Harvard University Extension School

Data Literacy in the Age of Machine Learning (MGMT E-5072), 2018 Fall Term
Web Conference: Thursdays, 5:30 – 7:30 pm ET, September 6 – December 13, 2018

Instructor: Dr. Jitendra Subramanyam (jsub@gartner.com) [Please include "E-5072" in the subject line]
Teaching Assistant: Mr. Sasha Karimi (sahandskarimi@gmail.com)

Course Description

This course is a practical introduction meant to help business executives understand key concepts and techniques in data science and immediately apply them to business problems. It is not for engineering or computer science students seeking to learn the theoretical (and mathematical) underpinnings of machine learning.

The course is divided into four parts.

- *Part 1: The Mechanics of Prediction*. In Part 1 we'll dive right into machine learning, unpacking the key concepts (spoiler: there are just a few and they're simple) and demystify what really happens when machines learn. We'll apply these concepts to make *predictions* from real datasets. We'll cover the basic techniques of machine learning – regression and logistic regression – and get a feel for the practical things that data scientists do. We will cover the various models in machine learning – Clustering, Dimensionality Reduction, Support Vector Machines, Neural Networks, Decision Trees, and Ensemble models – and learn why we need this variety.
- *Part 2: The Science of Machine Learning*. In Part 2 we'll learn to systematically evaluate the performance of machine learning models. We'll understand how to define performance and measure it. We'll use this knowledge to not only build the right machine learning models but build them right.
- *Part 3: The Art of Machine Learning*. In Part 3 we'll tackle the art of constructing the right features for our models to get the most predictive bang for our data buck.
- *Part 4: Select Topics in Machine Learning*. Finally, in Part 4 we'll cover select topics in machine learning such as ensemble models, spam filters, learning from images and text, and the promise of deep learning.

Outline of the Syllabus

1. The Mechanics of Prediction
    a. Introduction to machine learning
        i. What's different about machine learning?
        ii. How does it work?
        iii. The types of machine learning
        iv. The business value of machine learning
    b. Predicting a numerical value
        i. Linear regression (using single and multiple features)
        ii. Non-linear regression (using single and multiple features)
    c. Measuring the performance of regression models
        i. Training and Test data sets
        ii. Performance measures
    d. Predicting a categorical value – part 1
        i. Logistic regression (using single and multiple features)
        ii. Non-linear logistic regression (using single and multiple features)
    e. Predicting a categorical value – part 2
        i. Support Vector Machines
        ii. Neural Networks
        iii. Decision Trees
        iv. Random Forests
    f. Measuring the performance of logistic regression models. Precision, recall, and accuracy.

2. The Science of Machine Learning
    a. Preparing data for prediction
        i. Data pre-processing and cleaning
        ii. Systematic visualization of data sets
        iii. Systematic investigation of data sets
    b. An overview of models
        i. The menu of different machine learning models, why we need them, and when to use them. (The problem of having enough data; the no-free-lunch theorem.)
        ii. Overfitting and underfitting; what they are and how to handle them.
        iii. Comparing rule-based, statistical, and machine learning-based approaches to prediction.
    c. Finding the optimal hyper-parameters of a model
        i. Models, parameters, hyper-parameters
        ii. Training, validation, and test datasets
        iii. Model complexity
        iv. Tuning a model's hyper-parameters (including k-fold cross validation)
    d. Measuring and improving model performance
        i. Precision, recall and other measures of prediction capability (again)

ii. Bias and variance
iii. Learning curves to determine model bias and variance
iv. Techniques for improving model performance

3. The Art of Machine Learning
   a. Feature selection
   b. Feature engineering

4. Select Topics in Machine Learning
   a. Unsupervised models -- similarity and clustering.
   b. The surprising power of ensemble models
   c. Building a spam filter using a Naïve Bayes model
   d. Learning from images and text
   e. The promise of deep learning

Prerequisites

This course does not contain any advanced mathematics. If you've taken the SAT or the GRE you've already come across mathematics that is much more advanced than anything you will need for this course. Alternatively, if you're comfortable working with spreadsheets (nothing fancy, just basic formulas and manipulations like sorting rows), you will be comfortable with all of the mathematics used in this course. Hands-on learning is encouraged using the *Orange* data science platform (https://orange.biolab.si/) – a visual way to solve machine learning problems without programming. For those with some programming knowledge of Python, we provide *Jupyter* notebooks that can be used to build, run, and experiment with machine learning models. Please note that Python knowledge is NOT a prerequisite for the course. The coursework DOES NOT require any Python programming.

