STA 35C: Statistical Data Science III

Lecture 3: Statistical Learning

Dogyoon Song

Spring 2025, UC Davis

Agenda

In the last lecture, we reviewed:

- Probability basics
- Conditional probability & Bayes' theorem
- Random variables
- Joint, marginal, and conditional distributions

Today, we will cover:

- More on probability with examples
- Statistical learning

Example: Expectation and variance

Example 1: Coin toss

$$p_X(x) = p(X = x) = \begin{cases} p & \text{if head } (x = 1), \\ 1 - p & \text{if tail } (x = 0). \end{cases}$$

- *Expectation:*

$$\begin{aligned} \mathbb{E}[X] &= \sum_x x \cdot p(x) \\ &= 0 \cdot (1 - p) + 1 \cdot p \\ &= p \end{aligned}$$

- *Variance:*

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \sum_x \left(x - p\right)^2 \cdot p(x) \\ &= (-p)^2 \cdot (1 - p) + (1 - p)^2 \cdot p \\ &= p(1 - p) \end{aligned}$$

- Alternatively:

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1 - p)$$

Example: Expectation and variance

Example 2: Gaussian random variable $X \sim N(0, \sigma^2)$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

- *Expectation:*

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x \cdot f_X(x)\ dx \\
&= \int_{-\infty}^{0} x \cdot f_X(x)\ dx + \int_{0}^{\infty} x \cdot f_X(x)\ dx \\
&= \int_{0}^{\infty} (-x + x) \cdot f_X(x)\ dx \\
&= 0
\end{aligned}$$

- *Variance:*

$$\begin{aligned}
\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 &= \int_{-\infty}^{\infty} x^2 \cdot f_X(x)\ dx \\
&= \sigma^2 \int_{-\infty}^{\infty} f_X(x)\ dx = \sigma^2 P(X \in \mathbb{R}) \\
&= \sigma^2
\end{aligned}$$

- Integration by parts: for $a \neq 0$ ($a = \frac{1}{2\sigma^2}$),

$$\int_{-\infty}^{\infty} x^2 e^{-ax^2} dx = \frac{x}{-2a}\ e^{-ax^2} \Big|_{-\infty}^{\infty} + \frac{1}{2a} \int_{-\infty}^{\infty} e^{-ax^2} dx$$

Example: Sum of random variables

Example 3: A mixture of two Gaussians

- Let $X \sim \mathit{Bern}(p)$, i.e., a Bernoulli random variable such that

$$p_X(x) = \begin{cases} p & \text{if head } (x = 1), \\ 1 - p & \text{if tail } (x = 0). \end{cases}$$

- Let $Y \sim N(0, \sigma^2)$, i.e., a Gaussian random variable with

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}}$$

We have seen that

- $\mathbb{E}[X] = p$ and $\mathrm{Var}(X) = p(1-p)$
- $\mathbb{E}[Y] = 0$ and $\mathrm{Var}(Y) = \sigma^2$

Suppose that $\mathrm{Cov}(X, Y) = 0$

Question: Compute

- $\mathbb{E}[cX + Y]$
- $\mathrm{Var}(cX + Y)$

Question: Draw the distribution of $X + Y$?

Example: Variance and covariance

Example 4: Consider a 2x2 contingency table as follows ($a \in [-1, 1]$)

X \ Y	-1	1	Marginal prob of X
1	(1 - a) / 4	(1 + a) / 4	1/2
-1	(1 + a) / 4	(1 - a) / 4	1/2
Marginal prob of Y	1/2	1/2	

- *Expectation of X:*

$$\mathbb{E}[X] = \sum_x x \cdot p_X(x) = (-1) \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 0$$

- *Variance of X:*

$$\mathrm{Var}(X) = \mathbb{E}[X^2] = (-1)^2 \cdot \frac{1}{2} + 1^2 \cdot \frac{1}{2} = 1$$

- *Covariance between X and Y:*

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \sum_{x,y} xy \cdot P_{X,Y}(x, y) \\ &= \frac{1+a}{2} - \frac{1-a}{2} \\ &= a \end{aligned}$$

Example: Variance and covariance

Example 4: Consider a 2x2 contingency table as follows ($a \in [-1, 1]$)

X \ Y	-1	1	Marginal prob of X
1	(1 - a) / 4	(1 + a) / 4	1/2
-1	(1 + a) / 4	(1 - a) / 4	1/2
Marginal prob of Y	1/2	1/2	

