Instructor:
Vu Dinh
Email: vucdinh@udel.edu
Office: Ewing Hall 312

Class Times: MWF 9:05am-9:55am, ISE Lab 222

Office Hours:

Tuesday 3pm-4pm and Wednesday 4:30pm-5:30pm, Ewing Hall 312, or by appointment.

Web page:

http://vucdinh.github.io/m637s19

Visit this page regularly. It will contain homework assignments, lectures, etc.

Prerequisites: Probability theory and basic statistics (e.g. MATH 350), Multivariable calculus (e.g. MATH 243), Linear Algebra (e.g. MATH 349), basic computing skills.

Optimization background (e.g. MATH 529) desirable but not necessary.

Reference: *An Introduction to Statistical Learning*, James, Witten, Hastie, and Tibshirani.

The pdf of the book is available for free at

http://www-bcf.usc.edu/~gareth/ISL/

Description: The course provides an introduction to the fundamental techniques used in data science. The main objective of the course is to develop a good practical knowledge and a mathematical understanding of the common tools that are used to analyze modern datasets. The course also provides hands-on experience in data analysis through practical homework and class projects.

Goals of the course:

- Become familiar with the basic methods used to analyze modern datasets.
- Understand the mathematical theory and the standard models used in data science
- Understand how to select a good model for data
- Be able to analyze datasets using a modern programming language such as Python or R

Topics:

- Analysis of the convergence and complexity of common algorithms
- Linear methods for regression (subset selection, ridge, lasso), logistic regression
- Linear discriminant analysis
- Support Vector Machines
- Cluster analysis (K-means, spectral clustering)

- Kernel Smoothing
- Principal component analysis
- Deep learning
- Cross-validation and Bootstrap
- Bayesian methods

Evaluation:

- Overall scores will be computed as follows:
  60% homework (theoretical + programming problems), 40% class project
- Here are the letter grades you can achieve according to your overall score. Subject to change (to your advantage only):

$\geq 95\%$ At least A

$\geq 90\%$ At least A-

$\geq 80\%$ At least B-

$\geq 70\%$ At least C-

$\geq 60\%$ At least D-

$< 60\%$ F

Data Analysis: We will use Python during the course. A good Python tutorial is available at

http://www.scipy-lectures.org/

The following may be useful to Matlab users:

http://mathesaurus.sourceforge.net/matlab-numpy.html

If you have never used Python before, I recommend using Anaconda Python 3.5

https://www.continuum.io/

It contains all the packages we will need.

Ethics: Please refer to UD's Guide to Academic Integrity

http://www.udel.edu/studentconduct/ai.html

In particular, note that copying solutions in whole or in part from other students or **any other source** without acknowledgement constitutes cheating. Any student found cheating risks automatically failing the class and will be referred to the Office of Student Conduct.