Statistical Learning

Sravani Vadlamani

01/11/2022

About Me

- Assistant Professor in Data Science & Business Analytics
- Education
  - PhD in Civil Engineering
  - MS in Geographic Information Systems
  - MS in Civil Engineering
  - Certificate in Statistics

- Transportation Engineer II at Maryland State Highway Administration (SHA)

About Me

- Research Interests
  - Application of data mining techniques to Transportation, Transportation Safety, Travel Demand Management, Transportation Operations
- Courses Taught
  - Statistics
  - Statistical Learning
  - Advanced Quantitative Methods
  - Data Analytics for Smart Cities & Transportation
  - Intelligent Mobility
  - Six Sigma
  - Capstone

Student Introductions

- Name
- Major
- Share something good/new/exciting that you experienced during the break ☺

Course Description

- This is an introductory-level course in supervised learning. Topics include **classification** and **regression**, **cross-validation** and **bootstrap**, **model selection**, **dimension reduction**, **tree-based methods**, **random forests and boosting, support-vector machines, principal components, and cluster analysis**. Students will have hands-on experience in model building, machine learning, and implementation.

Class Overview

- Statistical Learning refers to a vast set of tools for understanding data.
- Supervised vs Unsupervised tools
  - Predicting or estimating an output based on one or more inputs – supervised
  - Learn relationships and structure from data with inputs but no outputs – unsupervised
- Problems of this nature can occur in fields such as business, medicine, astrophysics and public policy.

Class Overview

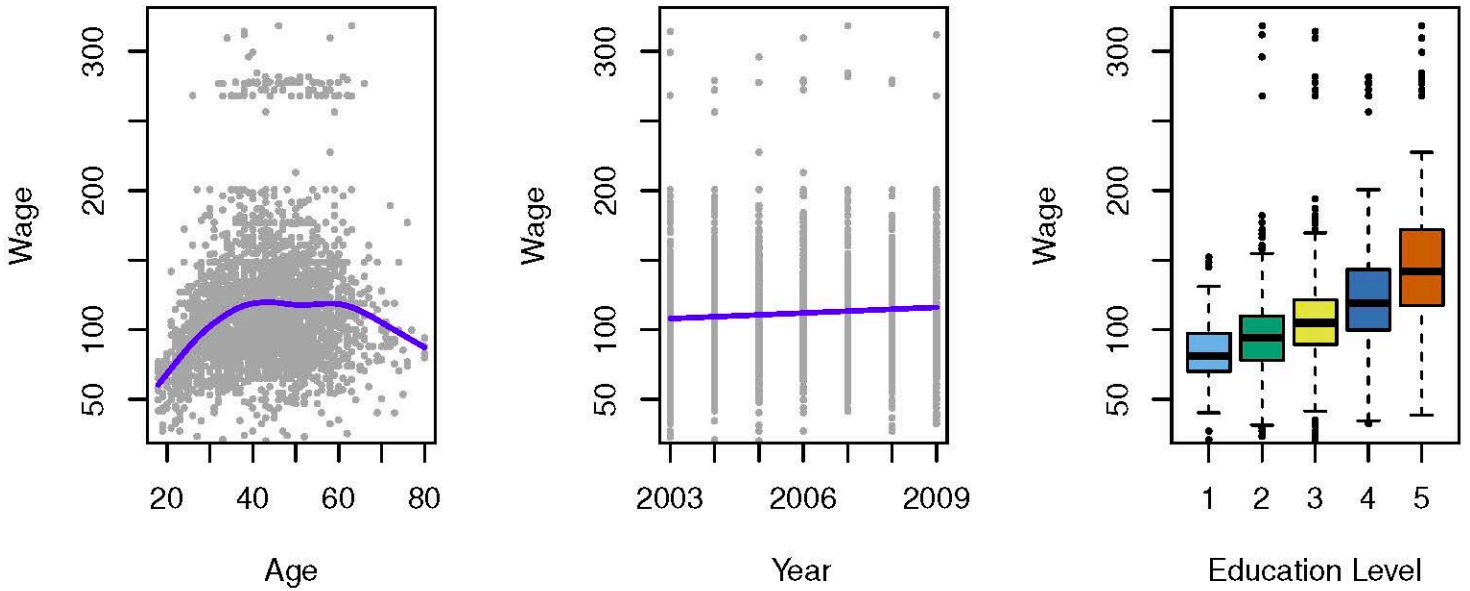
- Application of statistical learning methods to real-world problems.
- Many statistical learning methods are relevant and useful in a wide range of academic and non-academic disciplines, beyond just the statistical sciences.
- Statistical learning should not be viewed as a series of black boxes

