CMU ONLINE
Carnegie Mellon University

Syllabus Overview Sample: Gaining Insights through Statistical Modeling and Machine Learning

Program: Applied Data Science

Faculty: Julia Walchessen, Dietrich College

Course Number: 36641 **Units:** 6

Zoom Class: Weekday Evening, ET

Course Description

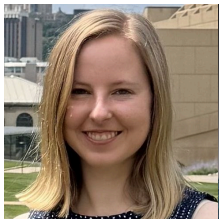
This course is all about the modeling of data, or in other words, using mathematical functions and computer algorithms to generate representations of the data-generating process. Statistical modeling and learning are elements that fall within the larger concept of artificial intelligence. Artificial intelligence is an idea—the idea that we can replicate human processes with computers—and data science provides the tools and methods to implement the idea. In this course, you will get hands-on experience performing exploratory data analysis and applying a suite of models, from linear or logistic regression through random forest, k-nearest neighbors, and support vector machine. You will generate reports akin to what we would write in real-life situations: we will be doing the actual work of data scientists.

Course Learning Objectives

By the end of this course, students will be able to:

- Explain the concepts of artificial intelligence, data science, and statistical modeling and learning (including machine learning) to others.
- Apply simple hypothesis tests to make statistical inferences.
- Describe what statistical models are and how to quantitatively assess how well they replicate the data-generating process.
- Apply optimization to estimate parameters in statistical models.
- Apply various metrics for assessing the performance of classification models.
- Apply a number of different commonly used machine learning models to data and interpret model output.

Faculty Introduction



I received a B.S. in mathematics and B.A. in statistics from the University of Chicago in 2020, and a Ph.D. in statistics from Carnegie Mellon University in 2025.

I work in the highly interdisciplinary area of spatial statistics---a field primarily focused on developing statistical methods for analyzing spatial data relevant to climate science, ecology, and epidemiology. My research interests lie primarily in adapting machine learning tools to enable classical statistical inference of complex, high-dimensional spatial data. Specifically, I use techniques from computer vision---a field in machine learning focused on processing, analyzing, and generating image data---because spatial data can often be viewed as image data. Through my career, I have traveled a decent amount including a year-long stay in Australia during my PhD and numerous conferences spanning four continents so far.

Course Topics

Topic	Learning Objectives
Exploratory Data Analysis	• Explain the concepts of artificial intelligence, data science, and statistical modeling and learning (including machine learning) to others. • Carry out exploratory data analyses on structured data. • Explain how to mitigate missing data. • Explain how to detect and mitigate anomalous data.
Dimension Reduction & Unsupervised Learning	• Use Principal Components Analysis to reduce the dimensionality of structured data. • Apply simple hypothesis tests to make statistical inferences given any column of data from a structured data table. • Apply commonly used unsupervised learning methods such as the K-means algorithm to detect distinct data groupings.
Linear Regression	• Describe what statistical models are and how to quantitatively assess how well they replicate the data-generating process. • Apply linear regression models to data with continuously valued response variables. • Explain why response variables may have to undergo numerical transformations, and how to affect those transformations. • Apply variable selection methods to data to improve model interpretability.
Extending the Ordinary Least	• Recognize and mitigate the effects that multicollinearity can have on statistical inference.

Squares Model	● Apply optimization to estimate parameters in statistical models. ● Utilize kernels to both estimate probability density functions and to learn local regression models. ● Apply statistical models to data that effectively capture data nonlinearities
Classification & Categorical Regression	● Apply various metrics for assessing the performance of classification models. ● Apply generalized linear models to two-category and multi-category data. ● Apply linear discriminant analysis and understand how it differs from logistic regression.
Machine Learning	● Explain how machine learning differs from classical statistical learning techniques like linear regression. ● Provide high-level algorithmic details about a number of commonly used machine learning models: trees, random forest, gradient boosting, support vector machine, and Naive Bayes. ● Apply a number of different commonly used machine learning models to data and interpret model output (to the extent possible).
Statistical Learning Projects	● Explain how deep learning fits within the overall context of statistical learning (optional). ● Appropriately implement methods of exploratory data analysis and statistical learning when analyzing data with either a quantitative or categorical response variable.

NOTE: This is a sample designed to provide an overview of topics covered in the course based on previous course syllabi. All students will be provided an updated and final syllabus at the start of the course. Faculty, content, and schedule are all subject to change.