STA 141A – Fundamentals of Statistical Data Science

Department of Statistics; University of California, Davis

Instructor: Dr. Akira Horiguchi (ahoriguchi@ucdavis.edu)
A01 TA: Zhentao Li (ztlli@ucdavis.edu)
A02 TA: Zijie Tian (zijtian@ucdavis.edu)
A03 TA: Lingyou Pang (lyopang@ucdavis.edu)

Section 5: Overview of Statistical Learning

Spring 2025 (Mar 31 – Jun 05), MWF, 01:10 PM – 02:00 PM, Young 198

Overview

Based on Chapters 1 and 2 of ISL book James et al. (2021).

- For more R code examples, see R Markdown files in `https://www.statlearning.com/resources-second-edition`

Section 5: Overview of Statistical Learning

- Supervised learning
- Unsupervised learning

A note on generating "random" numbers

How do `sample()` and `rnorm()`, for instance, really work?

- Are the generated values really random? Technically, NO.
- The generated values are based on a certain *seed* (a positive integer). The seed is the 'starting point' or an initial value. It is plugged into a certain function/generator leading to our generated value.
- **Setting a seed (e.g. `set.seed(23)` in R) at the beginning of your code will lead to the same sequence of "random" numbers. This helps to make your "random" results more reproducible, which aids with debugging, science replicability, etc. It is recommended that you do this.**
  - If you want a different sequence of "random" numbers, replace 23 with your favorite positive integer.
- The generated values are called *pseudo-random*, so "almost random", as they appear to be random.
- If the seed is not specified, it is usually based on milliseconds of the computer's current time. This helps make the numbers "more random."

Further detail is outside the scope of this class.

Statistical learning: supervised vs unsupervised

Statistical learning refers to a vast set of tools for understanding data.

- These tools can be classified as *supervised* or *unsupervised*.
- *Supervised* statistical learning: predict or estimate an output based on one or more inputs.
- *Unsupervised* statistical learning: learn relationship or structure among observations.
- (Are there outputs to "supervise" the learning task?)