Learning Objectives

By the end of this course you will be able to:

- List the types of problems that can be solved using machine learning.
- Understand the seven key steps to solving any machine learning problem.
- Apply machine learning techniques such as regression and classification to solve a variety of business problems using real-world data.
- Build strong intuitions about machine learning techniques by implementing them in a hands-on interactive environment that requires no programming.
- Determine efficient and effective ways to improve the predictions generated by machine learning models.
- Collaborate productively with your data science team by speaking the language of data and predictive modeling.

- Keep up with the rapidly progressing fields of machine learning and Artificial Intelligence.

Course Materials

This course requires students to work continually throughout the semester. It entails a fair amount of reading, working with data, reflection and discussion. Required readings will consist of a variety of blog posts and other articles/videos that can be accessed over the internet. You will also need to download and install the software to run the Orange machine learning platform on your local computer (https://orange.biolab.si/). All the materials required for this course are free.

Listed below is an optional book for the course. It is NOT required for the course but can serve as an alternative source for learning about the topics we cover in this course. The book can be purchased from many bookstores, including the Harvard Coop and online booksellers. You can also access the book online with your Harvard library credentials.

Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking
Provost, Foster and Fawcett, Tom (O'Reilly Media, Inc.)
ISBN: 978-1-36132-7

Grading

A student's final grade in this course will be based on the following weighting:

60%	Homework Assignments
20%	Group Assignments (details will be provided during the course of the term)
20%	Final Exam

Grades reflect the quality of a student's work submitted throughout the term according to the Harvard Extension School's grading standards (http://www.extension.harvard.edu/exams-grades-policies/grades).

This is a graduate-level course with graduate-level work. We expect active participation in class discussions and activities and high-quality written work. Much of a manager's success depends on communication; therefore, effective written communication on the homework assignments and the final exam will is essential. Written work should be clear, logical, grammatically correct, spell-checked, persuasive, supported by examples, and backed up by citations for any data, ideas or other content used. It should represent the student's best effort.

Please note that all homework assignments are due in the relevant course assignment folder (on the course Canvas website) at or before the deadline date and time. Late assignments will not be accepted.

Assignments

Homework Assignments
These assignments will vary in length and content. However, their objective is always the same – to ensure that you get a working knowledge of the material covered and enable you to explore topics that are not covered explicitly during class.

Group Assignment
Think of this as a substantial homework assignment. You'll work in collaboration with a group of your classmates to solve a problem from start to finish using a machine learning model. The assignment will give you a chance to exercise a majority of the techniques and concepts you've learned in the course.

Final Exam
The final exam will be a 2-hour exam that includes all the topics covered in the course lectures and the readings assigned in the course.

Academic Integrity

You are responsible for understanding Harvard Extension School policies on academic integrity (www.extension.harvard.edu/resources-policies/student-conduct/academic-integrity) and how to use sources responsibly. Not knowing the rules, misunderstanding the rules, running out of time, submitting the wrong draft, or being overwhelmed with multiple demands are not acceptable excuses. There are no excuses for failure to uphold academic integrity.

To support your learning about academic citation rules, please visit the Harvard Extension School Tips to Avoid Plagiarism (www.extension.harvard.edu/resources-policies/resources/tips-avoid- plagiarism), where you'll find links to the Harvard Guide to Using Sources and two free online 15-minute tutorials to test your knowledge of academic citation policy. The tutorials are anonymous open-learning tools.

