Sec 5: Main concepts of statistical learning

STA 141A – Fundamentals of Statistical Data Science

Instructor: Akira Horiguchi

Fall Quarter 2025 (Sep 24 – Dec 12)
MWF, 9:00 AM – 9:50 AM, TLC 1215
University of California, Davis

Outline

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

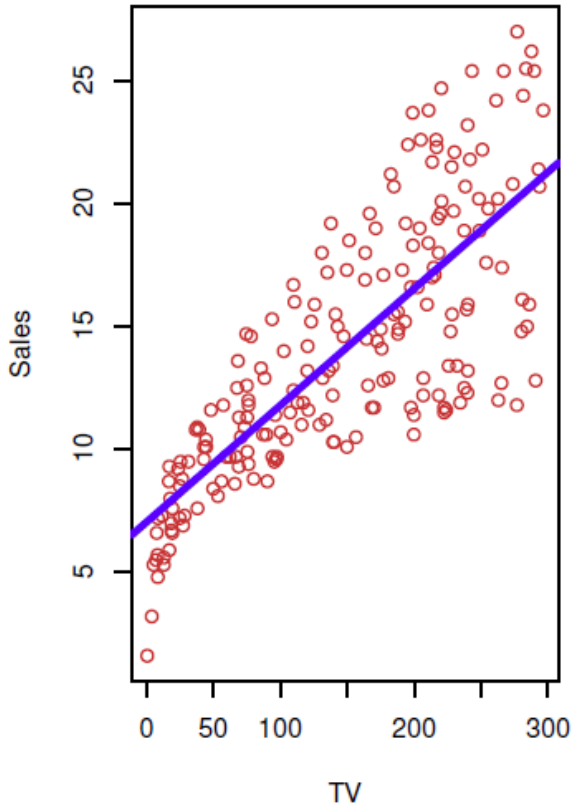
1. Motivation

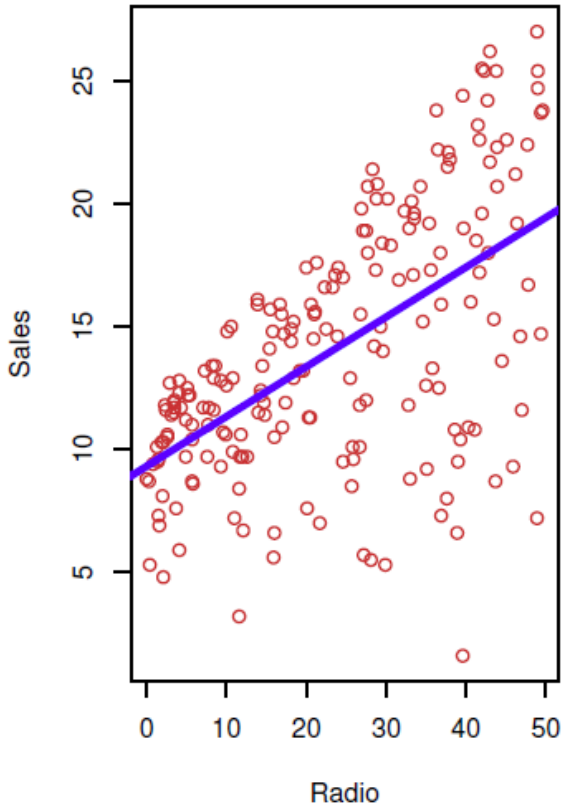
2. Unsupervised learning

3. Supervised learning
   - Learning goals and tasks
   - Choosing a model
   - Assessing model accuracy

Motivation

How does advertising affect sales? (supervised learning)





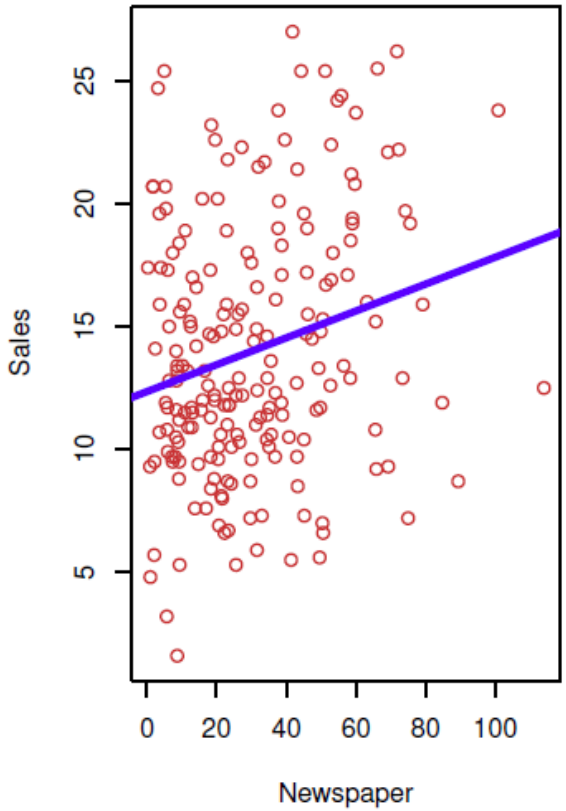


Figure: Image by James et al. (2021), based on the `Advertising` data set in R. The plot diplays `sales` in thousands of units depending on the input TV, `radio` and `newspaper` (advertising) budgets, in thousand dollars, for 200 different marktes.

Who will default on credit card payment? (supervised learning)

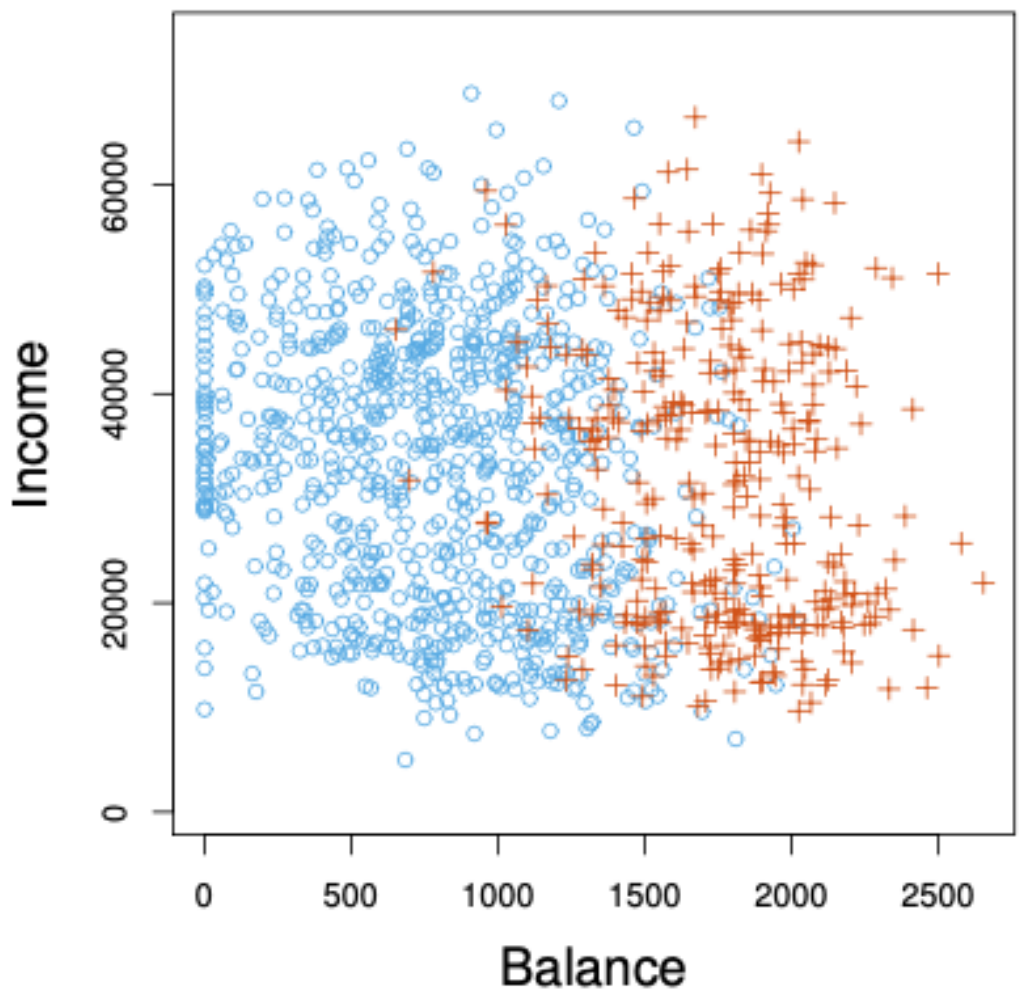


Figure: Image by James et al. (2021). The `Default` data set. The annual incomes and monthly credit card balances of a number of individuals. Orange +s indicate individuals who defaulted on their credit card payments; blue circles indicate individuals who did not default.