Section 5: Overview of Statistical Learning

Supervised learning

Example – linear regression

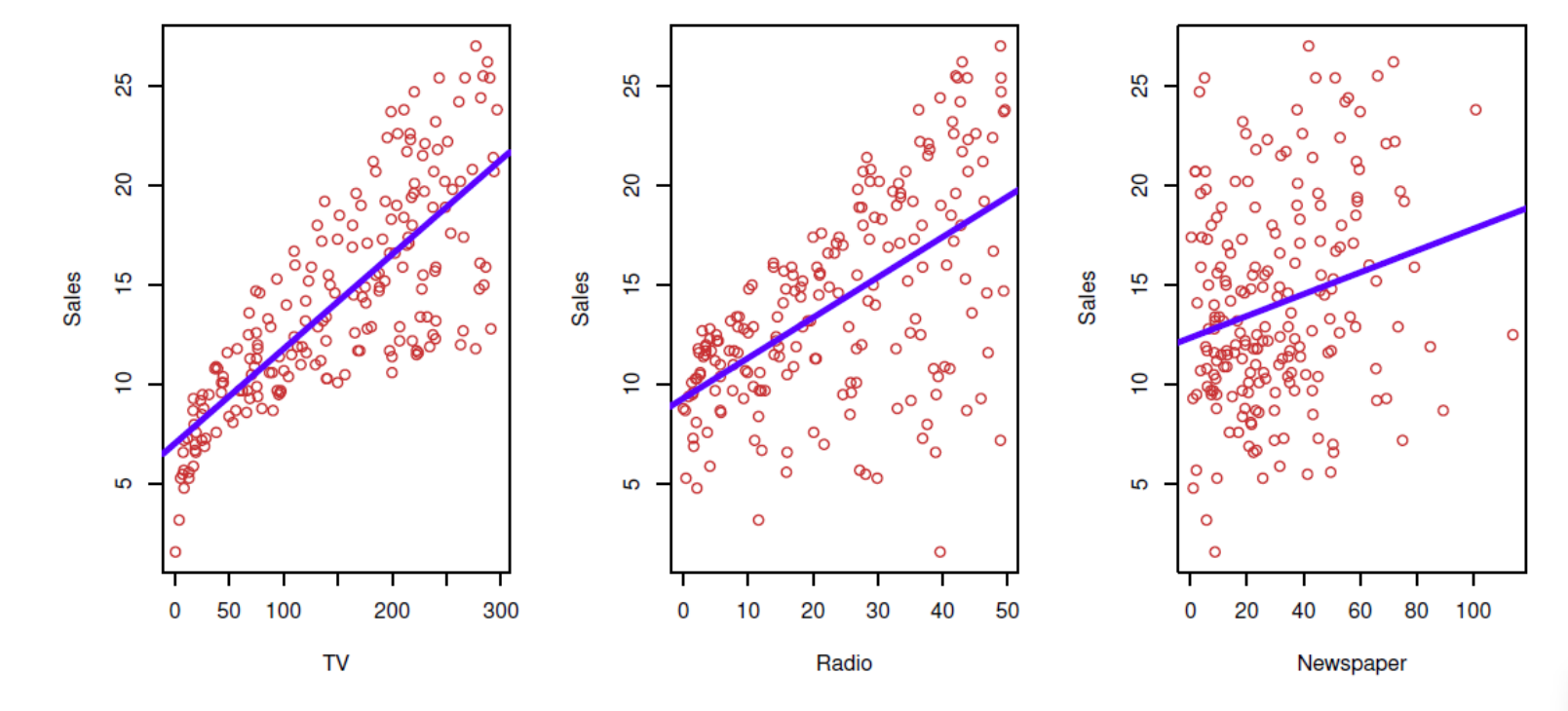

Figure 1: Image by James et al. (2021), based on the `Advertising` data set in R. The plot displays `sales` in thousands of units depending on the input TV, `radio`, and `newspaper` (advertising) budgets, in thousand dollars, for 200 different markets.