Additional Information

Workload. The value you receive from this course will be commensurate with the thought and effort that you put into the endeavor. Students should expect to spend 4-8 hours outside of class each week to read the assigned materials, reflect, complete assignments, and prepare for the next class session. More time will be required to do the team project.

On Time. Students are expected to arrive to the online classroom on time and stay for the duration of the class session. If you expect to be late or absent from class – or need to leave early – let the instructor know prior to the start of class.

Deadlines. All assignments must be submitted to the correct assignment drop box on the Canvas course website by the specified day and time and late submissions will not be accepted. If you experience any problems uploading your assignment to the Canvas drop box, you should email the document to the TA for this course, Sasha Karimi. Please note that, if you email the assignment because you cannot upload it, the email and the relevant attachment must be received on or before the assignment deadline to be accepted. Should you experience any internet problems, please call/leave a message for the teaching assistant (Sasha Karimi) – this call should occur *before* the submission deadline passes. If you are absent the day an assignment is due, the assignment is still due at the specified day and time. True medical or family emergencies will be dealt with on a case by case basis.

Professional Conduct. Professional behavior is expected throughout the class. This means respectful communication both inside and outside of class. During discussions, civil discourse should be maintained at all times and comments should be aimed at moving the discussion forward. This does not mean that students must always agree with others since reasoned, respectful dissention may be part of the discovery process and lead to previously unconsidered options.

Disability Services. The Extension School is committed to providing an accessible academic community. The Accessibility Office offers a variety of accommodations and services to students with documented disabilities. Please visit www.extension.harvard.edu/resources-policies/resources/disability-services- accessibility for more information, or contact the Accessibility Services office at Accessibility@dcemail.harvard.edu or (617) 495-4024.

Syllabus

Session	Date	Topics	Read and Do BEFORE the Class Session	Deadlines (All Dates/Times are US ET)
1	September 6	**Introduction to Machine Learning** ▪ Overview of the course ▪ A tour of Orange ▪ Visualizing data using Orange	▪ Complete pre-course survey (in Canvas) ▪ Make sure that you have a working headset for the online sessions ▪ Read through course syllabus in its entirety ▪ View Frank Chen's video, "The Promise of Machine Learning" (https://vimeo.com/215926017) ▪ View John Launchburg's video, A DARPA Perspective on AI (https://www.youtube.com/watch?time_continue=5&v=-O01G3tSYpU)	**Pre-Course Survey (Due by 10 am on September 8)**

2	September 13	**Hands On - The Nuts and Bolts of Machine Learning** ▪ How machine learning really works ▪ How it compares with other approaches	▪ Download and start up Orange (https://orange.biolab.si/download/) ▪ View the first 4 Orange training videos at https://www.youtube.com/watch?v=HXjnDIgGDuI&list=PLmNPvQr9Tf-ZSDLwOzxpvY-HrE0yv-8Fy ▪ Ensure that you have access to the Jupyter notebooks on GitHub	
3	September 20	***Predicting Numerical Values 1*** ▪ Regression with a single feature	▪ Read the HBR article, The Simple Economics of Machine Intelligence (https://hbr.org/2016/11/the-simple-economics-of-machine-intelligence)	**Homework Assignment 1 (Due by 10 am on September 20)**
4	September 27	***Predicting Numerical Values 2*** ▪ Regression with multiple features ▪ Non-linear regression	▪ Read the lecture notes on single-feature regression. ▪ Run a regression using Orange and a dataset of your choice.	
5	October 4	***Predicting Categorical Values 1*** ▪ Logistic regression with two features ▪ Logistic regression with multiple features ▪ Non-linear logistic regression	▪ Read the lecture notes on regression with multiple features. ▪ Read the lecture notes on non-linear regression. ▪ Read about the key types of data science projects to work on (https://www.dataquest.io/blog/build-a-data-science-portfolio/)	**Homework Assignment 2 (Due by 10 am on October 4)**
6	October 11	***Predicting Categorical Values 2*** ▪ Measuring the performance of regressors and classifiers ▪ Particularly useful models for non-linear logistic regression – Support Vector Machines, Neural Networks, Decision Trees	▪ Read the lecture notes on logistic regression. ▪ Run a logistic regressions (linear and non-linear) using Orange and a dataset of your choice.	
7	October 18	***The Science of Machine Learning 1*** ▪ Systematic visualization and investigation of data sets ▪ The variety of models for machine learning ▪ Overfitting and underfitting ▪ Rules-based versus machine learning-based approaches to prediction	▪ Read the lecture notes on measuring the performance of regressors and classifiers. ▪ Read Marc Andreesson's article, "This is Probably a Good Time to Tell You…" (http://blog.pmarca.com/2014/06/13/this-is-probably-a-good-time-to-say-that-i-dont-believe-robots-will-eat-all-the-jobs/) ▪ Read Kevin Kelly's article, "The Myth of Superhuman AI" (https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai/)	**Homework Assignment 3 (Due by 10 am on October 18)**
8	October 25	***The Science of Machine Learning*** 2 ▪ Training, validation and test datasets ▪ Models, parameters and hyper-parameters ▪ K-fold cross validation ▪ Validation curves	▪ Read the lecture notes on the science of machine learning, part 1. ▪ Select a dataset or two in coordination with your team. ▪ Determine the business problem(s) you want to investigate. ▪ Get a sense of how you'll use the data to solve the business problem(s).	**Homework Assignment 4 (Due by 10 am on October 25)**

