**Harvard University Extension School**

**Data Literacy in the Age of Machine Learning (MGMT E5072), 2018 Fall Term**

**Web Conference: Thursdays, 5:30 – 7:30 pm EDT, August 28 – December 13, 2018**

Instructor: Dr. Jitendra Subramanyam ([jsub@gartner.com](mailto:jsub@gartner.com))

Teaching Assistant: Mr. Sasha Karimi ([sahandskarimi@gmail.com](mailto: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 scaling machine learning systems and the promise of deep learning.

**Outline of the Syllabus**

1. The Mechanics of Prediction
   * Introduction to machine learning
     + What's different about machine learning?
     + How does it work?
     + The types of machine learning
     + What can we expect from machine learning?
   * Predicting a numerical value
     + Linear regression (using single and multiple features)
     + Non-linear regression (using single and multiple features)
   * Predicting a categorical value
     + Logistic regression (using single and multiple features)
     + Non-linear logistic regression (using single and multiple features)
   * Preparing data for prediction
     + Data pre-processing and cleaning
     + Systematic data visualization
     + Systematic investigation of data
   * Overview of different machine learning models, why we need them, and when to use them. SVMs, neural networks, decision trees, ensemble models.
   * Unsupervised models -- similarity and clustering.
   * Learning from images and text.
2. The Science of Machine Learning
   * Finding the optimal hyperparameters of a model
     + Models, parameters, hyperparameters
     + Training, validation, and test datasets
     + Model complexity
     + Tuning a model's hyperparameters
   * Measuring model performance
     + Precision, recall and other measures of prediction capability
     + Bias and variance
     + Learning curves to determine model bias and variance
     + Techniques for improving model performance
3. The Art of Machine Learning
   * Feature selection
   * Feature engineering
4. Select Topics in Machine Learning
   * Building a machine learning pipeline
   * 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.