Advanced Data Analysis Methods

Johanna Devaney – johanna.devaney@brooklyn.cuny.edu

DATA 71200 (61133): Wednesdays, 4:15 - 6:15 PM, Rm. 3212

Course Description

This course will provide you with skills necessary to apply machine learning techniques to data, and interpret and communicate their results. You will also begin to develop intuitions about when machine learning is an appropriate tool versus other statistical methods. This course will cover both supervised methods (e.g., k-nearest neighbors, naïve Bayes classifiers, decision trees, and support vector machines) and unsupervised methods (e.g., principal component analysis and k-means clustering). The supervised methods will focus primarily on "classic" machine learning techniques where features are designed rather than learned, although we will briefly look at recent deep learning models with neural networks. This is an applied machine learning class that emphasizes the intuitions and know-how needed to get learning algorithms to work in practice, rather than mathematical derivations. The course will be taught in Python, primarily using the scikit-learn library.

Course Objectives

By the end of the course, you will be able to

- articulate the main assumptions underlying machine learning approaches
- demonstrate the basic principles of dataset creation
- articulate the importance of data representations
- evaluate machine learning algorithms
- articulate the difference between supervised and unsupervised learning
- apply a range of supervised and unsupervised learning techniques

Grade Breakdown

Class Participation	10%
Datacamp Assignments	25%
Project 1: Data set creation	15%
Project 2: Supervised learning	15%
Project 3: Unsupervised learning	15%
Final paper	20%

Required Text

- Guido, Sarah and Andreas C. Müller. (2016). *Introduction to Machine Learning with Python*, O'Reilly Media, Inc. [IMLP]

Recommended Texts

- Géron, Aurélien. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow'* O'Reilly Media, Inc. [HOML] *The first edition of the book is available online at* *https://www.lpsm.paris/pageperso/has/source/Hand-on-ML.pdf*
- Hastie, Trevor, Jerome H. Friedman, and Robert Tibshirani. (2009). *The Elements of Statistical Learning*, Springer-Verlag New York. [TESL] *The book is available online at* *https://web.stanford.edu/~hastie/Papers/ESLII.pdf*

Class Website

Class-related information, including course schedule, assignment details, resources, and a copy of this syllabus, is available at `https://data71200sp20.commons.gc.cuny.edu/`.

Datacamp

You will be assigned a number of tasks to complete on Datacamp throughout the course. The link to the course's Datacamp page will be emailed to you.

GitHub

Class notes will be available on GitHub (`https://github.com/jcdevaney/data71200sp20`). You will each be required to create your own GitHub repository for the class to host your projects and final paper.

Grade Component Details

Class Participation: 10%

The participation grade is a combination of attendance (including arriving on time); attentiveness, engagement, and participation during class; and general preparedness for class discussions.

Datacamp Assignments: 25%

These projects are hands-on activities designed to both provide coding background and reinforce the concepts covered in class.

Project 1 (Dataset creation): 15%

Curation and cleaning of a labeled data set that you will use for the supervised and unsupervised learning tasks in project 2 and 3. The dataset can built from existing data and should be stored in your GitHub repostiory.

Project 2 (Supervised learning): 15%

Application of two supervised learning techniques on the dataset you created in Project 1. This assignment should be completed as a Jupyter notebook your GitHub repository.

Project 3 (Unsupervised learning): 15%

Application of two unsupervised learning techniques on the dataset you created in Project 1. This assignment should be completed as a Jupyter notebook your GitHub repository.

Final Paper: 20%

A 5–8 page paper describing the work you did in projects 1–3 (your dataset and your supervised and unsupervised experiments). The paper should describe both what you did technically and what you learned from the relative performance of the machine learning approaches you applied to your dataset. This assignment should be posted as a PDF in your GitHub repository.

Course Schedule

Date	Topic	Readings Due
29-Jan	Introduction	
5-Feb	What is Machine Learning?	[1] Ch 1: "The Machine Learning Landscape" [HOML, 1–31] [2] Jordan, Michael I. and Tom M. Mitchell. (2015). "Machine Learning: Trends, perspectives, and prospects" *Science* 349, 255—-60. `http://www-cgi.cs.cmu.edu/~tom/pubs/Science-ML-2015.pdf`
12-Feb	No class	
19-Feb	Getting Started with ML	Ch 1: Introduction [IMLP, 1–25]
26-Feb	Inspecting Data	Ch 2: End-to-End Machine Learing Project [HOML, 33–66]
4-Mar	Representing Data	Ch 4: Representing Data/Engineering Features [IMLP, 213–55]
11-Mar	Evaluation Methods	Ch 5: Model Evaluation [IMLP, 257–310]

Course Schedule (con't)

18-Mar	Supervised Learning	Ch 2: Supervised Learning (k-Nearest Neighbors, Linear Models) [IMLP, 27–46] – *Project 1 Due*
25-Mar	Supervised Learning	Ch 2: Supervised Learning (Naïve Bayes Classifiers and Decision Trees) [IMLP 47–93]
1-Apr	Supervised Learning	Ch 2: Supervised Learning (Support Vector Machines and Uncertainty estimates from Classifiers) [IMLP 93—106, 121–31]
7-Apr	Unsupervised Learning *Conversion Day*	Ch 3: Unsupervised Learning (Dimensionality Reduction Feature Extraction, and Manifold Learning) [IMLP, 133-170] – *Project 2 Due*
8-Apr	No class	
15-Apr	No class	
22-Apr	Unsupervised Learning	Ch 3: Unsupervised Learning (Clustering) [IMLP, 170-211
29-Apr	Deep Learning	[1] Ch 2: Neural Networks/Deep Learning [IMLP 106–21] [2] Ch 10: Introduction to Artificial Neural Neetworks [HOML, 253-273]
6-May	Ethics	Bostrom, Nick, and Eliezer Yudkowsky. (2014). "The ethics of artificial intelligence." *The Cambridge Handbook of Artificial Intelligence.* 316–34. `http://faculty.smcm.edu/acjamieson/s13/artificialintelligence.pdf`- *additional readings may be added – Project 3 Due*
13-May	Ethics	West, Sarah Myers, Meredith Whittaker, and Kate Crawford. (2019). "Discriminating systems: Gender, race and power in AI." AI Now Institute, 1–33. `https://ainowinstitute.org/discriminatingsystems.pdf` - *additional readings may be added*
20-May	Project Due	

Important Dates

Monday, January 27	First day of Spring 2020 classes
Sunday, February 2	Last day to add a course
Wednesday, April 1	Last day to withdraw from a Spring course with a "W" grade
Tuesday, April 7	Conversion Day – Classes follow a Wednesday Schedule
Friday, May 15	Reading Day
Saturday, May 16	Final Examinations Begin
Friday, May 22	Final Examinations End / End of Spring Term