Syllabus ¶

Bus241f: Machine Learning and Data Analysis for Business and Finance ¶

Key information ¶

Instructor ¶

- Blake LeBaron
- blebaron@brandeis.edu
- http://www.brandeis.edu/~blebaron
- Sachar 204

- Office hours:
    - Tuesday 2:00-3:00pm
    - Wednesday 12:30-1:30pm

TA ¶

- Mohammed AlMehdar
- mmehdar@brandeis.edu
- Office hours: TBA

Times: ¶

Class Times: Monday/Wednesday 5:00-6:20 (Module 2 starts Wednesday Oct 24, No class on Monday, Oct 29)

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 in both cross sectional and time series environments. 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 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, can be concurrent)
    1. Random variables, expectations, PDF's, CDF's
    2. Linear regression (Ordinary least squares)
2. Bus215f: Python for Business and finance or good working Python knowledge
        1. Knowledge of : Numpy, SciPy, and Pandas
        2. FIN285a is another course covering this material
        3. Online options such as **Datacamp** may be useful
3. Basic calculus (about 1 semester, undergrad level)
4. Basic matrix algebra (know what a matrix is, and the rules for matrix multiplication)

Required Readings: ¶

1. (MG) Muller and Guido, *Introduction to Machine Learning with Python: A guide for Data Scientists*, O'Reilly, 2017.

Optional books: ¶

1. (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.)
2. (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).)
3. (EH) Efron/Hastie, *Computer Age Statistical Inference*, Cambridge University Press, 2016. (A more recent book on computational methods in modern statistics. Not totally machine learning, but very useful to many of the methods we will talk about in this class. I will refer to it occasionally.)
4. (HTF) Hastie, Tibshirani, Friedman, *The Elements of Statistical Learning: Data Minining, Inference, and Prediction*, Springer, 2009. (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. While it is complete, it is getting a little dated.)
5. There are new books on machine learning appearing everyday, and I may add some other recommendations as the semester proceeds.

```
Occasionally, the notes will give pointers to these readings.
They will use the initials.  The pointers to
the primary text (MG) are always given.
```

Required Software ¶

1. Python 3.6 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. Final exam (50%): Exam week, December 13-20th
3. Group project (30%)

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 ¶

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 ¶

- Wednesday, Aug 29: First class
- Monday, September 3: No class (holiday)
- Thursday, September 6, Brandeis Monday
- Mon/Tue, Sept 10/11: No class (holiday)
- Wed, Sept 19: No class (holiday)
- Mon, Sept 24: No class (holiday)
- Tuesday, Sept 25: Brandeis Monday
- Monday, Oct 1: No class (holiday)
- Wednesday, October 24th: Module 2 Wed classes start
- Monday October 29th, no module 2 classes
- Wed-Friday, Nov 21-23: No class (holiday)
- Monday, December 10: Last class
- Final exams, December 13-20th

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. Classifying data sets: Tall, wide, and dense data
    5. 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

4. Supervised learning (skim MG chapter 1)

    1. Model complexity (MG 27-32)

        1. Overfitting (figure 2-1)
        2. Training/validation/testing

    2. Simplest models (MG 32-51)

        1. Getting started with scikit-Learn
        2. K nearest neighbors
        3. Application: Iris data (MG 13-25)
        4. Linear regression

    3. Controlling model complexity (MG 51-58)

        1. Regularization
        2. Ridge regression
        3. Lasso regression

    4. Classification

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

    5. Decision trees (MG 72-94)

        1. Controlling complexity again
        2. Feature importance and reading trees
        3. Bagging predictors
        4. Multiple trees (The Random Forest)
        5. Boosting

    6. Nonlinear models and kernels (MG 94-106)

        1. Kernelized support vector machines
        2. The kernel trick
        3. Kernel ridge regression

5. Predictive objectives (MG chapter 5)

        1. Classification issues
        2. Evaluation metrics
            1. More cross-validation
6. Neural networks and deep learning (MG pp 106-121)
    1. Activation functions
    2. Tuning networks
    3. Regularization
7. Unsupervised learning
    1. Tool overview
    2. K-means
    3. Evaluation
8. Advanced topics (may not get here)
    1. Text analysis/sentiment
    2. Recommender systems
    3. Measuring classifier uncertainty (MG pp 121-129)

Navigation