Focus Areas

- Supervised vs Unsupervised Learning
- Regression vs Classification
- Bias-Variance Trade off
- Prediction Accuracy vs Model Interpretability
- Tree Based Methods
- Introduction to Deep Learning (if time permits)

Example 1: Wage Data

- Examine factors that affect wages for a group of men from the Atlantic Region of the US. We wish to understand the impact of employee's age, education and calendar year on wage.

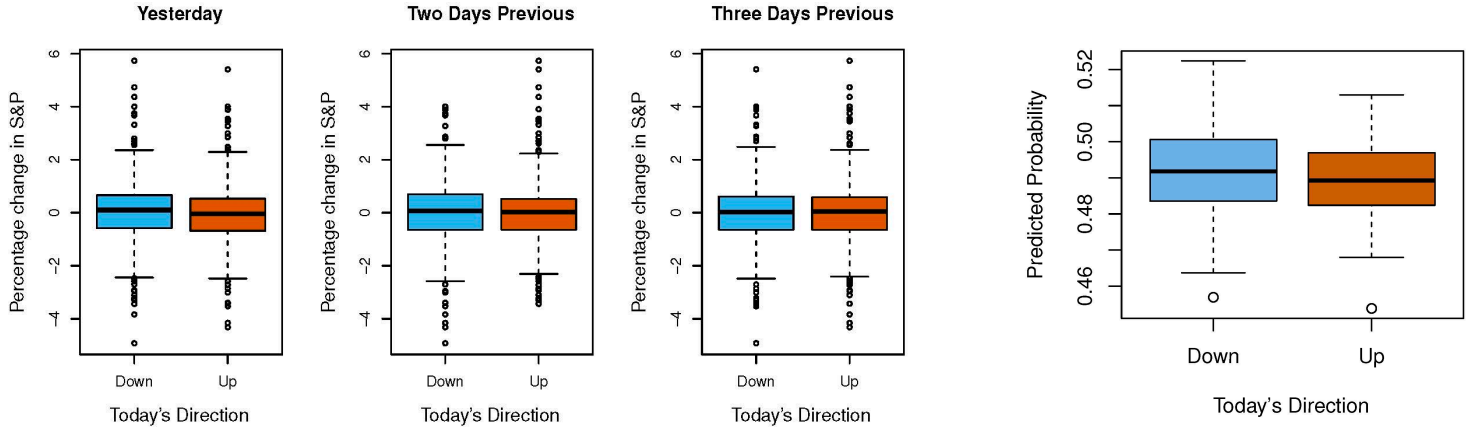


Chapter 3: Linear Regression to predict wage

Chapter 7: Non-linear model to address the relation between wage and age

Example 2: Stock Market Data

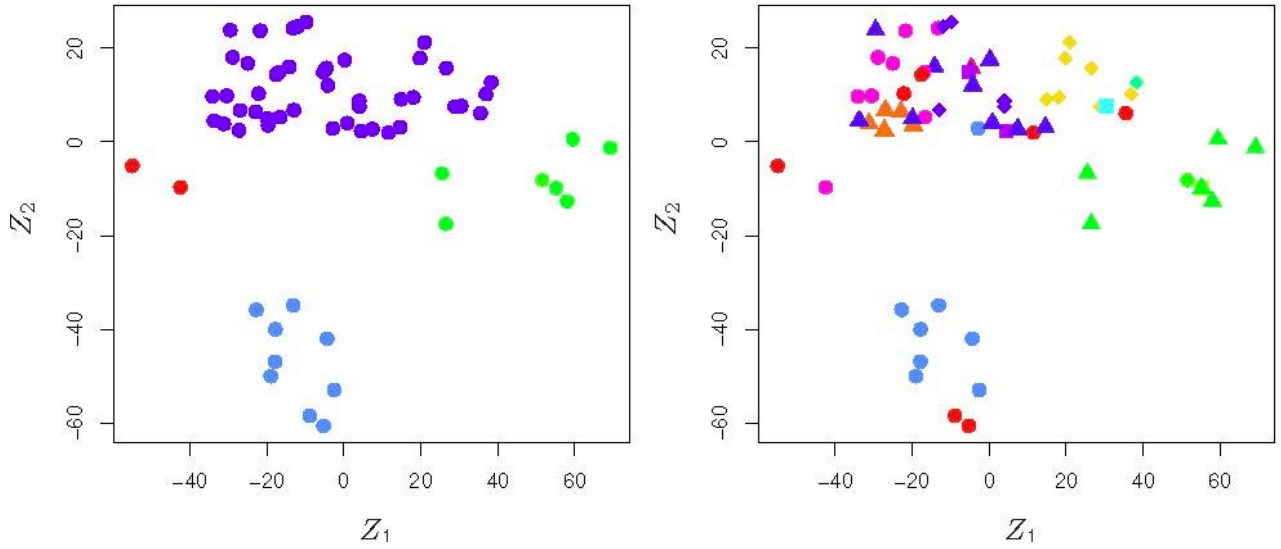
- Examine a stock market dataset that contains the daily movements in the Standard & Poor's 500 (S&P) stock index over a five-year period between 2001 and 2005.
- Goal: Predict whether the index will increase or decrease on a given day using the past 5 days' percentage change.



Chapter 4:
Classification Problem

Example 3: Gene Expression Data

- NCI60 dataset contains 6830 gene expression measurements for each of 64 cancer cell lines
- Goal: Determine if there are groups, or clusters based on their gene expression measurements.



Other Statistical Learning Problems

- Identify the risk factors for prostate cancer
- Predict if someone will have a heart attack based on demographics, diet history and clinical measurements
- Customize email spam detection system
- Identify numbers in a handwritten zip code
- Classify pixels in an image

Course Learning Outcomes

- Explain statistical learning methodology.
- Implement the techniques covered, interpret and understand results, and validate models.
- Monitor performance of ongoing implementations where appropriate.
- Effectively communicate the results of model implementation orally and in writing.