Example – Non-linear regression with two input variables

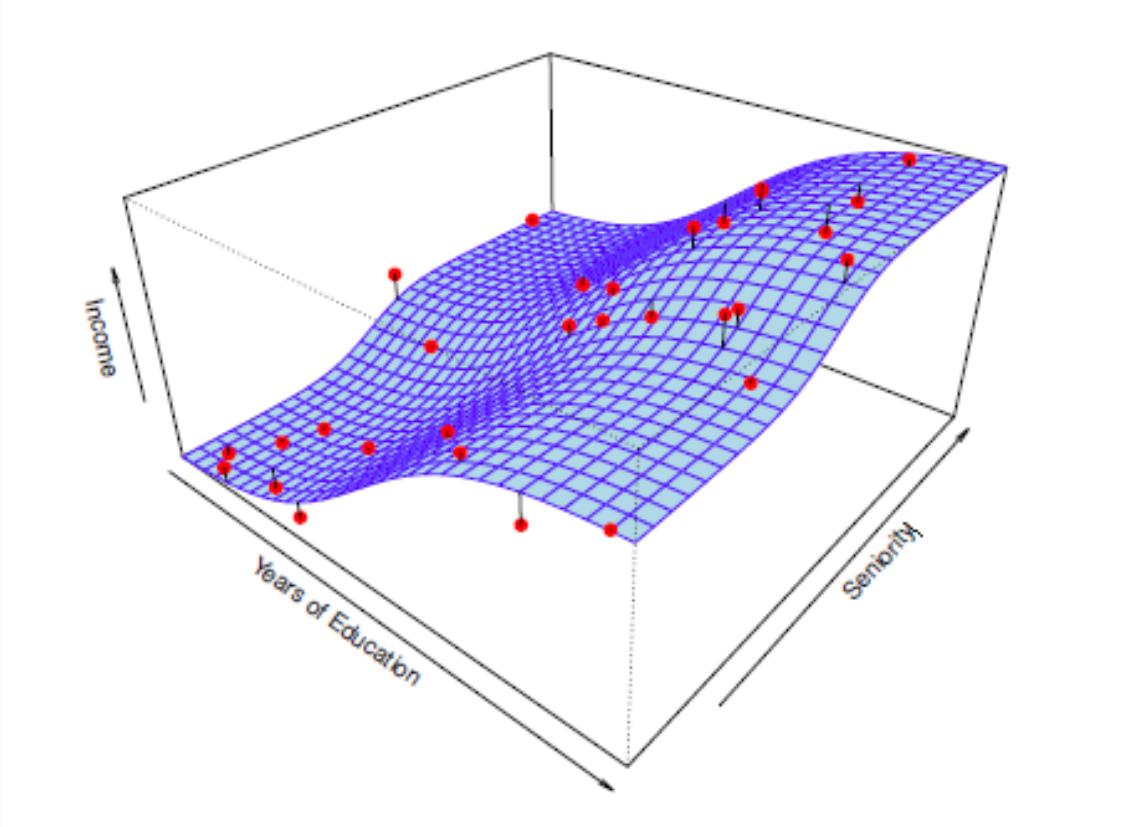

Figure 2: Image by James et al. (2021), based on the `Income` data set in R. The `income` is displayed as a function of `years of education` and `seniority`.

Some notation

Recall: predict or estimate an output based on one or more inputs.

- *Input variables* are called *predictors*, *independent variables*, or *features*; denoted by X, and X_1, X_2, X_3 etc. if there is more than one.
- *Output variable* is called *response* or *dependent variable*; denoted by Y.

Example: In Figure 1, the predictors are TV, `radio`, `newspaper`, denoted by X_1, X_2, X_3, respectively, and the response is `sales`, denoted by Y.

Model

Suppose we observe a quantitative (i.e., numeric) response Y, and predictors $X_1, \ldots, X_p$ for some $p \in \mathbb{N}$.

- We assume that there is some relationship between the response Y and the vector of predictors $X = (X_1, \ldots, X_p)$, namely

$$Y = f(X) + \varepsilon. \tag{1}$$

  - f denotes a fixed but unknown function of $X_1, \ldots, X_p$.
  - ε is a random *error term*, which is independent of X, with $E(\varepsilon) = 0$.

- Two main reasons to estimate f: *prediction* and *inference*.

Goal: predict response Y at a set of inputs X.

- If X is available, because the error term averages to zero, we can predict Y using

$$\hat{Y} = \hat{f}(X), \tag{2}$$

  where $\hat{f}$ denotes the estimate for f, and $\hat{Y}$ the resulting prediction for Y.
- For prediction tasks, $\hat{f}$ is often treated as a *black box* – does it accurately predict Y?

Example: The blue surface in Figure 2 is an estimate $\hat{f}$ for the unknown function f describing the relationship of the predictors `years of education` and `seniority` to the response `income`:

$$\widehat{\texttt{income}} = \hat{f}(\texttt{years of education, seniority}).$$

Inference

Goal: learn relationship between response Y and inputs $X_1, \ldots, X_p$.

- One may want to answer:
  - Which predictors are associated with the response?
  - What is the relationship between the response and each predictor?
  - Can we use a linear equation to describe the relationship between $X_1, \ldots, X_p$ to Y, or is there a more complex relationship?
- Knowing more about f allows us to ask questions about Y, such as:
  - What value of $(X_1, \ldots, X_p)$ maximizes Y?
  - How much is Y affected by each predictor X_i?
    E.g., in Figure 1 we might have
    - 60% of $Var(\texttt{sales})$ can be explained by TV budget,
    - 30% of $Var(\texttt{sales})$ can be explained by `Radio` budget,
    - 8% of $Var(\texttt{sales})$ can be explained by `Newspaper` budget,
    - remaining 2% can be explained by $X_4, X_5, \ldots, X_p$
  - These questions can be difficult to answer if f is highly non-linear!
  - `https://www.climateinteractive.org/en-roads/`

How to estimate f?

Always assume we have observed a set of n different data points.

- These are called the *training data* because these observations will train our method on how to estimate f.

