Syllabus

Bus241f(2): Machine Learning and Data Analysis for Business and Finance: Spring 2020

Key information

Instructor

- Blake LeBaron
- blebaron@brandeis.edu
- http://www.brandeis.edu/~blebaron
- Sachar 204
- Office hours: Monday 12:30-1:30pm, Thursday, 11:30-12:20pm

TA’s

- Section 1: Zijing (Jimmy) Hu, zijinghu@brandeis.edu
- Section 2: Zhichao(Simon) Zhang, zhichaozhang@brandeis.edu

Times:

Class Times: M/W 9:30–10:50(section 1), 11:00–12:20(section 2), Lee Hall

Detailed information

Course Description

This course is a general topics course on machine learning tools, and their implementation through Python, and the Python package, Scikit Learn. Students will finish the class with a basic understanding of how to execute predictive analytic algorithms. They will also have a good sense for how to evaluate and test their predictive models. The course is statistical in nature, but will use only basic statistics from a standard one semester econometrics class. Given the time horizon (6–7 weeks) the course can only provide a birds eye view of the many different ML technologies that are available. There is not enough time to cover all the tools in use today. Finally, the course assumes a good working knowledge of the Python programming language at the start. Online courses in Python may be acceptable to meet this requirement.
Learning Goals

1. Basic data processing and handling with Python/Pandas
2. Machine learning tools available in Scikit Learn
3. Implementation of machine learning algorithms
4. Testing and evaluating forecasts/predictions (cross validation)
5. Presenting/describing results (graphics)

Prerequisites:

1. ECON213a/ECON184a (equivalent to most undergrad 1 semester classes in econometrics)
   1. Random variables, expectations, PDF’s, CDF’s
   2. Linear regression (Ordinary least squares)
   3. Intro to machine learning topics: Ridge and Lasso regression
   4. Logistic regression
2. Bus215f: Python for Business and finance or good working Python knowledge
   1. Knowledge of: Numpy, SciPy is useful
   2. We will also be using: matplotlib, and pandas, but we'll go over these
   3. FIN285a is another course covering this material
   4. Online options such as Datacamp may be useful
3. Basic matrix algebra is not required, but is useful (know what a matrix is, and the rules for matrix multiplication)

Required Readings:


Optional books:

1. (G) Geron, Hands-on Machine Learning with Scikit-Learn, Keras and Tensorflow, O'Reilly, 2019, second edition. (This book covers a few more topics than our current book, especially neural networks. It may be the textbook in Fall 2020.)
2. (M) McKinney, Python for Data Analysis: Data Wrangling with Pandas, Numpy, and IPython, O'Reilly, 2017, second edition. (This book is a must have for Python data analytic types. Covers all the necessary extensions to Python needed for data.)
3. (JWHT) James, Witten, Hastie, Tibshirani, An Introduction to Machine Learning, Springer, 2013. (A kind of easy to access overview of machine learning along with R code. We will refer to this a few times in the class. This book
provides a lot of technical **math foundations** which are not present in (MG.)

Available at [JWHT](#)


5. (HTF) Hastie, Tibshirani, Friedman, *The Elements of Statistical Learning: Data Minining, Inference, and Prediction*, Second edition, Springer, 2017. (The mathematical core of machine learning. Available online as a pdf file. The level of technical rigor of this book is well beyond this course, but if you need more, this is the place to go.) [HTF](#)

Occasionally, the notes will give pointers to these readings. They will use the initials. The pointers to the primary text (MG) are always given. (JWHT) will refer to the Ja This is optional, but may be useful to have. (PDF is online.) (M) Again, this book is very useful for any data analytics serious perso

---

**Required Software**

1. Python 3.7 and the entire Anaconda suite of tools. (This is open source and runs on all major operating systems.)
2. **Textbook code and workbooks**

**Data resources**

1. **Machine learning repository**
2. [Kaggle](#)

**Grading**

Grades will be based on:

1. Problem sets (20%)
2. Midterm exam (35%): Monday, April 6th, in class
3. Final exam (45%): Monday, May 4th, 1:30-4:30pm (standard final exam slot)

---

**Rules and responsibilities**
**Communications**

You are responsible for all announcements and materials in class. Also, much of the information in class will be sent over Latte and the class website.

**Rules specific to Bus241f**

- **Exams**
  - Your own work.
  - Closed book (no notes).
  - No laptops, no cell phones, no calculators, no pda’s.