Flow cytometry (unsupervised learning)

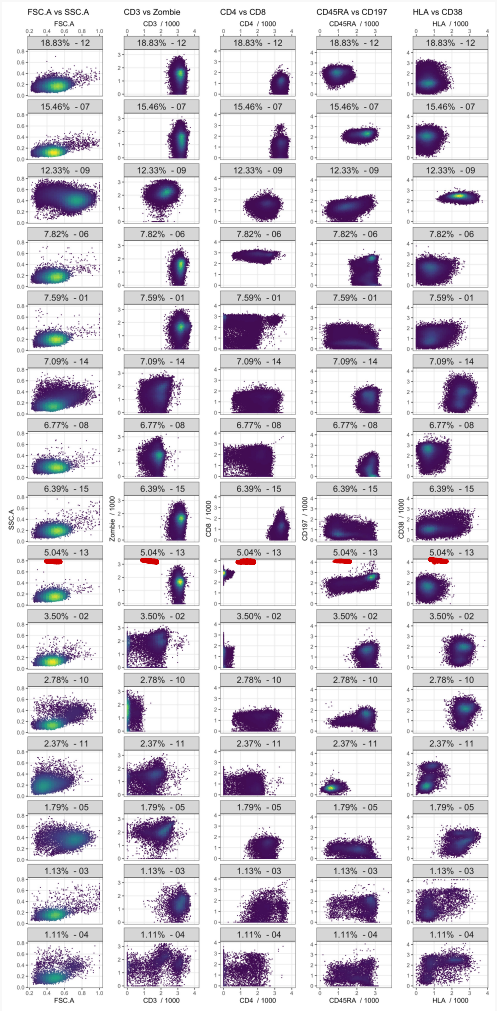


Figure: Image by Horiguchi et al. (2024) – https://projecteuclid.org/journals/bayesian-analysis/advance-publication/A-Tree-Perspective-on-Stick-Breaking-Models-in-Covariate-Dependent/10.1214/24-BA1462.full

Statistical learning: supervised vs unsupervised

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

- *Supervised* statistical learning: predict or estimate an output based on one or more inputs. (STA 142A)
- *Unsupervised* statistical learning: learn relationship or structure among observations. (STA 142B)
- (Are there outputs to "supervise" the learning task?)

Unsupervised learning

Overview

Recall: learn relationship or structure among observations. Example tasks:

- *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.
- *Cluster analysis*: partition observations $X_1, \ldots, X_n$ into distinct groups.

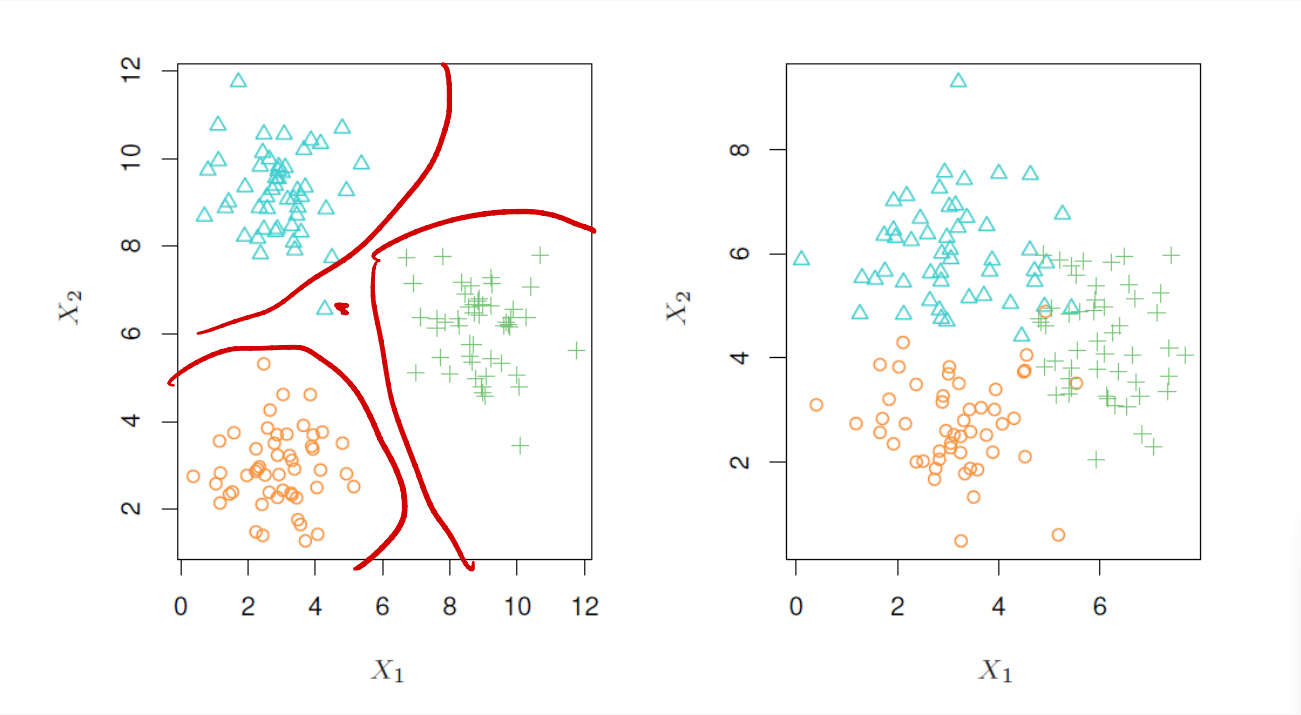


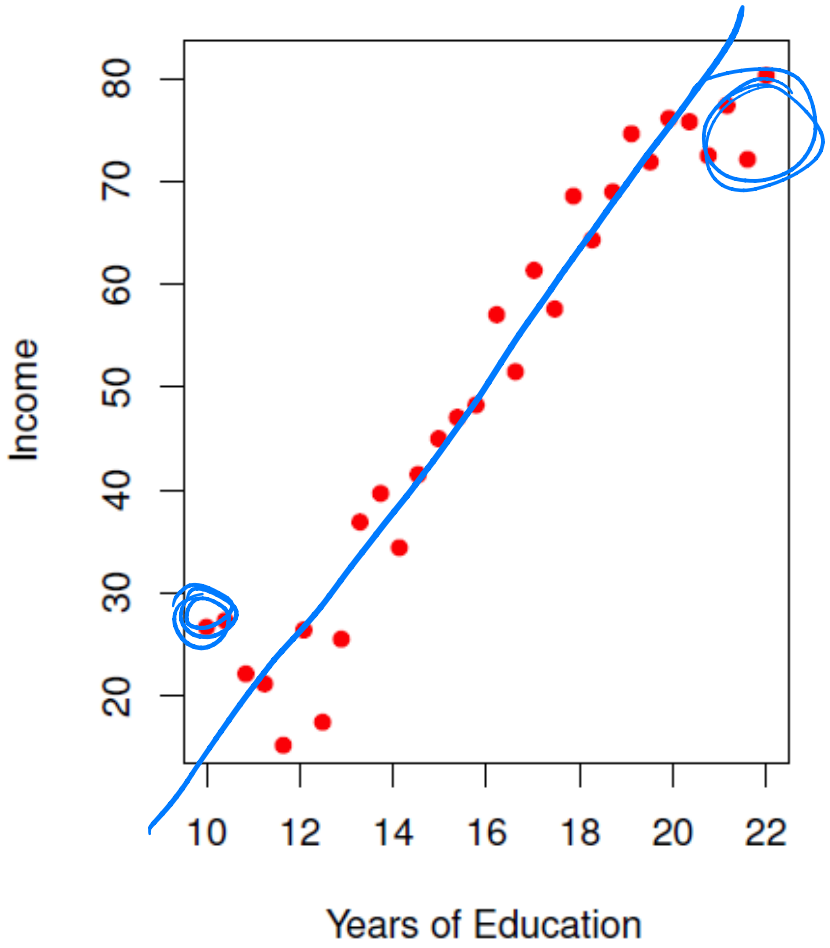
Figure: Image by James et al. (2021). Clustering in a data set involving three groups.