More precisely:

- We index the observations by $i = 1, \ldots, n$.
- We index the inputs/predictors by $j = 1, \ldots, p$.
- Let x_{ij} represent the value of the jth predictor for the ith observation.
- Let y_i represent the response variable for the ith observation.
- Then our training data consist of $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$.

Goal: apply a statistical learning method to the training data to estimate the unknown function f.

- I.e., find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observed (X, Y).
- Most statistical learning methods for this task can be characterized as either *parametric* or *non-parametric*.

Parametric methods involve two steps:

1. Define what estimates $\hat{f}$ of the function f are allowed to look like. E.g., allow $\hat{f}$ to be a linear function of $X = (X_1, \ldots, X_p)$:

$$\hat{f}(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p. \tag{3}$$

2. Use the *training data* to *fit*/*train* the model. For the linear model (3), we have to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ so that $Y \approx \hat{f}(X)$.

Called *parametric* because $\hat{f}$ is determined by finitely many parameters.

- Advantage: Reduces the problem of estimating an arbitrary p-dimensional function f to the easier problem of estimating only $p + 1$ parameters.
- Disadvantage: The allowed form of $\hat{f}$ likely will not match true form of f. One could try to use more *flexible* models (however, this could lead to *overfitting* the data, i.e. the models follows the errors too closely).
- The left plot in Figure 1 estimates the function f by

$$\hat{f}(\texttt{TV}, \texttt{Radio}, \texttt{Newspaper}) = \hat{f}(\texttt{TV}) \approx 6 + \frac{1}{20} X_1. \tag{4}$$

(Language: “make an assumption about the functional form of f”)

Non-parametric (or *nonparametric*) methods

Non-parametric methods: $\hat{f}$ is determined by infinitely many parameters.

- Misnomer: non-parametric $\neq$ no parameters!
- Advantage: more flexibility in what $\hat{f}$ is allowed to be $\Rightarrow$ more likely that $\hat{f}$ better estimates f.
- Disadvantage: requires a large number of observations to accurately estimate f.
- Disadvantage: inference is often more difficult.

Example: (next slide)