Required Materials

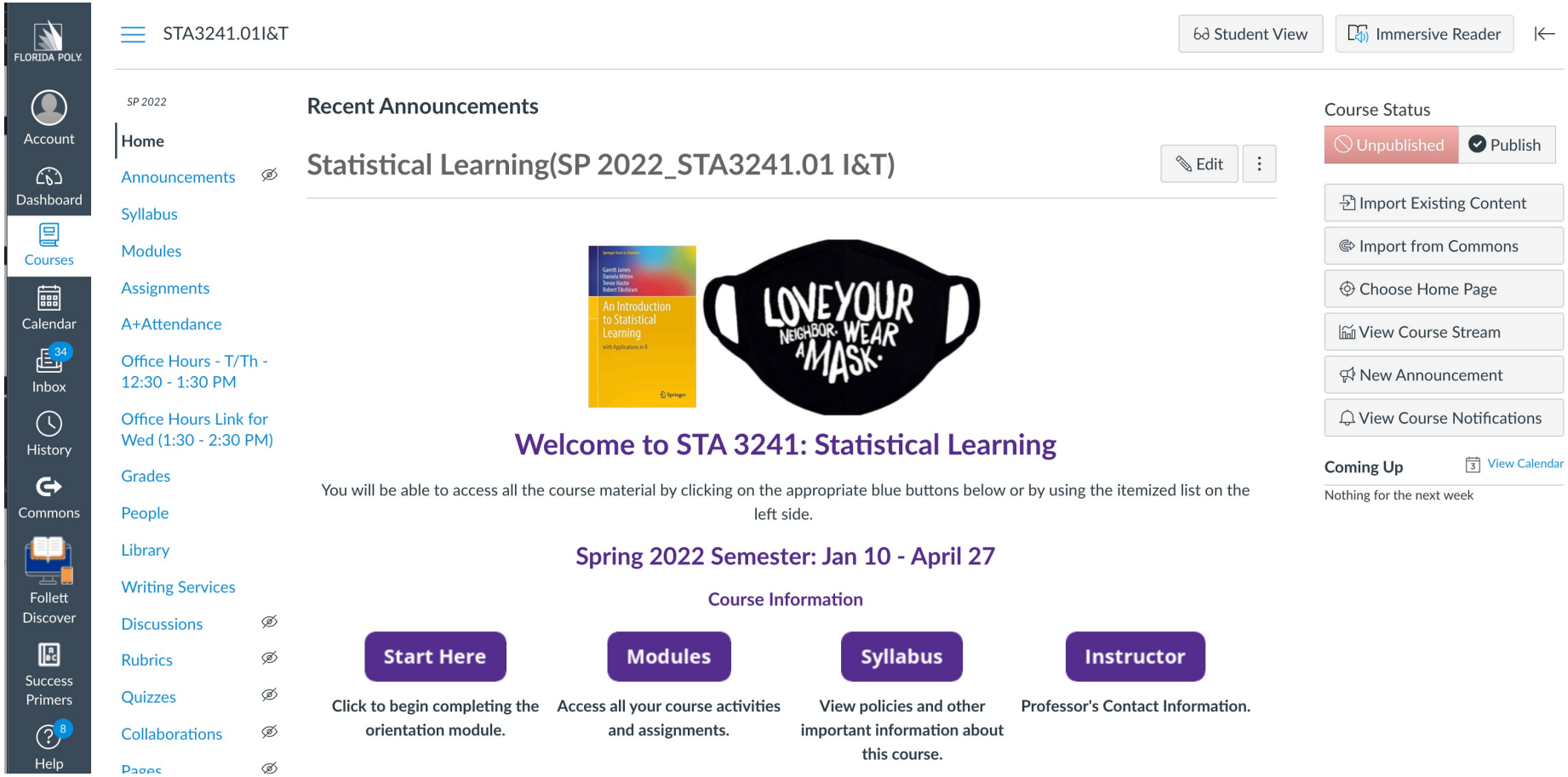
- (ISLR) Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. “An Introduction to Statistical Learning - with Applications in R” Second Edition. Available at: https://www.statlearning.com/ ISBN-13: 978-1071614174