Supervised learning

Supervised learning

Learning goals and tasks

Non-linear regression



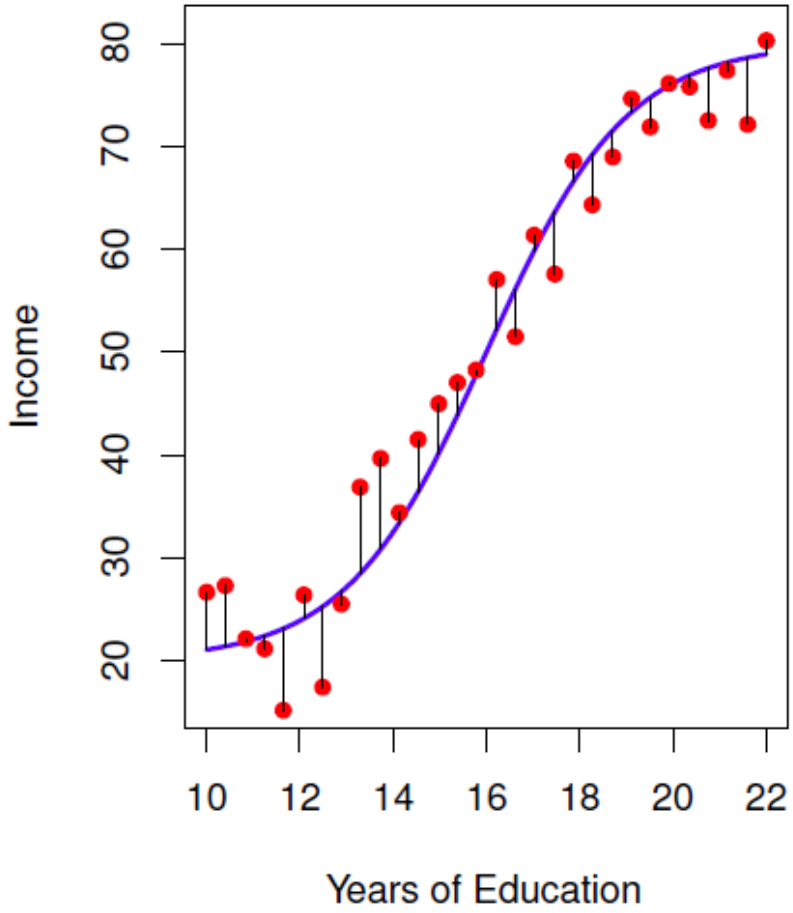


Figure: Image by James et al. (2021), based on the `Income` data set in R. The red dots are the observed values of `income` in tens of thousand dollars and `years of education` for 30 individuals.

A three-dimensional plot

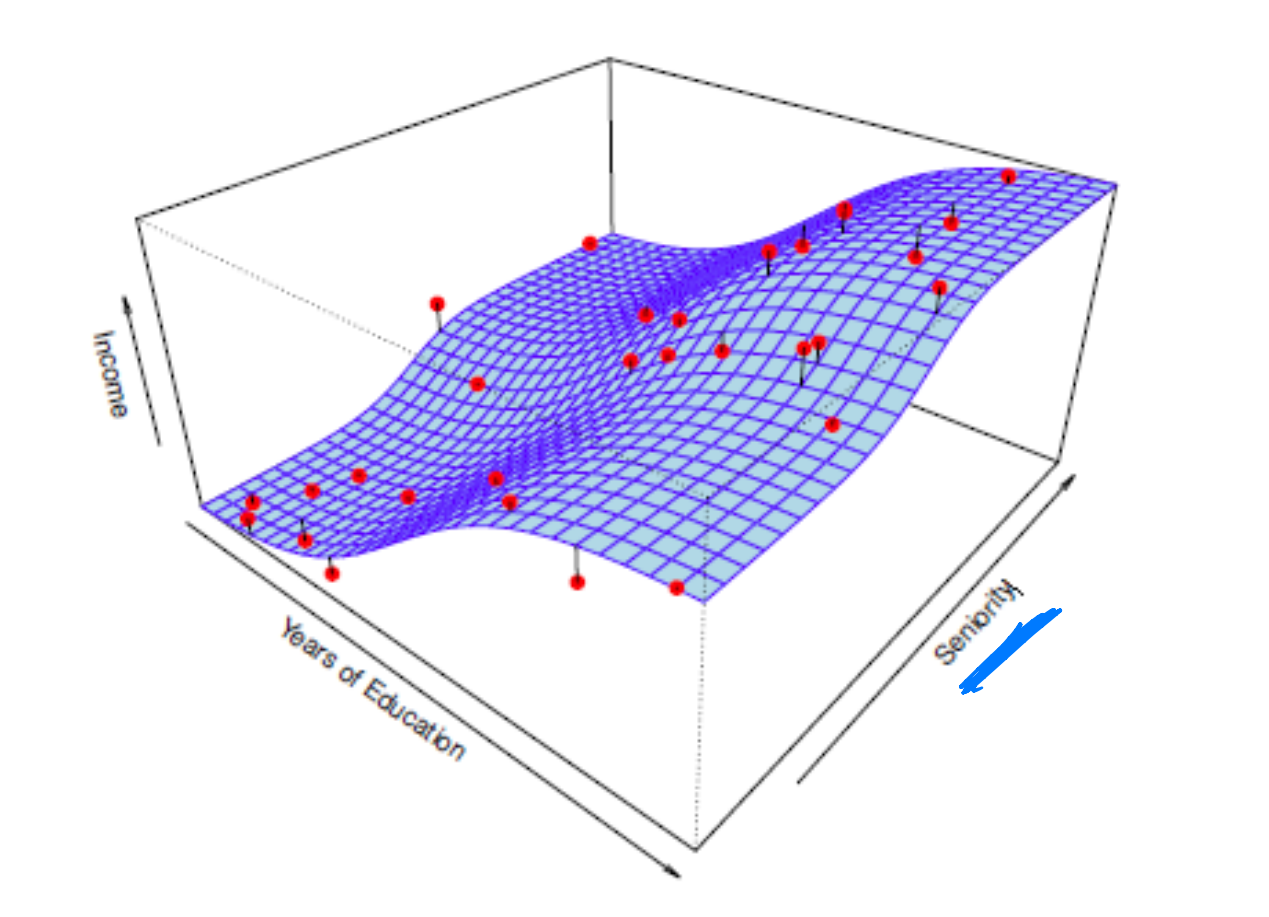


Figure: Image by James et al. (2021), based on the `Income` data set in R. The `income` is displayed as a function of `years of education` and `seniority`.

