STA 35C: Statistical Data Science III

Lecture 4: Statistical Learning

Dogyoon Song

Spring 2026, UC Davis

Announcements

Homework 1 is due Thu, Oct 1, 11:59 PM PT

- Please submit on time and follow the submission instructions
- Discussion section tomorrow: Thu, 11:00–11:50 AM (TLC 2212)
- Instructor office hours: Wed, 1–2 PM (MSB 1143)
- TA office hours: Thu, 12–1 PM (MSB 1117)
- Feel free to post questions on Piazza (the earlier, the better)

Agenda

Last time:

- Bayes' rule
- Random variables
  - Distribution functions
  - Expectation & variance

Today:

- Complete probability review: joint distribution, covariance, sums of random variables
- Statistical learning
  - Motivating examples
  - Supervised vs. unsupervised learning
  - Prediction vs. inference
  - Parametric vs. nonparametric methods

Recap: Bayes' theorem and random variables

Bayes' theorem states that *Posterior* = *Likelihood* × *Prior* / *Evidence*, i.e.,

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)} \qquad \text{where} \qquad P(B) = P(B \mid A)P(A) + P(B \mid A^c)P(A^c)$$

- For example, A: nature/model & B: data

Random variable: a map $X : \Omega \to \mathbb{R}$ assigns a real number to each outcome $\omega \in \Omega$

- The **PMF** of a discrete random variable p_X: $p_X(x) = P(X = x)$
- The **PDF** of a continuous random variable f_X: $P(a \leq X \leq b) = \int_a^b f_X(x)\, dx$
- The **CDF** of either variable F_X: $F_X(x) = P(X \leq x)$
- **Expectation** and **variance** summarize center and spread
  - Useful identities follow from the linearity of expectation

Distribution of multiple random variables

Let X, Y be discrete random variables

Joint distribution:

- The joint PMF of X and Y satisfies $p_{X,Y}(x, y) = P(X = x,\ Y = y)$
- The joint CDF of X and Y is defined by $F_{X,Y}(x, y) = P(X \leq x,\ Y \leq y)$

Marginal distributions: The marginal PMFs of X and Y are:

$$p_X(x) = \sum_y p_{X,Y}(x, y), \qquad p_Y(y) = \sum_x p_{X,Y}(x, y)$$

Conditional distribution: For $p_Y(y) > 0$, the conditional PMF of X given $Y = y$ is

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} = \frac{p_{X,Y}(x, y)}{\sum_{x'} p_{X,Y}(x', y)}$$

- **Independence:** X and Y are independent if $p_{X,Y}(x, y) = p_X(x)\, p_Y(y)$ for all x, y

For continuous RVs, PMFs are replaced by PDFs and sums are replaced by integrals

Joint, marginal, conditional distributions: Example

Example (One fair die)

Suppose rolling a fair die and let

$$X = \mathbf{1}\{\text{roll is even}\}, \qquad Y = \mathbf{1}\{\text{roll} > 3\}.$$

$$(X, Y) = \begin{cases} (0,0) & \text{if } \omega \in \{1,3\} \\ (0,1) & \text{if } \omega \in \{5\} \\ (1,0) & \text{if } \omega \in \{2\} \\ (1,1) & \text{if } \omega \in \{4,6\} \end{cases} \implies$$

	$Y=0$	$Y=1$	$p_X(x)$
$X=0$	$1/3$	$1/6$	$1/2$
$X=1$	$1/6$	$1/3$	$1/2$
$p_Y(y)$	$1/2$	$1/2$	

So the conditional probability

$$p_{X|Y}(1 \mid 1) = \frac{p_{X,Y}(1,1)}{p_Y(1)} = \frac{1/3}{1/2} = \frac{2}{3}$$

Also, X and Y are not independent, since

$$p_{X,Y}(1,1) = \frac{1}{3} \neq p_X(1)p_Y(1) = \frac{1}{4}$$

Covariance and correlation

The **covariance** between X and Y measures how they vary together:

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

- $\mathrm{Cov}(X, Y) > 0$: large values of X tend to occur with large values of Y
- If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$; the converse is false in general

