STA 35C: Statistical Data Science III

Lecture 1: Introduction and Overview

Dogyoon Song

Spring 2026, UC Davis

Agenda

- Course overview
- Course logistics
- Intro to probability

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
    - understanding key ideas, how and when to apply them, and their limitations

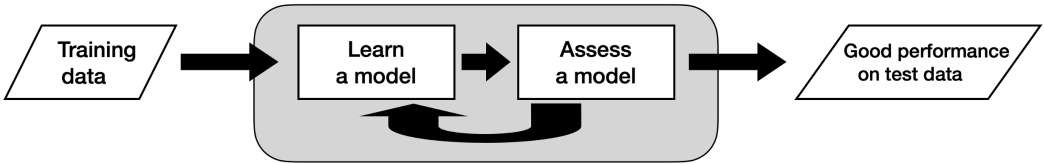
What is statistical learning?

- **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**: Predict Y (given X), 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 or combine a few 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 handle uncertainty
- 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, model selection, ...)
- Our focus is on learning **first principles**; we **do not** cover advanced machine learning techniques (e.g., deep neural networks, large language models)

Course logistics

- **Instructor:** Dogyoon Song
  - E-mail: dgsong@ucdavis.edu
  - Office hours: Wed, 3:30–4:30pm (or by appointment) at MSB 4220
- **TA:** Yanhao Jin
  - E-mail: yahjin@ucdavis.edu
  - Office hours: TBA
- **Lectures:** Monday, Wednesday and Friday, 1:10–2:00 PM at Wellman Hall 6
- **Lab/Discussions:** Tuesday 8:00–8:50 AM / 9:00–9:50 AM at TLC 2212 (run by TA)
- **Online platforms**
  - Course webpage: Lecture notes, homework, supplementary materials, etc.
  - Canvas: Homework submission, solutions and grades, lab materials
  - Piazza: Announcements and discussion
  - E-mail: Private matters only (**do not** send messages on Canvas)

Course content & prerequisites

Course content:

- Probability basics
- Intro to supervised learning
  - Fundamental concepts
  - Regression
  - Classification
- Model assessment, selection, and inference
  - Cross-validation and the bootstrap
  - Model selection and regularization
  - Simultaneous inference
- 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
- Try "Homework 0" for self-assessment

"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, March 31, 2026).
- If you need additional help, please feel free to attend office hours and consult with the instructor or TA during the first week

Grading

- **Homework**: 35%
  - 7 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**: 25%
  - Two in-class midterms (Fri, April 24 & Fri, May 15)
  - The lower can be dropped
  - **No make-up exams offered**
- **Final exam**: 40%
  - Friday, June 5, 1:00-3:00 PM
- **Participation**: up to 3% extra
  - Contribute in person to a class discussion during a lecture

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

Software: R and RStudio

- **R** is a free, open-source **statistical programming language** for data analysis:
  - Interactive environment for data wrangling, modeling, and visualization
  - Highly extensible via packages
  - In this course, you will use R for homework and labs

- **Where to run R / Computer access**:
  - You can install R and RStudio on your own computer
  - Alternatively, refer to UC Davis Cloud Services
  - If there is a problem, let us know immediately—resources are available to help

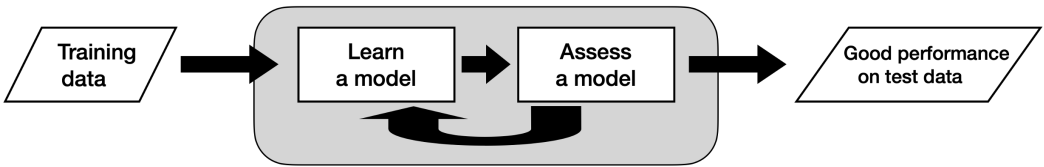
- **Lab/discussion section**:
  - Held in TLC 2212, where computers are available
  - You may also bring your own laptop (please charge it beforehand)
  - Lab/discussion and TA office hours are the best places to get help with R

Motivation: Why probability?

Recall the goals of statistical learning:

- Predict Y given X (learn a function $f : X \to Y$)
- Identify patterns in data X

Standard workflow:



Key challenge:

- We aim for good predictions or insights on ***new, unseen* data**
- How should we assess a model, given that training data $\neq$ test data?

Challenges in statistical learning and probabilistic solutions

Probabilistic tools and viewpoints offer a formal way to manage and quantify uncertainty

In particular,

- **Issue/Need:** Training data $\neq$ test data
  - *Solution:* Assume training and test data are *randomly drawn* from same distribution
- **Issue/Need:** Uncertainty in prediction
  - *Solution:* Model Y as a *random variable*, and predict Y conditioned on X
- **Issue/Need:** Choosing among many models
  - *Solution:* Update our belief about the "best model" based on observed data

We will discuss these aspects in more detail throughout the course

Probability in everyday examples

- Coin toss
  - Possible outcomes are Head or Tail; each has probability 0.5
- Die roll
  - Possible outcomes $\{1, 2, \ldots, 6\}$; each has probability $1/6$
- Y chromosomes in the US childbirths[1]
  - About 51.2% of births are to babies with Y chromosomes, and 48.8% do not
  - The probability of having a baby with a Y chromosome is 0.512
- Commute time
  - Commute may take 30 minutes on average, but can vary with traffic, weather, etc.
- Subjective probabilistic beliefs
  - You may personally estimate the likelihood of a stock price rising or falling, based on your own analysis or expert opinions
  - This kind of probability reflects beliefs rather than strict long-run frequencies

[1] Source: CDC National Vital Statistics Reports, Births: Final Data for 2023

Formalizing probability: Sample space and events

- **Outcome:** A possible result of an experiment or trial
- **Sample space**: the set of all possible outcomes, often denoted by Ω
  - e.g., $\{H, T\}$, $\{1, 2, 3, 4, 5, 6\}$
- **Event:** a subset of Ω
  - e.g., $\emptyset$, $\{H\}$, $\{T\}$ $\{H, T\}$, $\{6\}$, $\{1, 2\}$, $\{2, 4, 6\}$
- **Probability**[2]**:** a map P that assigns a number in $[0, 1]$ to each event such that
  - $P(\Omega) = 1$;
  - For disjoint events $A_1, A_2, \ldots$, $P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$
    - Simply put, if $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$
    - Example: $A_1 = \{1, 2\}$, $A_2 = \{6\}$

[2] A formal, mathematically rigorous definition of probability measure is beyond the scope of STA 35C

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 is not appropriate
- **Immediate next steps (Week 1)**:
  - Refresh core probability concepts (from STA 35A/B)
  - Deepen understanding of conditional probability and explore Bayesian ideas
- **Before next lecture**:
  - Complete "Homework 0" and seek help if needed
  - Ensure you have access to R and RStudio