Combined plots

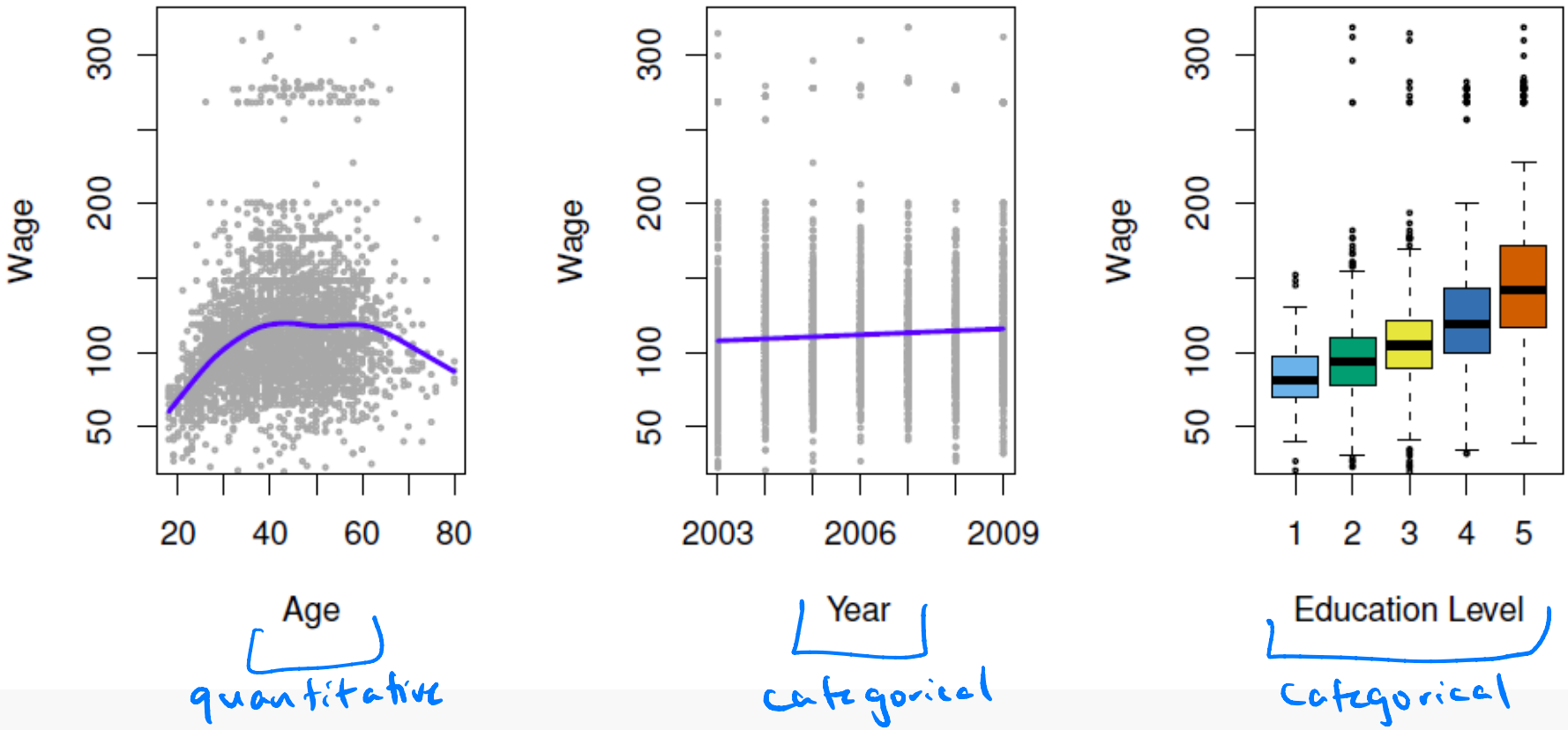


Figure: Image by James et al. (2021), based on the Wage data set in R. The wage is displayed as a function of age, year and education.

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 numeric response Y and p predictors $X_1, \ldots, X_p$.

- We assume that there is some relationship between Y and $X = (X_1, \ldots, X_p)$:

$$Y = f(X) + \varepsilon. \quad \text{"epsilon"} \tag{1}$$

  - 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*.

Prediction – i) Idea

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 6 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`:

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

Prediction – ii) Accuracy

$\hat{Y} = \hat{f}(X)$ $\quad Y = f(X) + \varepsilon$

The accuracy of $\hat{Y}$ for predicting Y depends on *reducible error* and *irreducible error*.

- For a fixed estimator $\hat{f}$ and X, we get

random fixed

$$E\left[\left\{Y - \hat{Y}\right\}^2\right] = E\left[\left\{f(X) - \hat{f}(X) + \varepsilon\right\}^2\right]$$

$E[\varepsilon^2] = Var(\varepsilon) + [E\varepsilon]^2$

$(a+\varepsilon)^2 = a^2 + \varepsilon^2 + 2a\varepsilon$

where $a = \{f(X) - \hat{f}(X)\}$

$$= E\left[\left\{f(X) - \hat{f}(X)\right\}^2\right] + E\left[\varepsilon^2\right] + 2E\left[\left\{f(X) - \hat{f}(X)\right\}\varepsilon\right]$$

$2\{f(x) - \hat{f}(x)\}\, E\{\varepsilon\}$

$$= \underbrace{\left\{f(X) - \hat{f}(X)\right\}^2}_{\text{reducible error}} + \underbrace{Var(\varepsilon)}_{\text{irreducible error}}.$$

End of 10/22 lecture

- A more appropriate learning technique might reduce the *reducible error*.
- *irreducible error* comes entirely from the inherent observation noise ε; independent of how we estimate f.
  - Even if we would estimate f perfectly, i.e. $\hat{Y} = f(X)$, there is still some *irreducible* prediction error from ε, since $Y = f(X) + \varepsilon$.

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/`

Regression vs. classification

Describe problem based on whether the *response* Y is quantitative or qualitative.

- *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).

Regression vs classification

- Problems w/ a quantitative response: usually referred to as *regression* problems.
- Problems w/ a qualitative response: usually referred to as *classification* problems.

Further notes:

- Distinction is not always crisp; is logistic regression classification or regression?
- Whether the predictors $X_1, \ldots, X_p$ are quantitative or qualitative is less important.

Example of classification – *K*-nearest neighbors

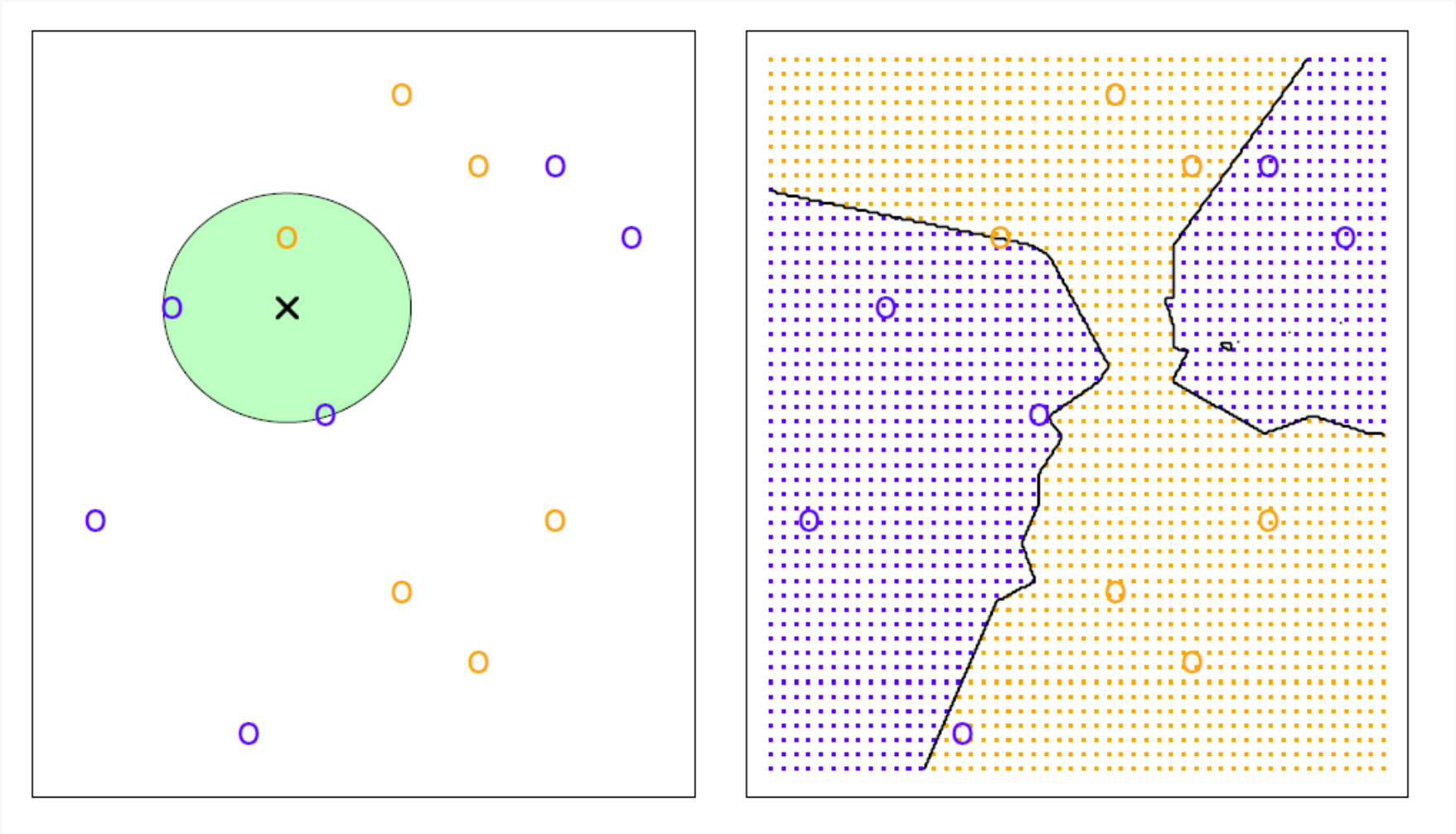


Figure: Image by James et al. (2021). Classification using the *K*-nearest neighbor approach with $K = 3$. **Left**: We assign a new observation (this is the "x") to the class for which most of three neighbors of "x" belong to. **Right**: A decision line/region for how we would assign a new element for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

Supervised learning

Choosing a model

Estimate f

Supervised learning: estimate unknown function f using n observed data points

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top \in \mathbb{R}^p$.