- Supplementary materials: “R for Data Science” by Garrett Grolemund and Hadley Wickham accessible at https://r4ds.had.co.nz/

Grade Breakdown

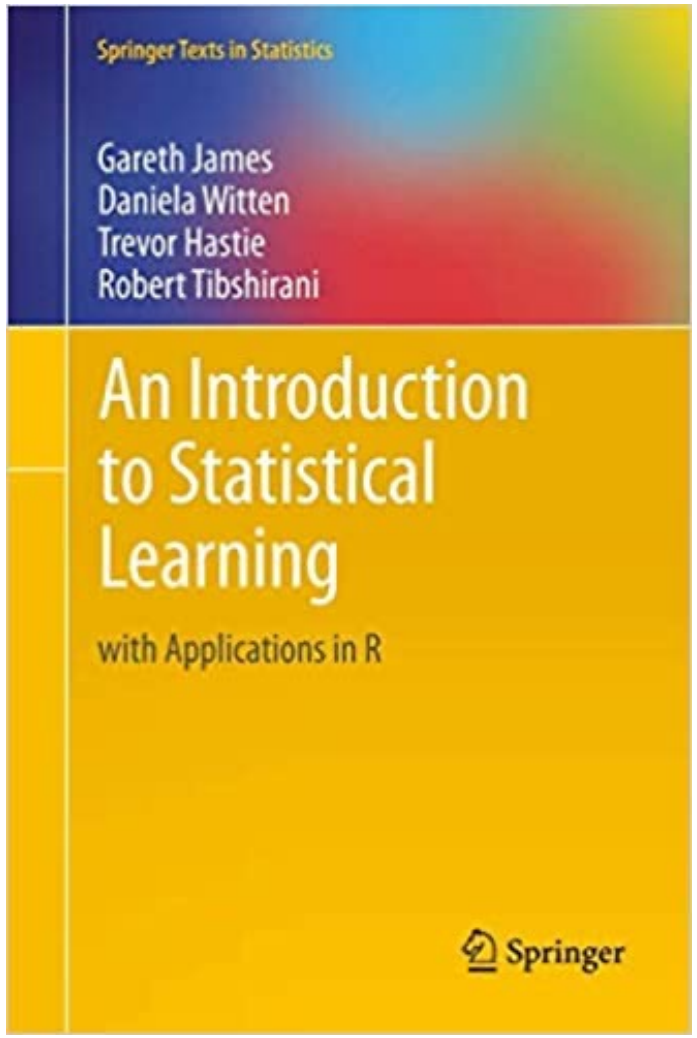
Attendance & Participation	10%
Assignments	20%
Quizzes	10%
Data Analysis Project	20%
Midterm Exam	20%
Comprehensive Final Exam	20%
Total	100%

