paper of interest: Markets are efficient iff P=NP (http://arxiv.org/pdf/1002.2284v2.pdf)
Thursday	October 4	About Data Adjusting for Dividends and Splits Missing Data Fill forward and fill back The fundamental law of active portfolio management	
Tuesday	October 9	Integrate your technical indicator with Event Studies (HW 6 overview)	HW 5, Create a Technical Indicator Due

Thursday	October 11	Introduction to ML Regression Classification Exam 1 preview Data we will use for ML	
Friday	October 12	Drop Day	
Tuesday	October 16		Project 1, Market Simulator Due
Thursday	October 18	Exam 1 (30 minutes) Instance-based learning Linear regression KNN introduction Use technical indicator to generate trades (HW 7 overview)	
Tuesday	October 23	KNN continued Compare LR and KNN for time series prediction (Project 2)	HW 6, Event Study Due HW 7, Technical Indicator generates trades due
Thursday	October 25	Assessing a learning algorithm. Cross validation Roll forward cross validation	
Tuesday	October 30	Overfitting.	

November

DATE		TOPICS	NOTES
Thursday	November 1	Balch on travel Ensemble learners Bagging Boosting	
Tuesday	November 6	Decision Trees How to use one if you have one How to learn one	
Thursday	November 8	Random Trees	Project 2, KNN due

		Random forests (Project 3)	
Tuesday	November 13	Tricks with leaves NaNs Regression Leaves	
Thursday	November 15	News & Prices (Prof Jonathan Clarke guest talk)	
Tuesday	November 20	Yes we will have class on this day Integrating your learner with a trader (Project 4)	
Thursday	November 22	Thanksgiving (no class)	
Tuesday	November 27	More on integrating learner with trader (Project 4)	Project 3, Random Forests Due
Thursday	November 29	Feature selection	

December

DATE	----------	TOPICS----------	NOTES----------
Tuesday	December 4	Exam 2 preview	
Thursday	December 6	Exam 2	
Tuesday	December 11	Final Exam Week No Class	
Thursday	December 13	Final Exam Week No Class	Project 4, Combine ML & Trading Due

Slide Decks

- [001-ML4T-Intro.pdf (http://wiki.quantsoftware.org/images/1/1f/001-ML4T-Intro.pdf)]
- [003-ML4T-Inside-Hedge-Fund.pdf (http://wiki.quantsoftware.org/images/c/c3/003-ML4T-Inside-Hedge-Fund.pdf)]
- [005-ML4T-Assessment-of-Fund.pdf (http://wiki.quantsoftware.org/images/1/11/005-ML4T-Assessment-of-Fund.pdf)]

Homework & Projects

- 2012Fall7646 Homework 1: Create a portfolio of 4 equities, and assess it in excel or libre office. Due 11:55PM, September 6.
- 2012Fall7646 Homework 2: Write a Python program to investigate different one day trading methods. Due 11:55PM, September 26.

- 2012Fall7646 Homework 3: Write a Python program to to assess real stock data. Due 11:55PM, October 2.
- 2012Fall7646 Homework 4: Optional homework: Assess correlation and Beta. Due 11:55PM, October 2.
- 2012Fall7646_Project_1: Project 1: Build a market simulator. Due 11:55PM, October 11.
- 2012Fall7646 Homework 5: Create an indicator and plot it, Due 11:55, October 9.

Other Resources

- python tutorial for timeseries
- wikipedia page on KNN (http://en.wikipedia.org/wiki/Nearest_neighbor_search)
- Andrew Moore's slides on KNN (http://www.cs.cmu.edu/afs/cs/Web/People/awm/tutorials/mbl.html)
- Culter paper on random forests (http://www.interfacesymposia.org/I01/I2001Proceedings/ACutler/ACutler.pdf)
- original paper on how to build a single decision tree (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.167.3624&rep=rep1&type=pdf)
- See also Resources on this wiki.
- 2010_GT_Course
- 2011_GT_Course

Projects from Past Versions of the Course

These are from last year. We will be changing them for this year.

- 2011Fall7646 Homework 1: Write a Python program to print the first 100 prime numbers, Due at 11:55, Thursday September 1.
- Reading assignment: Chapters 1 & 2 of Grinold & Kahn text book. Due before Thursday class, September 1.
- 2011Fall7646 Homework 2: Get your account set up at gekko.cc.gatech.edu, Due at 11:55, Tuesday September 13.
- 2011Fall7646 Homework 3: Generate some of your own plots, Due at 11:55, Thursday September 15.
- 2011Fall7646 Homework 4: Compute and plot Bollinger bands. Due at 11:55, Monday October 3.
- 2011Fall7646 Project 1: Event studies. Due at 11:55, Tuesday October 25.
- 2011Fall7646 Project 2: KNN Learner. Due at 11:55, Thursday November 3.
- 2011Fall7646 Homework 5: Compare two implementations of KNN. Due at 11:55, Tuesday November 15.
- 2011Fall7646 Homework 6: Classify News Articles. Due at 11:55, Tuesday November 22.
- 2011Fall7646 Project 3: Find the best indicators. Due at 11:55, Wednesday, December 7.
- 2011Fall7646 Project 4: Classify news articles. Due at 11:55, Wednesday December 14.