- **Problem sets**
  - Hand in your own work.
  - Can talk and assist each other.
  - Use all resources.

- **Group projects**
  - Own work for the group.
  - Hand in one writeup per group.

- **Laptops**: Please bring to class if you want to.

**Academic Integrity**

You are expected to be honest in all of your academic work. Please consult Brandeis University Rights and Responsibilities for all policies and procedures related to academic integrity. Students may be required to submit work to TurnItIn.com software to verify originality. Allegations of alleged academic dishonesty will be forwarded to the Director of Academic Integrity. Sanctions for academic dishonesty can include failing grades and/or suspension from the university. Citation and research assistance can be found at LTS – Library guides.

**Work Load**

Success in this two-credit course is based on the expectation that students will spend a minimum of 9 hours of study time per week in preparation for class (readings, papers, discussion sections, preparation for exams, etc.)
Disability Statement

Brandeis seeks to welcome and include all students. If you are a student who needs accommodations as outlined in an accommodations letter, please talk with me and present your letter of accommodation as soon as you can. I want to support you. In order to provide test accommodations, I need the letter more than 48 hours in advance. I want to provide your accommodations, but cannot do so retroactively. If you have questions about documenting a disability or requesting accommodations, please contact Student Accessibility Support (SAS) at 781.736.3470 or access@brandeis.edu. If you are a student with a documented disability on record at Brandeis University and wish to have a reasonable accommodation made for you in this class, please see me immediately.

Fall calendar dates

- First day of classes: Jan 13
- Last day of classes: Apr 29
- Final exams: May 4–12
- No class:
  - January 20
  - February 17–21
  - April 8–16
- Brandeis days
  - April 7: Thursday schedule
- Module dates:
  - Module II: Starts, March 11, ends April 29th

Course Outline

1. Introduction: What is Machine Learning? What is Artificial Intelligence?
2. Landscape of problems
   1. Supervised versus unsupervised learning
2. Classification versus forecasting
3. Time series/ cross section
4. Predictive modeling/ policy intervention

3. Python basics (very short) (MG chapter 1: 1–12), (M, chapter 4, 8)
   1. Anaconda
   2. Spyder
   3. Numpy, Scipy
   4. Matplotlib
   5. Scikit Learn
6. Knowledge in this section assumes information in McKinney, 2nd edition, in the following sections:
   1. Chapter 1
   2. Chapter 2
   3. Chapter 3
   4. Chapter 4

7. Other chapters in the book are useful, but not required:
   1. Plotting: chapter 9
   2. Data/Pandas: chapter 6–8
   3. Time series: chapter 11

4. Model complexity (MG pp 27–32)
   1. Overfitting (figure 2–1)
   2. Training/validation/testing

5. Simplest models (MG pp 32–51, MG skip chapter 1)
   1. Getting started with scikit-Learn
   2. K nearest neighbors (MG pp 37–46)
   3. Linear regression (MG pp 47–51)

6. Controlling model complexity
   1. Regularization
   2. Ridge regression (MG pp 51–55)
   3. Lasso regression (MG pp 55–58)
   4. The Bias/Variance trade off

7. Classification
   1. Logistic regression (MG 58–65)
   2. Linear discriminant
3. Support vectors (SVC) (MG 58–65)
4. Naive Bayes (MG 70–72, skim)
5. Multi(k>2) class problems (MG 65–68)

8. Decision trees (MG chapter 2, pp 72–85)
   1. Implementation
   2. Feature importance
   3. Time series example
   4. Forecast variable usage

9. Model combinations (MG chapter 2, pp 85–94)
   1. Bagging predictors
   2. Random forests
   3. Boosting

10. Nonlinear models and kernels (MG chapter 2, pp 94–106)
    1. Kernalized support vector machines
    2. Kernel ridge regression

11. Predictive objectives (MG chapter 5, pp 277–305)
    1. Classification issues
    2. Evaluation metrics
    3. ROC curves and AUC

12. Neural networks and deep learning (MG chapter 2, pp 106–121)
    1. Activation functions
    2. Tuning networks
    3. Regularization

    1. Tool overview
    2. K–means clustering
    3. Evaluation

14. Advanced topics (may not get here)
    1. Text analysis/sentiment (MG pp 325–336)
    2. Recommender systems
    3. Measuring classifier uncertainty (MG pp 121–129)