- We index the observations by $i = 1, \ldots, n$.
- We index the inputs/predictors by $j = 1, \ldots, p$.
- Let $x_{ij} \in \mathbb{R}$ represent the value of the jth predictor for the ith observation.
- Let $y_i \in \mathbb{R}$ represent the response variable for the ith observation.

Choosing a model

A statistical learning method will use the data to estimate the unknown f.

- I.e., compute a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observed (X, Y).
- What kind of function is $\hat{f}$ allowed to be? Dictated by our chosen learning method.
- Examples: linear model

$$\hat{f}_{lm}(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

  or linear model plus two-way interactions

$$\hat{f}(X) = \hat{f}_{lm}(X) + \beta_{1,2} X_1 X_2 + \beta_{1,3} X_1 X_3 + \cdots + \beta_{2,3} X_2 X_3 + \cdots + \beta_{p-1,p} X_{p-1} X_p$$

  or quadratic model

$$\hat{f}(X) = \hat{f}_{lm}(X) + \beta_{1,2} X_1^2 + \beta_{2,2} X_2^2 + \cdots + \beta_{p,2} X_p^2$$

Pros and cons to enlarging the class of possible functions that we allow $\hat{f}$ to be:

- Advantage: larger class of possible $\hat{f} \Rightarrow$ 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.

Choosing a model: too much flexibility (1-D)

Consider polynomial regression

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: too much flexibility (2-D)

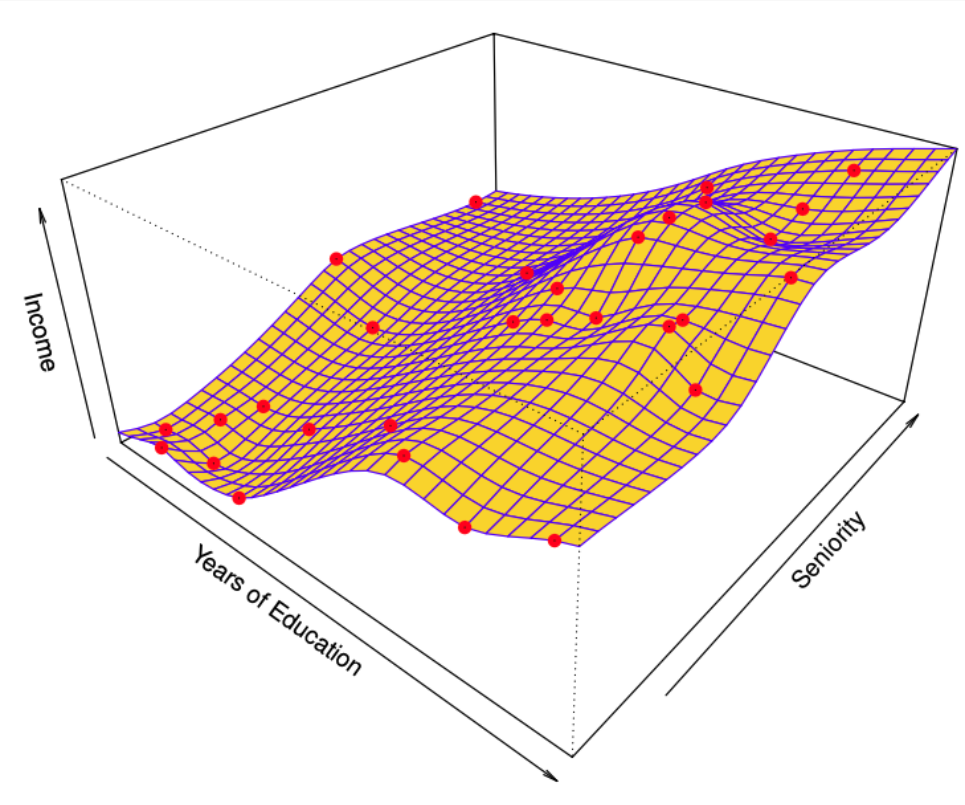


Figure: Image by James et al. (2021), based on the `Income` data set in R. "A rough thin-plate spline fit to the Income data from Figure 2.3. This fit makes zero errors on the training data."

Choosing a model: interpretability vs flexibility

Are we more interested in inference or prediction?

- Simpler models typically are easier to interpret and make inference on.
- Flexible models might more accurately predict Y but can be less interpretable.

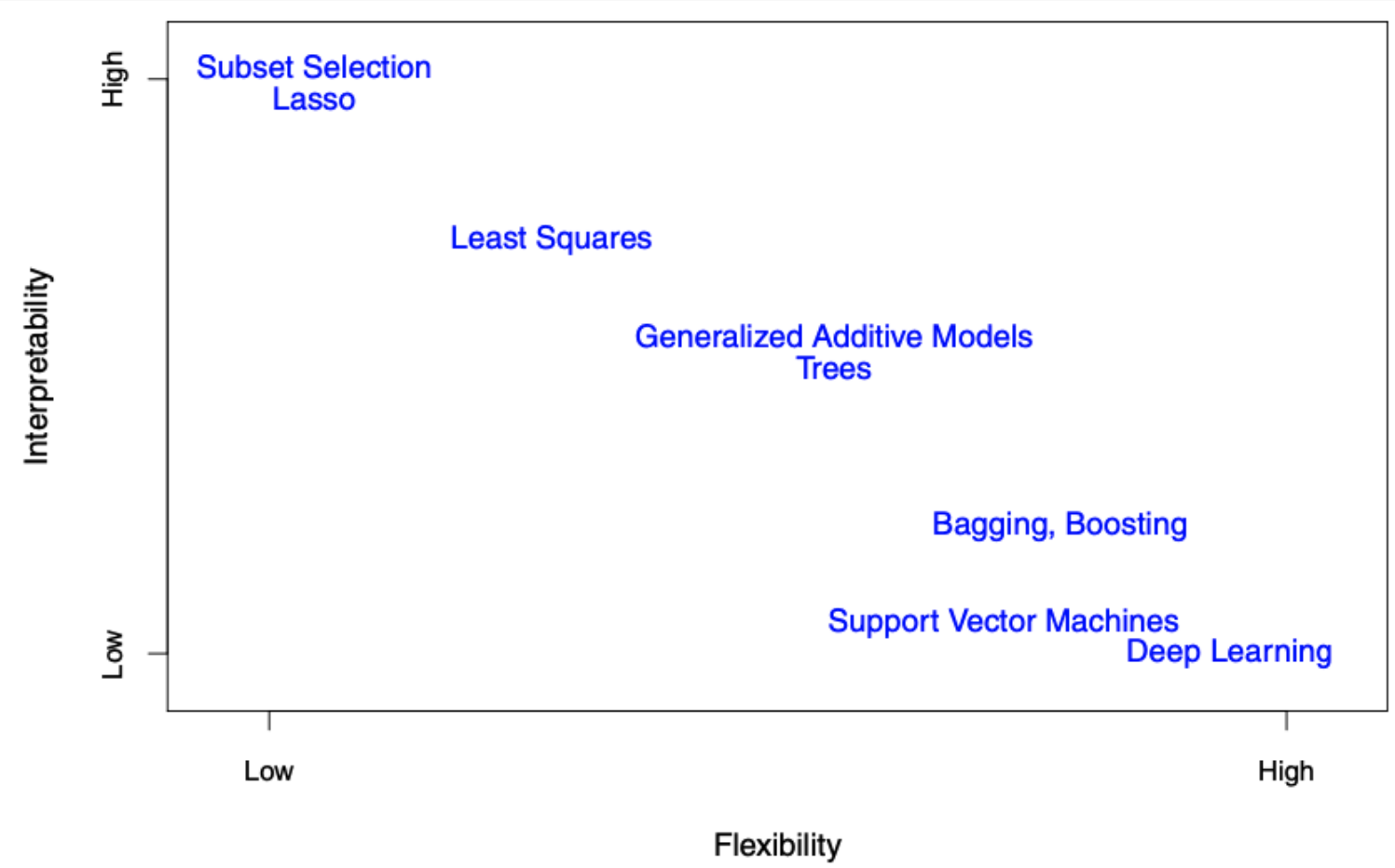


Figure: Image by James et al. (2021). No method dominates *all* other methods.