- $\mathbb{E}[X] = \mathbb{E}[Y] = 0$
- $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$
- $\mathrm{Cov}(X, Y) = a$

Thus,

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = a$$

- Q: Are X and Y independent[a]?
  - Yes if and only if $a = 0$
  - If X and Y are independent, then $\rho_{X,Y} = 0$
  - However, $\rho_{X,Y} = 0$ does *not* imply X and Y are independent

[a] Random variables X, Y are independent if $P_{X,Y}(A, B) = P_X(A)P_Y(B)$ for all A, B

Example: Variance and covariance

Example 4: Consider a 2x2 contingency table as follows ($a \in [-1, 1]$)

X \ Y	-1	1	Marginal prob of X
1	(1 - a) / 4	(1 + a) / 4	1/2
-1	(1 + a) / 4	(1 - a) / 4	1/2
Marginal prob of Y	1/2	1/2	

- Q: Is observing X useful in predicting Y?
- Q: How would you estimate a, or $\mathrm{sign}(a)$, from data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$?

Statistical learning

Let's begin with some examples

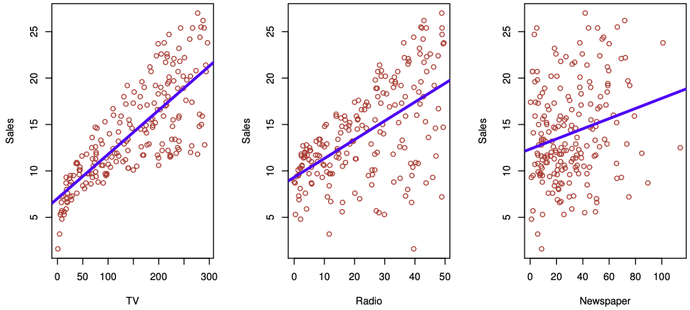


Figure: The `Advertising` data set shows `Sales` of a product in 200 different markets against advertising budgets for three media: `TV`, `Radio`, and `Newspaper` [JWHT21, Figure 2.1].

Want to know if there is an association between sales (Y) and advertising (X)
For example, can we predict `Sales` using `TV`, `Radio`, and `Newspaper`?

Statistical learning

Let's begin with some examples

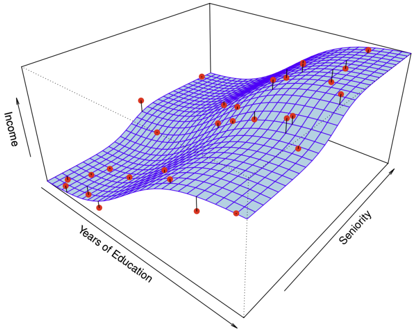


Figure: The simulated Income data set displays Income of 30 individuals as a function of Years of education and Seniority [JWHT21, Figure 2.3].

Want to know if there is an association between income (Y) and education/seniority (X)
For example, can we understand how Years of education affect Income?

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 $X = (X_1, X_2, \ldots, X_p)$

Assumption: There is some relationship between Y and X

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

where

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

Goal: Estimate f

Why estimate f?

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, and which are irrelevant
- *Example:* Seniority and years of education heavily affect income, but marital status typically does not

Understanding how X affects Y:

- If f is not too complex, we can interpret *how* each predictor affects Y
- *Example:* Measuring how an increase in TV advertising changes sales

Two main reasons to estimate f

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 cannot treat $\hat{f}$ as a black box; we need to know its exact form
- *Examples:*
  - Which media are linked to higher sales?
  - Which medium generates the largest boost in sales?
  - How much of an increase in sales is attributable to a given increase in TV advertising?

What is the smallest prediction error we can hope for?

The predictive accuracy of $\hat{Y} = \hat{f}(X)$ depends on two sources of error

- **Reducible error:** If $\hat{f}$ is not a perfect estimate of f, any inaccuracy introduces error
- **Irreducible error:** Even if $\hat{f} = f$, there is variability from ϵ
  - ϵ may include *unmeasured* variables important for predicting Y.
  - ϵ may also reflect *inherent* fluctuations (e.g., day-to-day or manufacturing variation).