Example – non-parametric

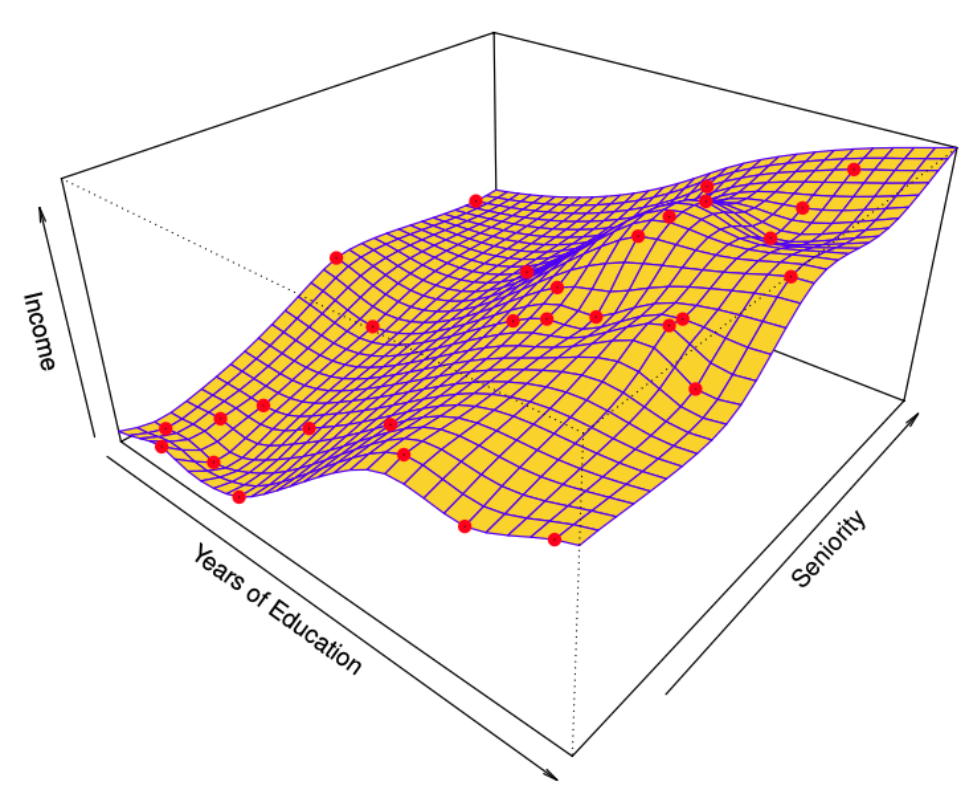

Figure 3: Image by James et al. (2021), based on the `Income` data set in R. "A rough thin-plate spline fit to the Income data from Figure 2.3. This fit makes zero errors on the training data." Here the yellow surface need only be continuous.

- Interpolation theorem: Any n bivariate data points (x_i, y_i) (where x_is are distinct) can be interpolated by a polynomial of order at most $n - 1$. `https://en.wikipedia.org/wiki/Polynomial_interpolation`

Choosing a model: interpretability vs flexibility

The choice of the model (e.g., model (1)) depends on whether we are interested in prediction, inference, or a combination of both.

- Simpler models typically are easier to interpret and make inference on.
- More flexible models might more accurately predict Y but are often less interpretable.
- Example: Simple linear regression allows an intercept and a slope (see Figure 1) compared to a quadratic/cubic/polynomial regression where even more parameters can be chosen (see Figure 2).

When choosing a statistical model for the data, there is usually a trade-off between interpretability and flexibility.

- One has to choose where to be on this spectrum (see next slide).