		▪ Measuring model bias and variance ▪ Learning curves ▪ Systematic ways to improve model performance	▪ Peruse the list of AI tools for businesses available today https://medium.com/imlyra/a-list-of-artificial-intelligence-tools-you-can-use-today-for-businesses-2-3-eea3ac374835. It might give you some ideas for your group project.	
9	November 1	▪ Open time for each team to review progress on final project. Opportunity for early presentation feedback. ▪ Unsupervised Models: Similarity and Clustering	▪ Read the lecture notes on the science of machine learning, part 2.	**Group Assignment 1 (Due by 10 am on November 1)**
10	November 8	***The Art of Machine Learning*** ▪ Feature Engineering: Feature selection, extraction, or transformation ▪ Feature compression	▪ View and experiment with neural networks on the TensorFlow Playground (http://playground.tensorflow.org)	**Homework Assignment 5 (Due by 10 am on November 8)**
11	November 15	***Select Topics in Machine Learning 1*** • The surprising power of ensemble models • Building a spam filter	▪ Read the lecture notes on feature engineering. ▪ Read David Brook's article, "What Data Can't Do" (http://www.nytimes.com/2013/02/19/opinion/brooks-what-data-cant-do.html) ▪ Read Brynjolffson and McAfee's article, "The Business of Artificial Intelligence" (https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence) ▪ Coordinate with your team ▪ Start planning out your analysis	**Homework Assignment 6 (Due by 10 am on November 15)**
12	November 29	***Select Topics in Machine Learning 2*** • Learning from images • Learning from text	▪ Read the lecture notes on similarity and clustering. ▪ Read Richard Weiss' article, "Cargo Cult of Data Science" http://blog.richardweiss.org/2017/07/25/data-science-in-organizations.html ▪ Read Michelle Nijhuis' article, "How to Call BS on Big Data" (http://www.newyorker.com/tech/elements/how-to-call-bullshit-on-big-data-a-practical-guide) ▪ Start iterating on your group project presentations	**Homework Assignment 7 (Due by 10 am on November 29)**
13	December 6	***Select Topics in Machine Learning 3*** • The promise of deep learning • Exam review	▪ Read the lecture notes on learning from text. ▪ View Frank Chen's video, "AI, Deep Learning, and Machine Learning" https://vimeo.com/170189199 ▪ View Andrew Ng's video, "Nuts and Bolts of Applying Deep Learning" (https://www.youtube.com/watch?v=F1ka6a13S9I)	**Group Assignment 2 (Due by 10am on December 6)**
14	**December 13**			**Final Exam**