Correlation is the normalized version of covariance:

$$\rho(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \in [-1, 1] \qquad (\text{when } \mathrm{Var}(X), \mathrm{Var}(Y) > 0).$$

Example (One fair die)

Recall the example from last time: $X = \mathbf{1}\{\text{roll is even}\}$, and $Y = \mathbf{1}\{\text{roll} > 3\}$

	$Y = 0$	$Y = 1$
$X = 0$	$1/3$	$1/6$
$X = 1$	$1/6$	$1/3$

- $\mathrm{Cov}(X, Y) = \frac{1}{12}$
- $\rho(X, Y) = \frac{1}{3}$

Sum of random variables: Expectation and variance are easy

Sums arise naturally when we combine several sources of randomness:

- Total number of heads in n tosses: $S_n = X_1 + \cdots + X_n$
- Total score, total revenue, total waiting time
- Total measurement error from several noisy components

The expectation and variance of $X + Y$ are often easy to compute:

- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$

$$\because \ \mathbb{E}[X+Y] = \sum_{x,y} (x+y)\, p_{X,Y}(x,y) = \sum_x x\, p_X(x) + \sum_y y\, p_Y(y)$$
$$= \mathbb{E}[X] + \mathbb{E}[Y]$$

- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$
  - *Exercise:* Verify this using properties of expectation
  - If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$, so $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Sum of random variables: Distributions can be harder

Even when $\mathbb{E}[X+Y]$ and $\mathrm{Var}(X+Y)$ are easy, the full distribution of $X+Y$ may not be

- We usually need the *joint distribution* of (X, Y), not just the separate marginals

Example (Same marginals, different joint)

Suppose X_1 and X_2 each take values 0 or 1 with probability $1/2$

- If X_1 and X_2 are independent, then

$$P(X_1 + X_2 = k) = \begin{cases} 1/4 & k = 0, \\ 1/2 & k = 1, \\ 1/4 & k = 2 \end{cases}$$

- If $X_1 = X_2$ always, then

$$P(X_1 + X_2 = k) = \begin{cases} 1/2 & k = 0, \\ 1/2 & k = 2 \end{cases}$$

$\Longrightarrow$ Marginals do *not* determine the distribution of the sum; the joint structure matters

Pop-up quiz

Consider two independent tosses of a fair coin with sample space

$$\Omega = \{HH, HT, TH, TT\}.$$

Define

$$X(\omega) = \text{number of Heads in } \omega.$$

Which statement is correct?

a). X is not a random variable because both HT and TH map to 1.

b). X is an event in Ω.

c). The sample space is $\{0, 1, 2\}$ because X takes values in $\{0, 1, 2\}$.

d). $\{X = 1\} = \{HT, TH\}$, so $P(X = 1) = 1/2$.

Answer: D.
A random variable is a function from outcomes to real numbers. Different outcomes are allowed to map to the same value, and $\{X = 1\}$ is an event (a subset of Ω); thus, the original sample space remains $\Omega = \{HH, HT, TH, TT\}$.

Motivating example: Predicting sales

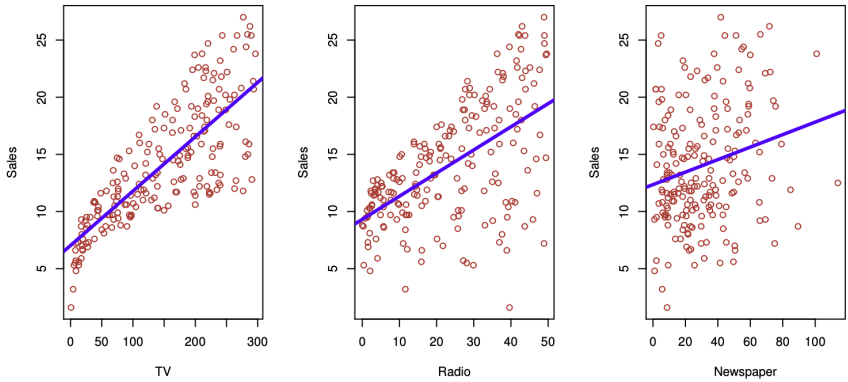


Figure: The `Advertising` data set shows `Sales` of a product in 200 different markets against advertising budgets for three media: `TV`, `Radio`, and `Newspaper` [JWHT21, Figure 2.1].