Choosing a model: interpretability vs flexibility

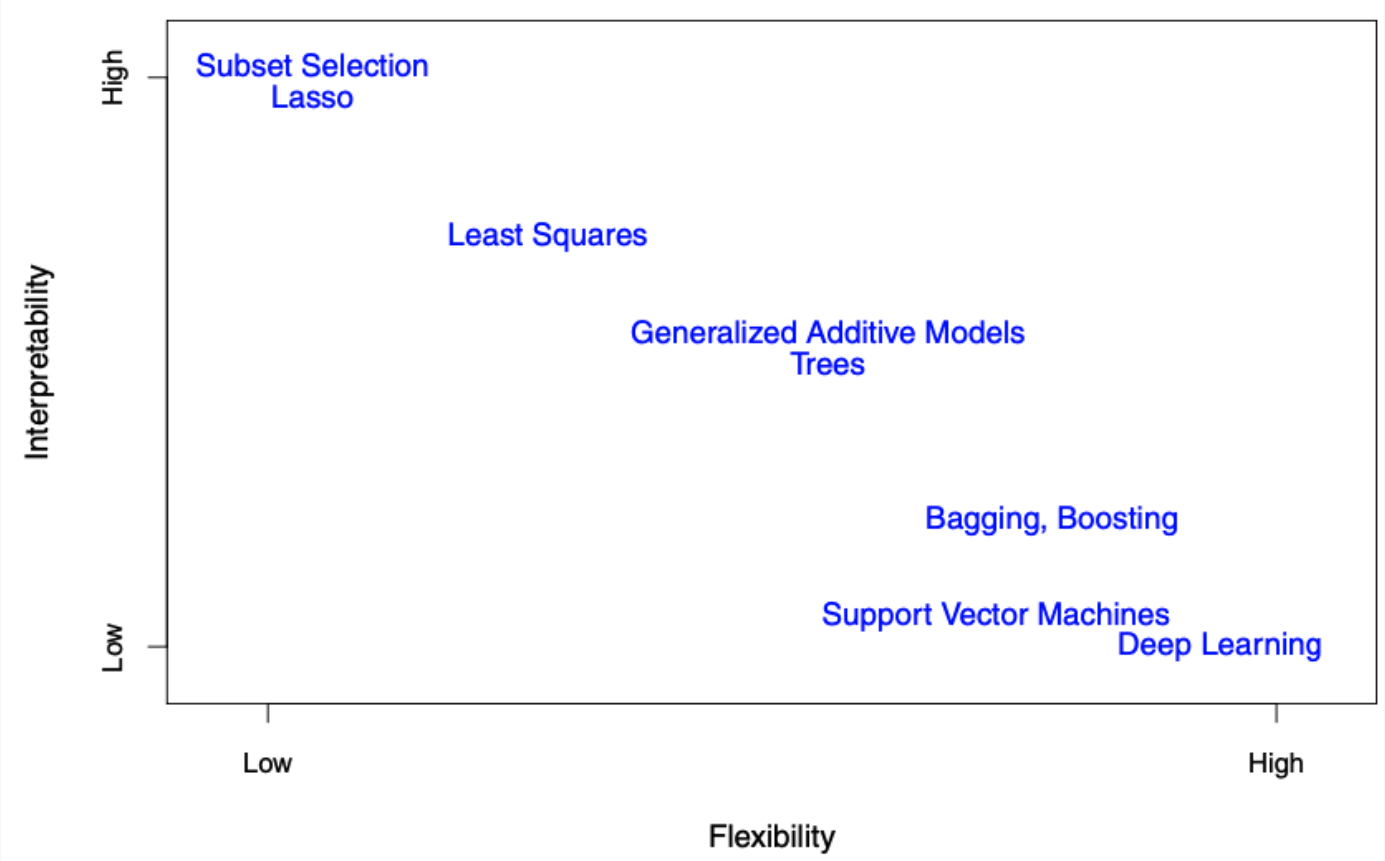

Figure 4: Image by James et al. (2021). "A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods."

Variables can be treated as:

- *quantitative* (numeric). E.g., age, height, income, house value, stock price.
- *qualitative/categorical* ("discrete" – value is one of K classes). E.g., marital status (married or not), brand of product purchased (brand A, B, or C).

One tends to select the statistical learning methods on the basis whether the response Y is quantitative or qualitative.

- Whether the predictors are quantitative or qualitative is less important.
- Problems with a quantitative response are usually referred to as *regression* problems.
- Problems with a qualitative response are usually referred to as *classification* problems.
  - Techniques include *logistic regression* and *K-nearest neighbors* (see Figure 5).
- The distinction is not always crisp; logistic regression can be thought of as either classification or regression.