Mathematically,

$$\begin{aligned}
\mathbb{E}(\hat{Y} - Y)^2 &= \mathbb{E}\big[(\hat{f}(X) - f(X) - \epsilon)^2\big] \\
&= \mathbb{E}\big[(\hat{f}(X) - f(X))^2\big] - 2\mathbb{E}\big[(\hat{f}(X) - f(X)) \cdot \epsilon)^2\big] + \mathbb{E}[\epsilon^2] \\
&= \underbrace{\left(\hat{f}(X) - f(X)\right)^2}_{\text{reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{irreducible}}
\end{aligned}$$

Goal: Estimate f that (1) minimizes the reducible error and (2) is interpretable

How do we estimate f?

- We will explore multiple approaches to estimate f from data
- We use *training data* $\{(x_1, y_1), \dots, (x_n, y_n)\}$ 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 $Y = \alpha + \beta X$)
        - Step 2: Use the training data to fit model parameters
    - **Non-parametric approach:**
        - Make no explicit assumption about the functional form of f
        - Instead, seek an $\hat{f}$ that fits data closely while remaining sufficiently smooth

Parametric vs. non-parametric methods

Parametric approach:

- Estimating parameters is easier than estimating an arbitrary function
- The assumed model may not match the true functional form of f
- Choosing a more flexible model can reduce bias but risks *overfitting*

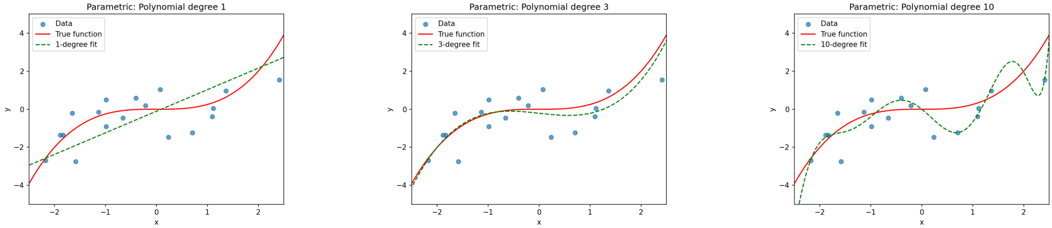
Non-parametric approach:

- Highly flexible; avoids the danger of using a wrong functional form
- Requires more data to accurately estimate a complex f
- Greater flexibility can also increase the risk of overfitting noise

Illustration of parametric methods

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

Parametric methods: polynomial regression $\hat{f}(x) = \sum_{j=0}^{d} \beta_j x^j$ (e.g., $\hat{f}(x) = \beta_0 + \beta_1 x$)



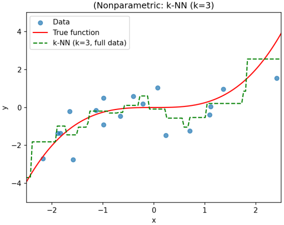
(a) Linear regression (deg 1) (b) Polynomial regression (deg 3) (c) Polynomial regression (deg 10)

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

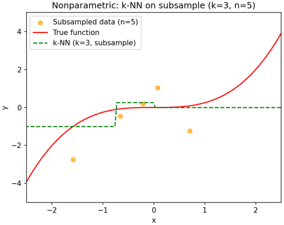
Illustration of non-parametric methods

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

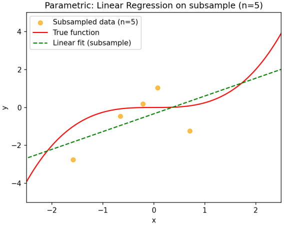
Non-parametric methods: k-nearest neighbors (k-NN)



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



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



(c) Linear regression (subsampled)

- Highly flexible; avoids the danger of using a wrong functional form
- Requires more data to accurately estimate a complex f
- Greater flexibility can also increase the risk of overfitting noise

Tradeoff: Prediction accuracy vs. model interpretability

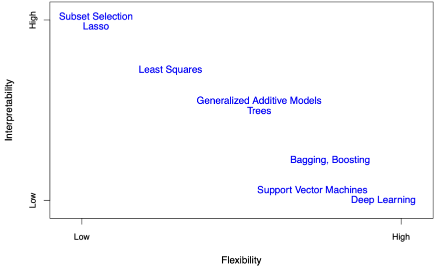


Figure: A representation of the tradeoff between flexibility and interpretability [JWHT21, Figure 2.7].

While more flexible methods can capture a much wider range of shapes to estimate f, we may still prefer more restrictive approaches because of:

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

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.