Choosing a model: final comments

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

Supervised learning

Assessing model accuracy

How do we assess model accuracy

Need to establish an error metric. Different for regression vs classification.

Regression: idea

In regression, most commonly use the *mean squared error* (MSE):

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i) \right)^2 , \tag{3}$$

for n data points $(x_1, y_1), \ldots, (x_n, y_n)$.

- MSE computed on the *training data* is called the *training MSE*, denoted by MSE_{train}.
- We want to choose an estimator $\hat{f}$ that produces a small MSE on data it wasn't trained on:

$$MSE_{test} = \frac{1}{m} \sum_{i=1}^{m} \left(y_{n+i} - \hat{f}(x_{n+i}) \right)^2 , \tag{4}$$

  for m data points $(x_{n+1}, y_{n+1}), \ldots, (x_{n+m}, y_{n+m})$. We call (4) the *test MSE*.

Emphasize: set of training data should be disjoint from set of test data.

Regression: a sketch

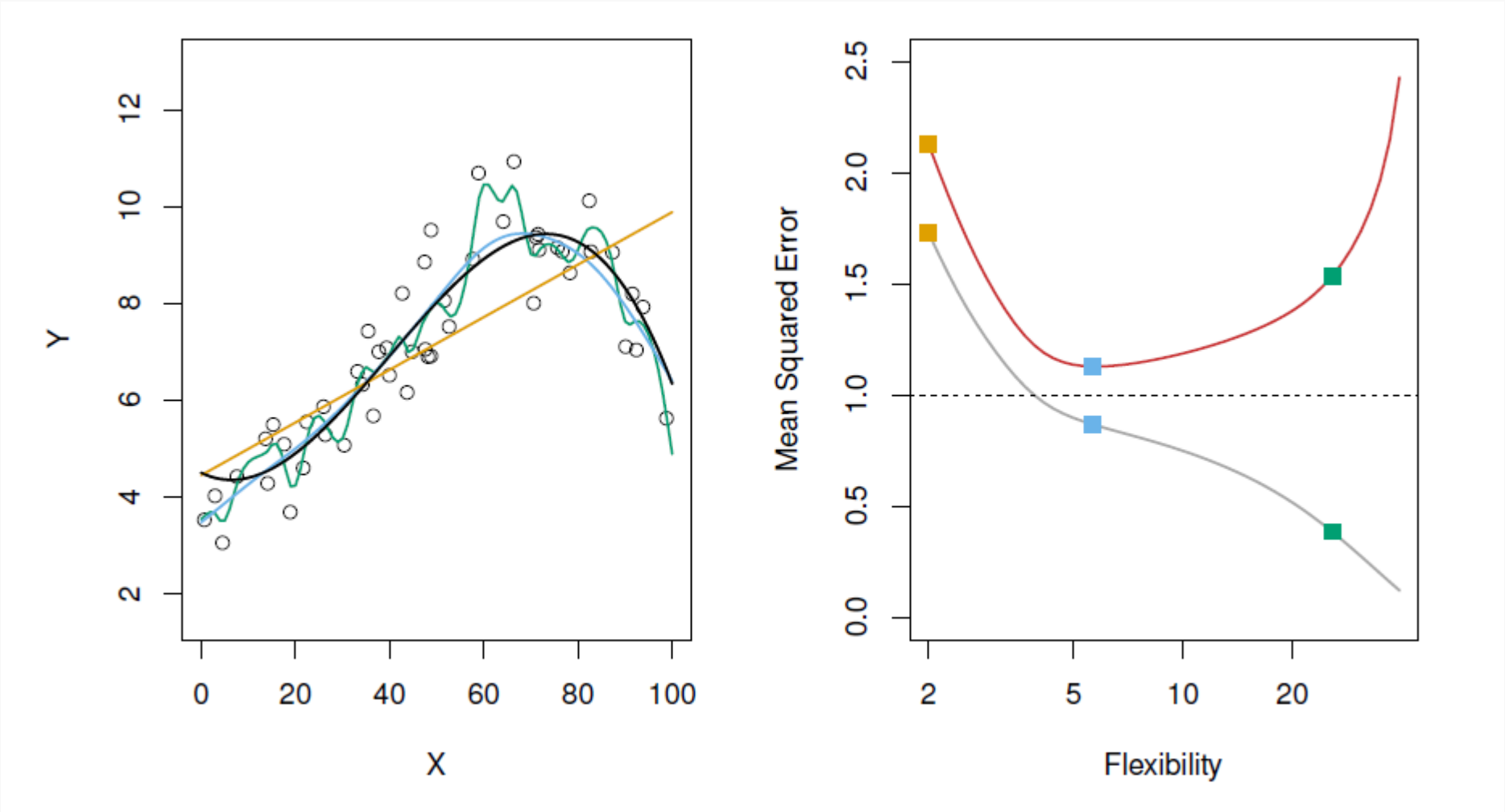


Figure: Image by James et al. (2021). **Left**: Data simulated from f (in black). Three estimates for f: linear regression (in orange), and two smoothing splines (in blue and green). **Right**: Training MSE (in grey), test MSE (in red), and minimum possible test MSE over all methods (dashed line). The squares represent the three fits from the left panel.

Regression: comments on the sketch

- The dashed line in Figure 8 indicates the irreducible error (this is $Var(\varepsilon)$) which is the lowest achievable test MSE among all possible methods.
- Training MSE *always* decreases as we increase model flexibility.
- The U-shape in the test MSE in Figure 8 indicates that it decreases up to a certain amount of flexibility, but gets worse afterwards.
- When a given method yields a small training MSE, but a large test MSE, we say that the data points are *overfitted*.