Example of classification – K-nearest neighbors

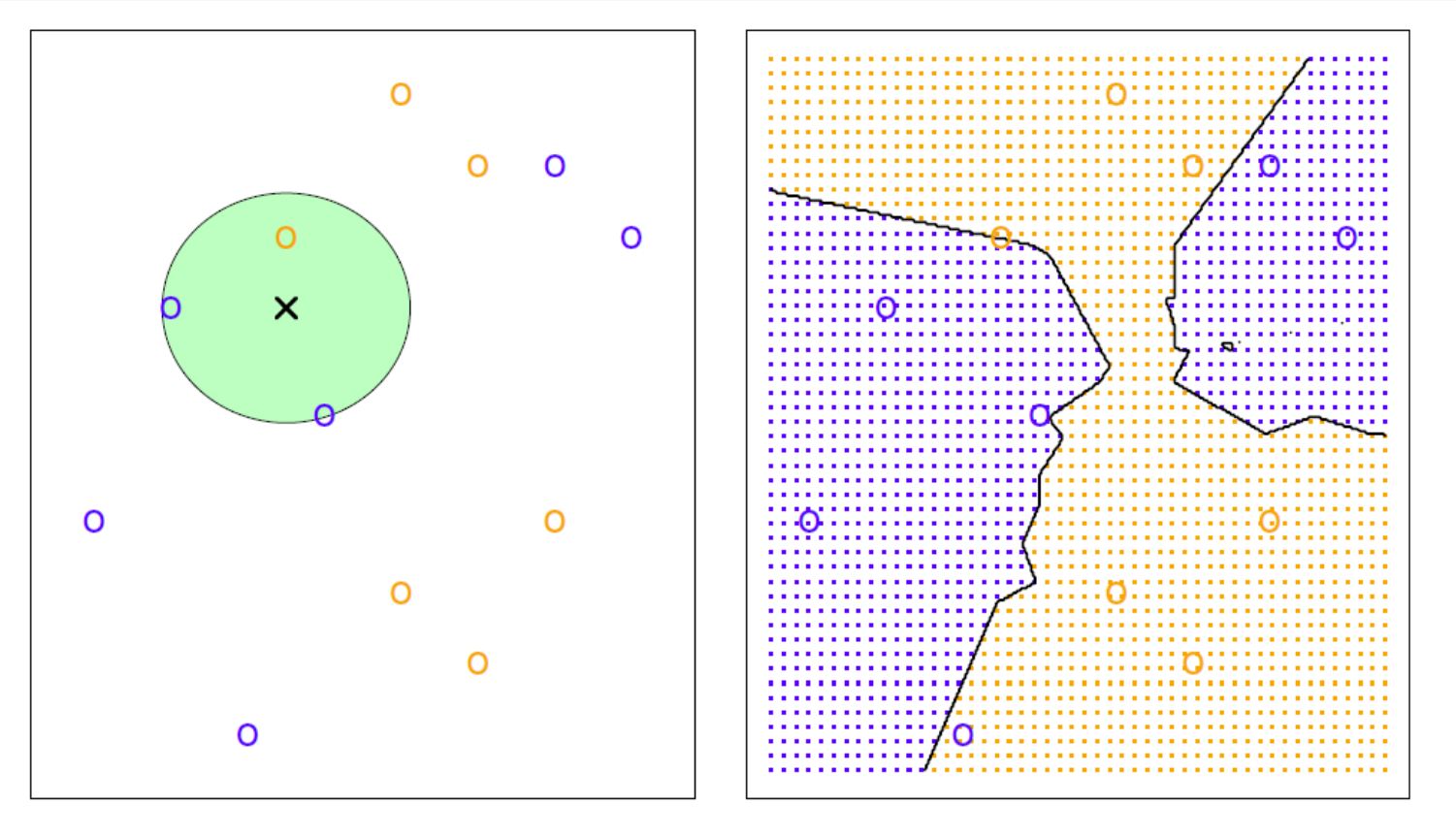

Figure 5: Image by James et al. (2021). Classification using the K-nearest neighbor approach with $K = 3$. Left: We assign a new observation (this is the "x") to the class for which most of three neighbors of "x" belong to. Right: A decision line/region how we would assign a new element for $K = 3$ if they would fall in a certain indicated region.

Section 5: Overview of Statistical Learning

Unsupervised learning

Overview

Recall: learn relationship or structure among observations. Examples include

- *Dimension reduction*: derive a low-dimensional set of features from higher-dimensional observations $X_1, \ldots, X_n$.
  - Uses: plotting 2-d representations of higher-dimensional data, regression.
  - *Principal components analysis* is a popular approach discussed later in this course.
- *Cluster analysis*: partition observations $X_1, \ldots, X_n$ into distinct groups.