Question: Can we predict `Sales` from the budgets `TV`, `Radio`, and `Newspaper`?

Motivating example: Income and predictors

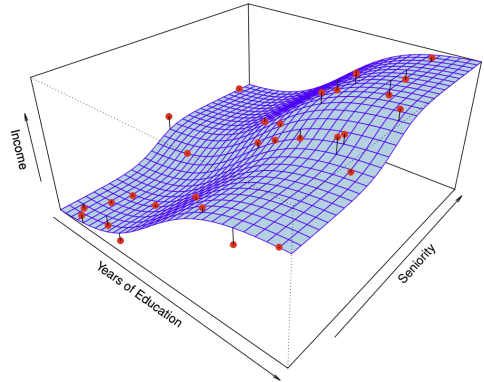


Figure: The simulated `Income` data set displays `Income` of 30 individuals as a function of `Years of education` and `Seniority` [JWHT21, Figure 2.3].

Question: Which of the predictors, e.g., `Years of education` and `Seniority`, are most strongly associated with `Income` and how?

Motivating example: Clustering unlabeled data

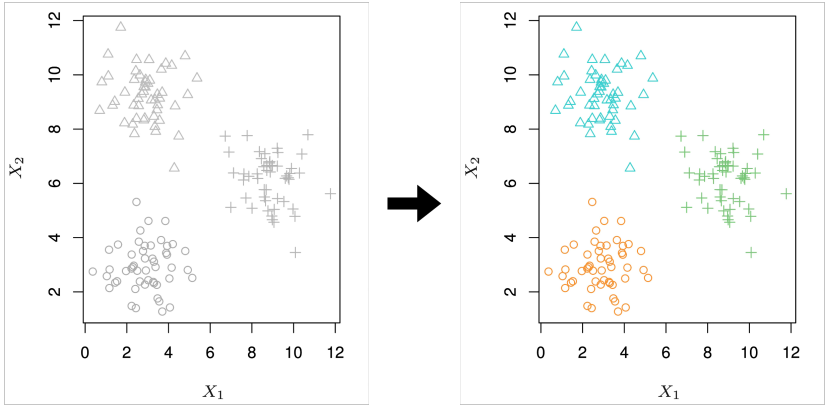


Figure: A synthetic data set with three underlying groups; colors are shown only for illustration. [JWHT21, Figure 2.8, modified].

Question: If there are no labels given, can we still discover meaningful groups?

Statistical learning: Supervised vs. unsupervised learning

Most statistical learning problems fall into two categories: supervised or unsupervised

In **supervised learning**:

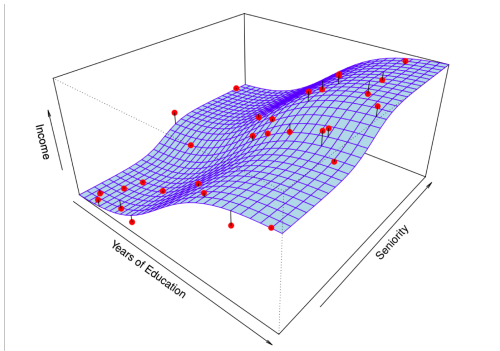
- Each predictor observation x_i is accompanied by a response y_i
- "Supervised" because the responses guide (supervise) the analysis
- Many classical statistical learning methods operate in the supervised learning domain
  - *Example:* linear regression, logistic regression, support vector machine, ...

In **unsupervised learning**:

- We have observations x_i but no response y_i
- "Unsupervised" because there is no response to guide the analysis
- Often used to explore relationships among observations or variables
  - *Example:* Cluster analysis, dimension reduction, ...