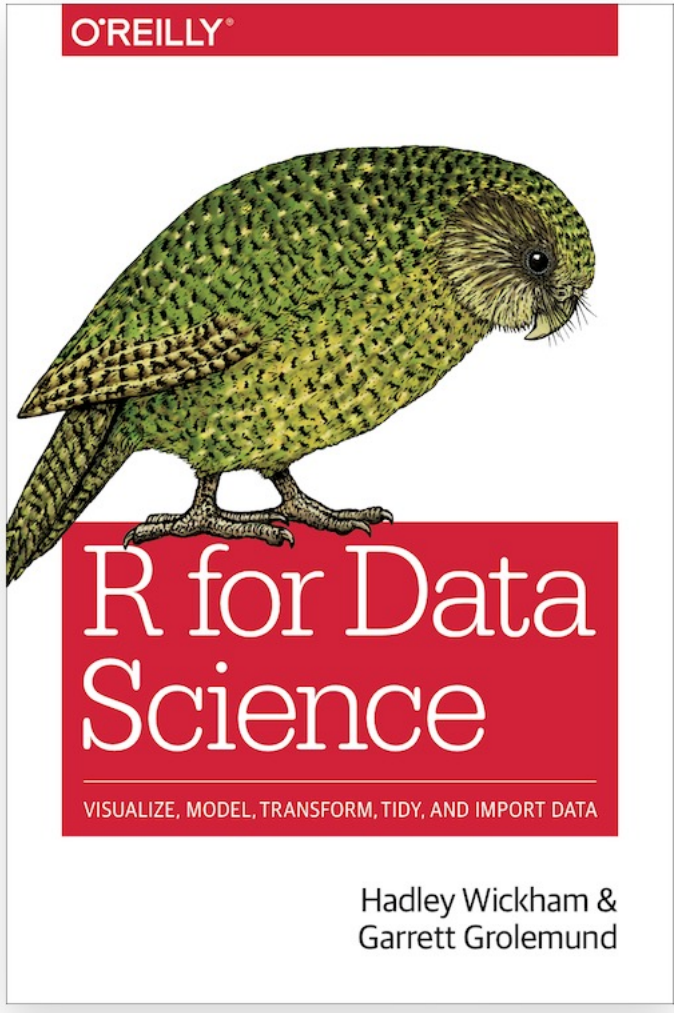
Course Page on Canvas

https://floridapolytechnic.instructure.com/courses/6098



Textbooks


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer


O'REILLY
R for Data Science
VISUALIZE, MODEL, TRANSFORM, TIDY, AND IMPORT DATA
Hadley Wickham &
Garrett Grolemund

Lecture Slides

- Will be posted on CANVAS
  - Try to post before class so you can take notes
- In-Class Notes
  - Will be posted after class
- R-Studio Notebooks
  - In class and lab assignments will be posted on CANVAS

Attendance

- We will use A+ attendance
  - Sharing the code with others is not allowed
  - Accounts for 10% of your final grade
  - Email instructor if you must miss a class

Support

- Office – IST 2040
- Office Hours
  - T/Th 12:30 – 1:30 PM
  - W 1:30 – 2:30 PM
  - In-Person
  - Virtual through Microsoft Teams (Links on Canvas)
  - By appointment
- Email: svadlamani@floridapoly.edu

Syllabus

- On Canvas

Action Items

- Orientation Quiz (graded)
- Primer Quiz (on canvas)
  - An indicator of your standing from prerequisite classes
  - Goal is to assess your understanding of the concepts and provide review material if needed