STA 35C: Statistical Data Science III

Lecture 1: Introduction and Overview

Dogyoon Song

Spring 2025, UC Davis

Agenda

- Course overview
- Brief intro to statistical learning
- Course logistics

What is statistical data science?

- **Statistics**: The study of collecting, analyzing, and drawing conclusions from data.
- **Data science**:
  - Interdisciplinary: statistical thinking + programming + databases
  - Emphasize real-world data wrangling, practical computing, and applications (science, business, sports, government, etc.)
- **STA 35 series**: Introductory data science courses from a statistical perspective, with an emphasis on computing.
  - **STA 35A**: Intro to statistics (probability, distributions, confidence intervals, hypothesis testing ,etc.)
  - **STA 35B**: Advanced R functionalities + additional statistical methods (linear regression, ANOVA, permutation tests, etc.)
  - **STA 35C**: Fundamentals of **statistical learning** methods
    - what the key ideas are, how and when to apply them, plus understanding their limitations

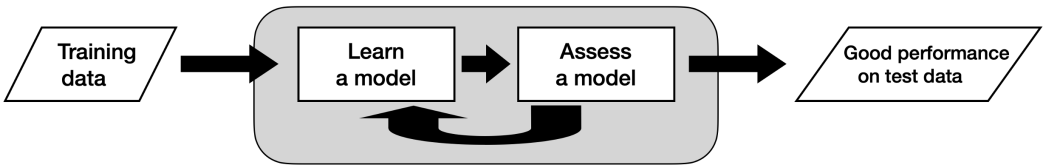
What is statistical learning?

- **Definition**: A set of tools for understanding data and making informed predictions
- **Examples**:
  - Identifying critical disease risk factors from large patient records
  - Predicting whether an event will occur (e.g., credit default, septic shock)
  - Classifying medical images or tissue cells (e.g., benign vs. malignant tumors)
  - Recognizing and localizing objects in images (e.g., for autonomous vehicles)
  - Evaluating impacts of new legislation or predicting unemployment rates
  - Modeling relationships among many variables to gain practical insights (e.g., which marketing strategies drive sales)
  - ...

Supervised vs. unsupervised learning

- **Supervised learning**
  - **Setup**: We have measurements of an outcome Y and predictors (features) X
  - **Goals**: Make accurate predictions of Y, or understand which X affects Y and how
  - **Examples**:
    - *Regression*: Forecast a product's sales next month
    - *Classification*: Predict if a customer will default on a loan
- **Unsupervised learning**
  - **Setup**: We only have features X, without any outcome variable
  - **Goals**: Discover hidden patterns or groupings in the data
  - **Examples**:
    - *Dimension reduction*: Extract a small subset or combine features for compression
    - *Clustering*: Cluster customers by purchasing behavior

Statistical learning and STA 35C



- **Core idea**: Learn a model from training data, evaluate its performance, and refine it
  - Aim for good predictions or insights on ***new, unseen* data**
  - Rely on probability and statistical principles to measure uncertainty and avoid overfitting
- In **STA 35C**, you'll learn the fundamentals of these methods
  - When and how to use different supervised or unsupervised learning methods
  - How to assess and interpret models (cross-validation, bootstrap, model selection)
  - Our focus is on **first principles**; we **do not** cover advanced machine learning techniques (e.g., deep neural networks, large language models)

Course content & prerequisites

Course content:

1. More on probability
2. Intro to supervised learning
   - Basic concepts
   - Regression
   - Classification
3. Model assessment, selection, and inference
   - Cross-validation and the bootstrap
   - Model selection and regularization
   - Simultaneous inference
4. Intro to unsupervised learning
   - Dimension reduction and PCA
   - Clustering

Prerequisite(s):

- STA 035B (C- or better)
- MAT 016B or 017B or 021B (C- or better)

These requirements are strict.

- If you don't meet prerequisites, please submit a petition ASAP

Course logistics

- **Instructor:** Dogyoon Song
  - E-mail: dgsong@ucdavis.edu
  - Office hours: Wed, 4-5pm (or by appointment) at MSB 4220
- **TA:** Soobin Kim
  - E-mail: sbbkim@ucdavis.edu
  - Office hours: Mon/Thu, 1-2pm (location: TBA)
- **Lectures:** Monday, Wednesday and Friday, 12:10-1:00 PM
- **Lab/Discussions:** Tuesday (run by Soobin Kim)
- **Online platforms**
  - Course webpage: Lecture notes, homework, supplementary materials, etc.
  - Canvas: Lab materials, homework submission (via Gradescope), solutions and grades
  - Piazza: Announcements and discussion
  - E-mail: Questions related to private matters only (**do not** send me messages on Canvas)

Grading

- **Homework**: 30%
  - Six homework assignments, excluding "Homework 0"
  - Assigned on Wednesday morning, due next Tuesday 11:59 pm PT
  - One homework with the lowest score can be dropped
  - **No late homework accepted for any reason**
- **Midterm exams**: 30%
  - Two in-class midterms (Fri, April 25 & Fri, May 16)
  - The lower can be dropped
  - **No make-up exams offered**
- **Final exam**: 40%
  - Friday, June 6, 1:00-3:00 PM
- **Participation**: up to 3% extra

See syllabus for full details and additional information (textbook, course policies, etc.)

"Homework 0" for self-assessment

- Complete the "Homework 0" for your self-assessment ASAP if you haven't yet
- It reviews key topics from STA 35A/B, and briefly check on your familiarity with R
- This will not be collected or graded, and no solutions will be provided.
- If you find any part challenging or need help with R or RStudio (e.g., installation), please review your STA 35A/35B notes, textbooks, or online resources, and **attend discussion sections tomorrow** (Tue, April 1, 2025).
- If you need additional help, please feel free to attend office hours and consult with the instructor or TA during the first week

Software: R and RStudio

- Data science teams often use a **mix of languages**, such as R, Python, or Julia.
- **R** is a free, open-source **statistical programming language** for data analysis:
  - Interactive environment for data wrangling, modeling, and visualization
  - Highly extensible via packages
- Basic interaction with R happens in the **R console** (terminal/command line).
- **RStudio** is a popular Integrated Development Environment (IDE) that:
  - Builds on top of the R Console
  - Provides menus, a file explorer, an editor, and other graphical tools
  - Streamlines the data science workflow

Computing setup

- **R programming**:
  - In this course, you will use R for homework and labs.
  - Lab/discussion sections and TA office hours are the best places to get help.
- **Where to run R**:
  - Use UC Davis JupyterHub, which has RStudio set up
  - Alternatively, you can install R and RStudio on your own computer
- **Computer access**:
  - You will need regluar, reliable access to a computer with a working browser or an up-to-date R/RStudio installation
  - If this is a problem, please let us know immediately—resources are available to help
- **Lab/discussion section**:
  - Held in TLC 2212, where computers are available
  - You may also bring your own laptop (please charge it beforehand)

Where we are headed

- **Throughout the course**:
  - Learn key ideas of regression, classification, clustering, and more
  - Practice implementing methods and interpreting results
  - Assess when each method is or isn't appropriate
- **Immediate next steps**:
  - Refresh core probability concepts (from STA 35A/B)
  - Deepen understanding of conditional probability and briefly explore the Bayesian ideas
- **Before next lecture**:
  - Complete "Homework 0" and seek help if needed
  - Ensure you have access to R and RStudio