Some problems lie between these two categories and the distinction can be less clear-cut

Illustration: Supervised vs. unsupervised learning



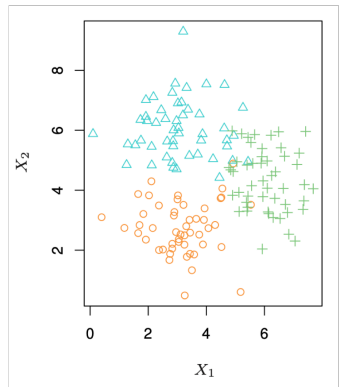


Figure: Supervised vs. unsupervised learning

For now, we focus on supervised learning; unsupervised learning comes later in the quarter

Statistical learning: Terminology and notation

Response (dependent variable) Y:

- The output variable we want to predict (e.g., `Sales`)

Predictors (independent variables, features) X:

- Input variables used to predict Y (e.g., `TV`, `Radio`, `Newspaper`)
- Often multiple predictors are collectively denoted by $X = (X_1, X_2, \ldots, X_p)$

A common working assumption is that there is some relationship between Y and X:

$$Y = f(X) + \epsilon,$$

where

- f is some fixed but ***unknown*** function.
- ϵ is a ***random*** noise, which has *mean zero*, and is *independent of* X.

Goal: Learn an estimator $\hat{f}$ of the unknown function f

Why learn f? Three concrete goals

Predicting Y:

- We often have input variables X but not the corresponding output Y
- With an estimate $\hat{f}$, we can *predict* Y at new points $X = x$ via $\hat{Y} = \hat{f}(x)$
- *Example:* $X =$ patient's blood sample, $Y =$ risk of a disease or adverse reactions

Identifying relevant predictors:

- We can determine *which* predictors among $X_1, \ldots, X_p$ are important in explaining Y
- *Example:* Some predictors may be strongly associated with income, while others add little predictive value

Understanding how Y changes with X:

- If f is not too complex, we can interpret *how* each predictor X is associated with Y
- *Example:* Understanding how predicted sales change as TV advertising increases, holding the other predictors fixed

These goals mostly fall under two broader objectives: *prediction* and *inference*

Two broad goals: Prediction vs. inference

Prediction

- *Objective:* Make accurate prediction of Y given X
- $\hat{f}$ can be treated as a "black box," prioritizing predictive accuracy over exact form
- *Examples:*
  - Which individuals, based on demographics, are likely to respond positively to a mailer?
  - Based on blood sample, is a patient at high risk of a severe adverse drug reaction?

Inference

- *Objective:* Understand the association between Y and X
- We need to interpret the estimated relationships, not just produce accurate predictions
- *Examples:*
  - Which media are linked to higher sales?
  - Which medium has the strongest association with sales?
  - How do predicted sales change with TV advertising, holding other predictors fixed?

Prediction error: Reducible vs. irreducible

Under the working model $Y = f(X) + \epsilon$, prediction error of $\hat{Y} = \hat{f}(X)$ has two parts:

- **Reducible error:** error from approximating f by $\hat{f}$
- **Irreducible error:** randomness in ϵ that cannot be predicted using X alone, hence no method can eliminate solely based on X
  - ϵ may include *unmeasured* variables important for predicting Y
  - ϵ may also reflect *inherent* fluctuations (e.g., day-to-day or manufacturing variation)

For a fixed fitted function $\hat{f}$ and an independent test observation at $X = x$,

$$\mathbb{E}\left[(Y - \hat{f}(x))^2 \mid X = x\right] = \underbrace{(f(x) - \hat{f}(x))^2}_{\text{reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{irreducible}}, \qquad (\because \mathbb{E}[\epsilon \mid X = x] = 0)$$

Goal: Learn $\hat{f}$ with low prediction error and interpretability suited to the task

How do we estimate f?

We will explore multiple approaches to estimate f from data

We use *training data* $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ to fit $\hat{f}$

Our goal: ensure $\hat{f}$ generalizes well to future data (X, Y)

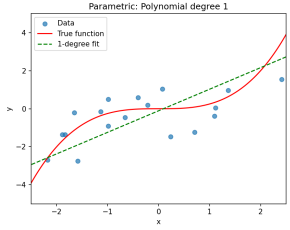
Most statistical learning methods are either **parametric** or **non-parametric**

- **Parametric (model-based) approach:**
  - Step 1: Assume a functional form (model) of f (e.g., linear $f(x) = \alpha + \beta x$)
  - Step 2: Use the training data to fit model parameters
- **Non-parametric approach:**
  - Do not fix a parametric form for f; assumptions such as smoothness still matter
  - Estimate f that fits data closely flexibly, while controlling complexity of $\hat{f}$ to avoid overfitting