Regression: another sketch

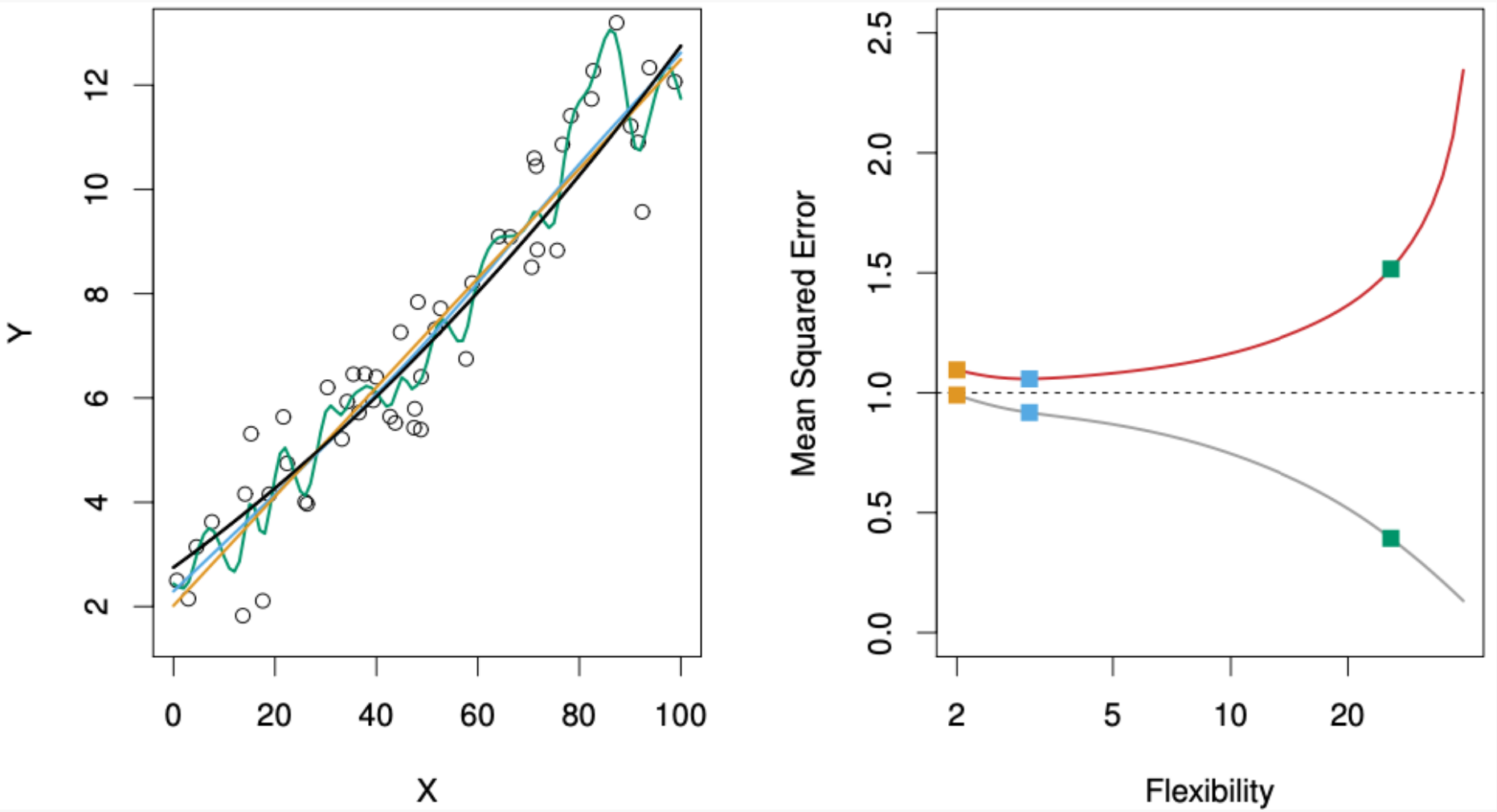


Figure: Image by James et al. (2021). Details are same as in previous figure, but using a different true f that is much closer to linear. In this setting, linear regression provides a very good fit to the data. U-shape is barely noticeable.

Regression: Bias-variance trade-off

U-shape results from two competing properties: *(squared) bias* and *variance*.

- Consider the *expected test MSE* at a new test data point (x, y).
  This is the average test MSE obtained if we repeatedly estimate f over *infinitely many training data sets drawn from some underlying data distribution.*
- The *expected test MSE* at (x, y) can always be decomposed into the sum:

$$E\left[\left\{y - \hat{f}(x)\right\}^2\right] = Var\left(\hat{f}(x)\right) + \left\{f(x) - E\left[\hat{f}(x)\right]\right\}^2 + Var(\varepsilon). \quad (5)$$

  - $Var\left(\hat{f}(x)\right)$ is the amount $\hat{f}$ would change by using different training data sets.
  - $\left\{f(x) - E\left[\hat{f}(x)\right]\right\}^2$ is the squared bias of $\hat{f}(x)$.
    - Squared bias can be interpreted as the error introduced by approximating f using the given model assumptions (which often do not capture the full complexity of f).
  - $Var(\varepsilon)$ is the variance of the error term ε (this is a property of the data, not of $\hat{f}$).
    - The expected test MSE can never be below $Var(\varepsilon)$. (Dotted horiz line in prev two figures.)

(5) is called a *bias-variance trade-off*. As a general rule, as model flexibility increases, the variance will increase and the (squared) bias will decrease.

Classification: training and test error rate

In classification, we quantify accuracy of $\hat{f}$ using the *training error rate*,

$$Err_{train} = \frac{1}{n}\sum_{i=1}^{n} I\left(y_i \neq \hat{f}(x_i)\right), \tag{6}$$

for n training data points $(x_1, y_1), \ldots, (x_n, y_n)$.

- y_i is the observed class given the observation x_i
- $\hat{f}(x_i)$ is the predicted class given the observation x_i
- *Indicator variable*

$$I\left(y_i \neq \hat{f}(x_i)\right) = \begin{cases} 1 & \text{if } y_i \neq \hat{f}(x_i) \\ 0 & \text{if } y_i = \hat{f}(x_i) \end{cases}$$

- Hence the training error rate is the mean number of wrong classifications.

A good classifier, however, produces a small *test error rate*

$$Err_{test} = \frac{1}{m}\sum_{i=1}^{m} I\left(y_{n+i} \neq \hat{f}(x_{n+i})\right), \tag{7}$$

where $(x_{n+1}, y_{n+1}), \ldots, (x_{n+m}, y_{n+m})$, with $m \in \mathbb{N}$, are test data. *Bias-variance trade-off* also appears in classification, as we will see.

Classification: Bayes classifier as a (theoretical) benchmark

The *Bayes classifier* assigns a predictor value x to the class

$$\arg\max_j P(Y = j|X = x). \tag{8}$$

where $P(Y=j|X=x)$ is the conditional probability that a point whose predictor value is x belongs to class j. E.g., predicting penguin species: Adelie, Chinstrap, or Gentoo.

- This classifier assigns each x to the most likely class given its predictor.
- It turns out that the Bayes classifier produces the lowest possible test error rate, called the *Bayes error rate* — analogous to the irreducible error discussed earlier.
- Can't compute the Bayes classifier if we don't know the cond. distribution of $Y|X$.
- Approximate the Bayes classifier by estimating the conditional probs $P(Y=j|X=x)$.

Classification: KNN – i) Idea

The *K-nearest neighbor (KNN)* classifier estimates $P(Y = j|X = x)$ by

$$\widehat{P(Y = j|X = x)} = \frac{1}{K} \sum_{i \in \mathcal{N}_K(x)} I(y_i = j), \tag{9}$$

where the set $\mathcal{N}_K(x)$ indexes the $K \in \mathbb{N}$ nearest neighbors of x.

- Ultimately, the KNN classifier will assign x to the *majority* class in $\mathcal{N}_K(x)$.

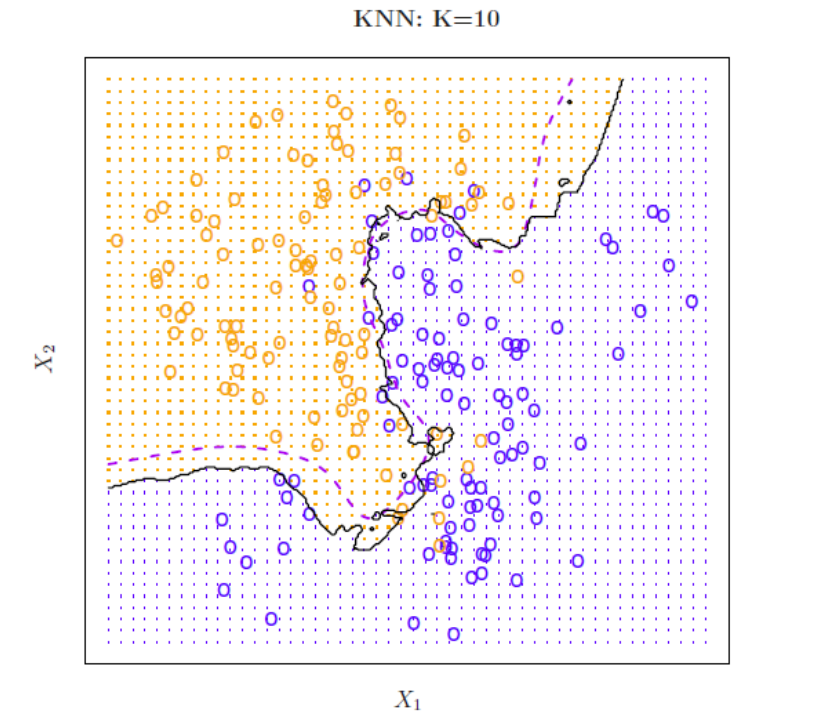


Figure: Image by James et al. (2021). The KNN decision boundary for $K = 10$ (black solid curve), and the Bayes decision boundary (purple dashed curve).

