E178/ME276DS Statistics and Data Science for Engineers

[Syllabus]

In this course you will learn how to use data to build models of systems. We will start by learning basic concepts from probability theory (distributions, conditional probability, independence, Bayes' theorem, etc.) and optimization theory (local vs. global solutions, first order necessary conditions, convexity, gradient descent). With this in place, we will then learn some important techniques from statistics: point estimation, maximum likelihood, and confidence intervals. These techniques are important and useful in their own right, but they also serve as an introduction to the second half of the course, where we will learn to build models of closed-box systems with inputs. In that context we will learn classical techniques from statistics, such as linear and logistic regression, as well as more computationally intensive "machine learning" techniques, such as neural networks, decision trees, ensemble methods, and support vector machines.
Throughout the course, we will test our modeling techniques using real-world datasets and Python packages. You do not need to be an expert in Python to take this course. Homeworks will involve both programming and theoretical work, and are due every two weeks.
This course has a single midterm and there is no final exam. Instead, you will have a final group project, which you will begin to work on following the midterm. There will also be weekly quizzes on the lectures from the previous week.

Here is the list of topics that we will cover in this course.

1. Probability theory
   - Single random variables: expected value, variance
   - Multiple random variables: covariance, correlation, conditional probability, independence.
2. Optimization theory
   - Problem formulation
   - Optimality conditions
   - Gradient descent and stochastic gradient descent
3. Statistics
   - The sample mean
   - Point estimation
   - Maximum likelihood estimation
   - MLE for Mixture Gaussian models and the EM algorithm
   - K-means clustering
   - Confidence intervals
4. Supervised learning
   - Minimization of the expected loss
   - Performance evaluation
   - Hyper-parameter tuning
5. The regression problem and Linear regression
   - Simple linear regression (solution and statistical properties)
   - Multiple linear regression (solution and statistical properties)
6. The classification problem and Logistic regression
7. Machine learning models
   - Neural networks
   - Decision trees and random forests
   - Gradient boosting and Adaboost
   - Support vector machines