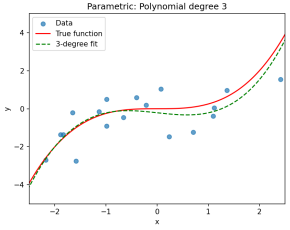
Illustration of parametric methods

Suppose that $Y = f(X) + \epsilon$ where $f(x) = \frac{1}{4}x^3$

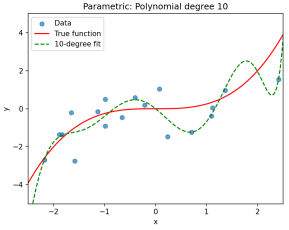
Parametric methods: e.g., polynomial regression $\hat{f}(x) = \sum_{j=0}^{d} \beta_j x^j$



(a) Linear regression (deg 1)



(b) Polynomial regression (deg 3)



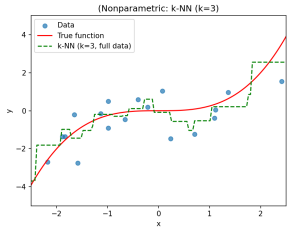
(c) Polynomial regression (deg 10)

- Strength: Estimating parameters $\beta_0, \ldots, \beta_d$ is easier than estimating an arbitrary function f
- Limitation: The assumed model may not match the true functional form of f
- Tradeoff: Choosing a more flexible model can reduce bias but risks *overfitting*

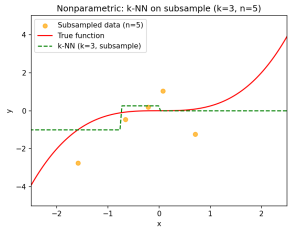
Illustration of non-parametric methods

Suppose that $Y = f(X) + \epsilon$ where $f(x) = \frac{1}{4}x^3$

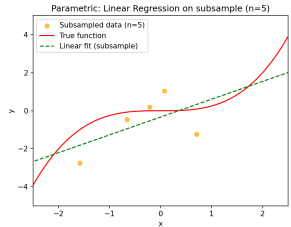
Non-parametric methods: e.g., k-nearest neighbors (k-NN)



(a) k-nearest neighbor ($k = 3$)



(b) k-NN ($k = 3$, subsampled)



(c) Linear regression (subsampled)

- Strength: Flexible; less tied to a specific functional form
- Limitation: Often needs more data to accurately estimate f and can be harder to interpret
- Tradeoff: Flexibility can increase overfitting risk; prediction can be computationally costly

Tradeoff: Flexibility vs. model interpretability

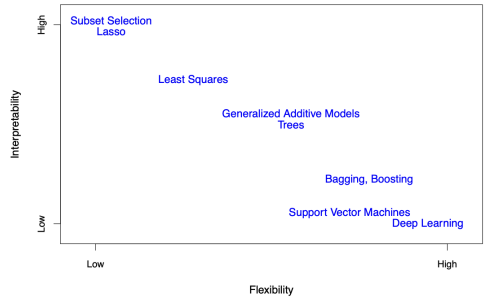


Figure: A representation of the tradeoff between flexibility and interpretability [JWHT21, Figure 2.7].

More flexible methods can capture a much wider range of relationships between X and Y, but we may still prefer more restrictive approaches because of:

- **Interpretability:** Restrictive (parametric) models are typically easier to interpret
- **Sample complexity:** Flexible models often require more observations
- **Risk of overfitting:** Very flexible methods can fit noise ϵ rather than true f

Wrap-up

- Supervised learning studies how predictors X can be used to predict or understand a response Y
- A common working model is $Y = f(X) + \epsilon$
  - f captures systematic signal
  - ϵ captures noise
- Supervised learning uses labeled responses Y; unsupervised learning looks for structure without them
- Two main goals are prediction and inference; the favored models depend on goals
- Parametric methods are usually simpler and more interpretable; non-parametric methods are more flexible but often need more data and can overfit

Suggested reading: [JWHT21, Ch. 2.1]

Next lecture: linear regression

References


Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
An Introduction to Statistical Learning: with Applications in R, volume 112 of *Springer Texts in Statistics*.
Springer, New York, NY, 2nd edition, 2021.