Classification: KNN – ii) Different *K*

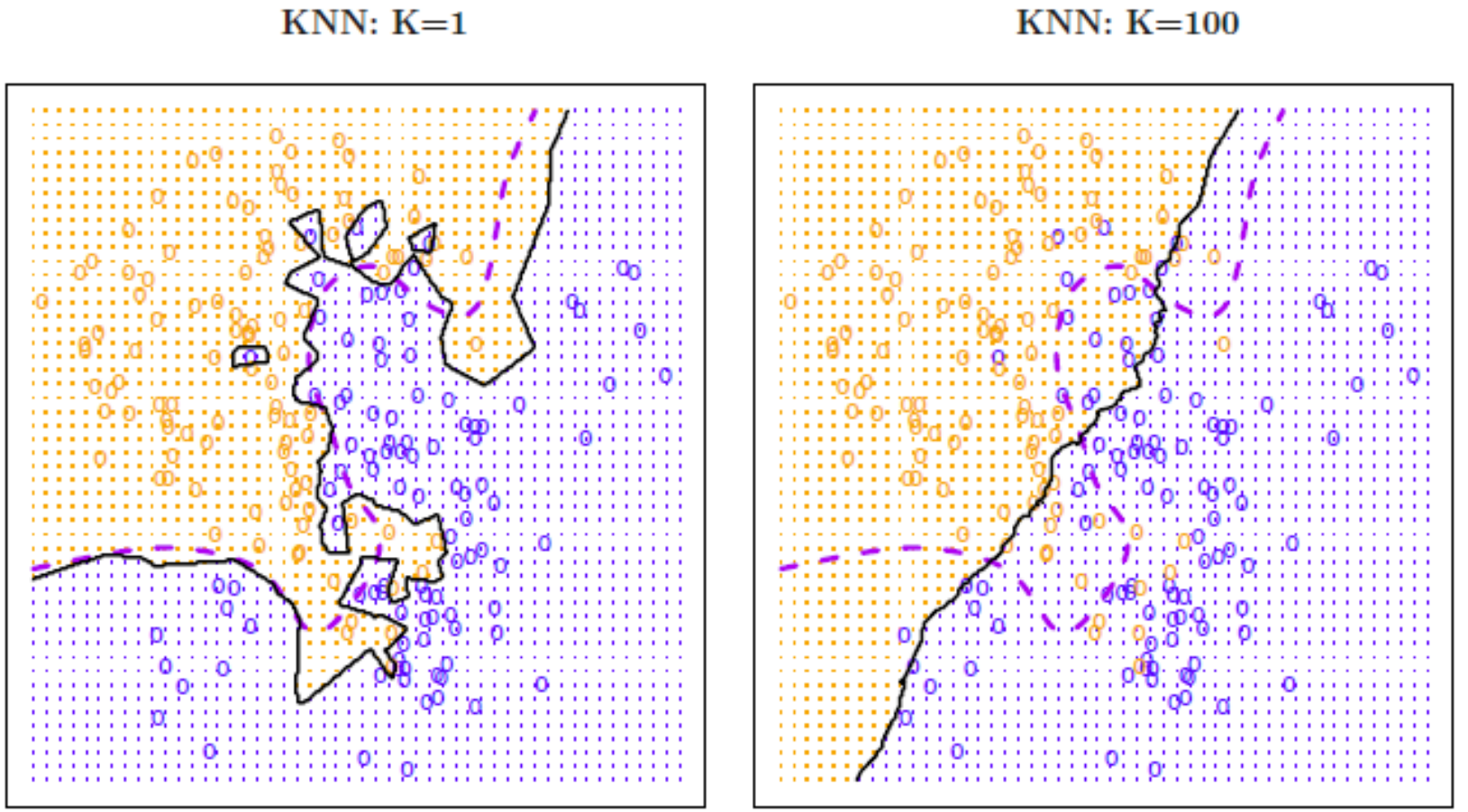


Figure: Image by James et al. (2021). KNN decision boundaries (black solid curve) for $K = 1$ and $K = 100$, and the Bayes decision boundary (purple dashed curve). The $K = 1$ decision boundary is too flexible, while the $K = 100$ boundary is not flexible enough.

Classification: KNN – iii) Bias-variance trade-off

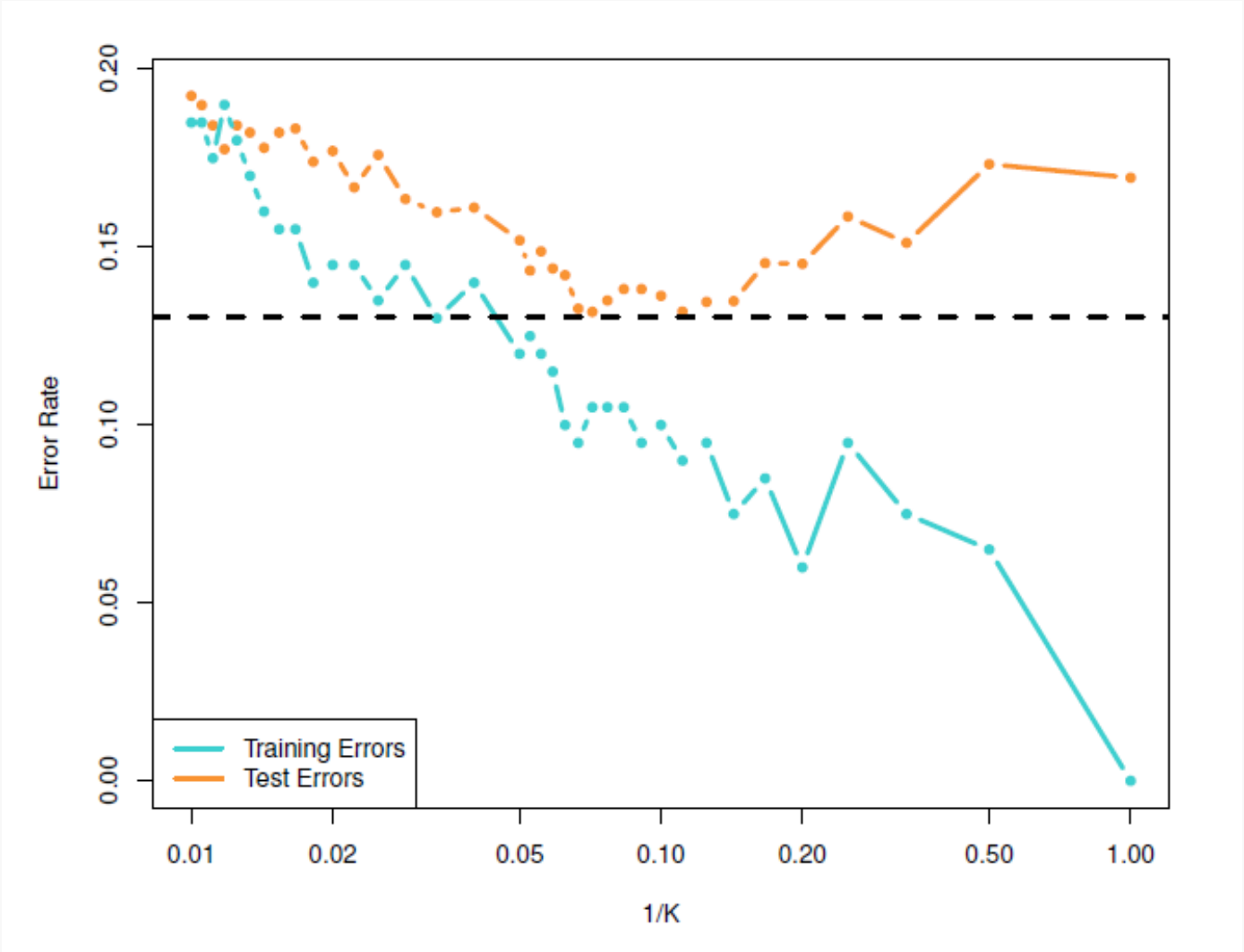


Figure: Image by James et al. (2021). The KNN training error rate (blue) and test error rate (orange) and the Bayes error rate (black horizontal dashed line). The jumpiness of the curves is due to the small size of the training data set.

General comments

We want to choose the method that produces the smallest test error.

- True regardless of regression or classification.