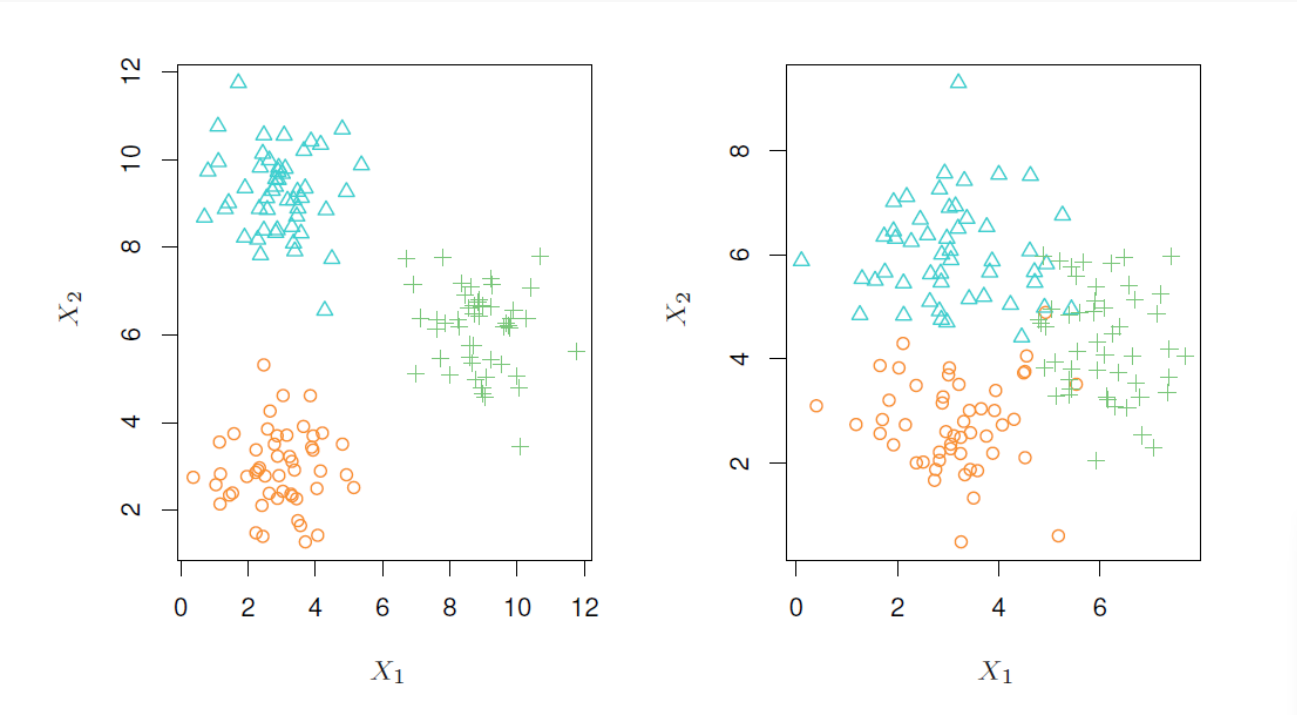

Figure 6: Image by James et al. (2021). Clustering in a data set involving three groups.

Example of clustering – flow cytometry

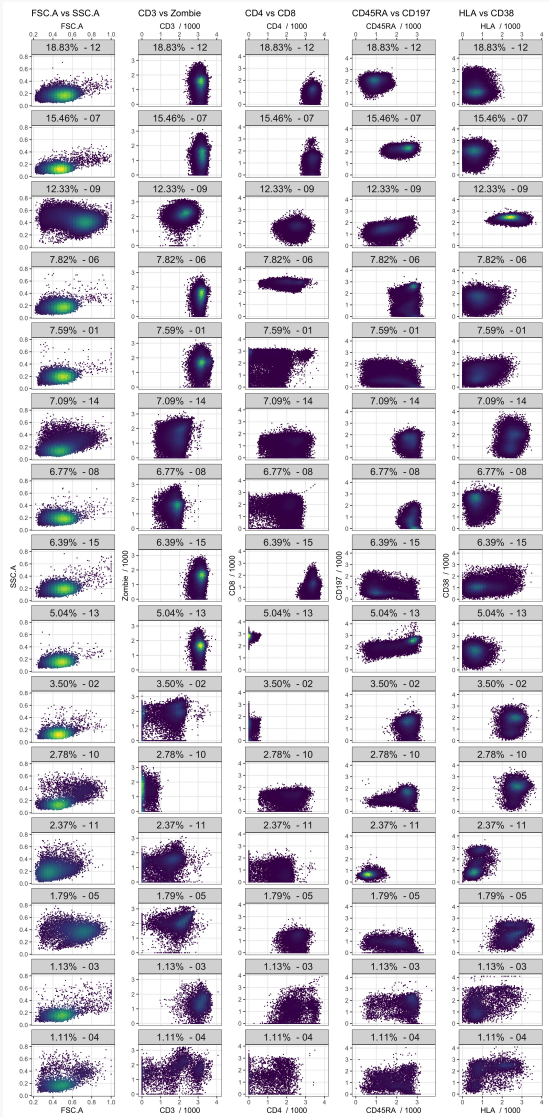

Figure 7: Image by Horiguchi et al. (2024) – `https://arxiv.org/abs/2208.02806`