

## Sec 4: Main concepts of statistical learning

STA 141A – Fundamentals of Statistical Data Science (UC Davis, Spring 2026)

**Instructor:** Akira Horiguchi

# Outline

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

## Motivation

### Unsupervised learning

### Supervised learning

Learning goals and tasks

Choosing a model

Assessing model accuracy

# How does advertising affect sales? (supervised learning)

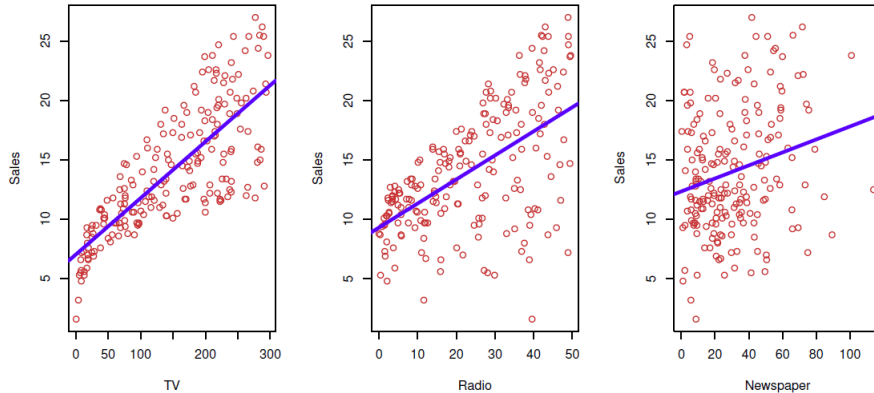


Figure 1: Image by James et al. (2021), based on the Advertising data set in R. The plot displays sales in thousands of units depending on the input TV, radio and newspaper (advertising) budgets, in thousand dollars, for 200 different markets.

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

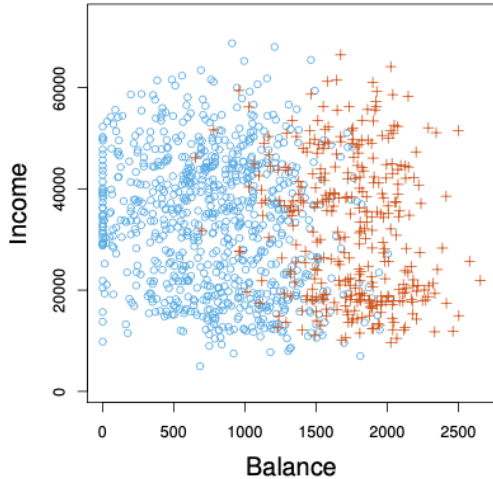


Figure 2: Image by James et al. (2021). The Default data set. The annual incomes and monthly credit card balances of a number of individuals. Orange +s indicate individuals who defaulted on their credit card payments; blue circles indicate individuals who did not default.

# Flow cytometry (unsupervised learning)

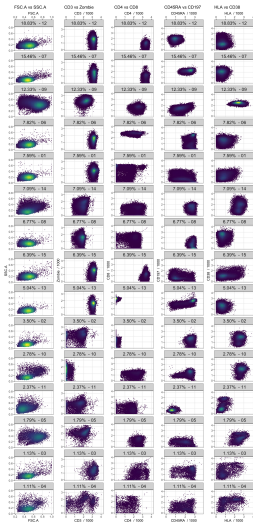


Figure 3: Image by [Horiguchi et al. \(2025\)](#)

# Statistical learning: supervised vs unsupervised

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

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

Motivation

**Unsupervised learning**

Supervised learning

Learning goals and tasks

Choosing a model

Assessing model accuracy

# Overview

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

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

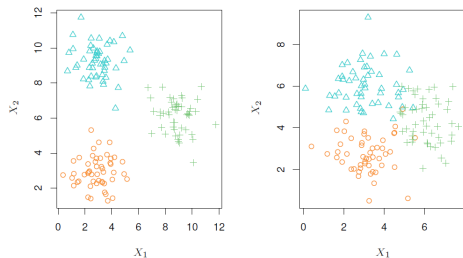


Figure 4: Image by James et al. (2021). Clustering in a data set involving three groups.

- Does heterogeneity exist in our data? (e.g. exam scores)
- We know that clusters exist; we want to accurately capture them. (e.g. flow cytometry)

Motivation

Unsupervised learning

**Supervised learning**

Learning goals and tasks

Choosing a model

Assessing model accuracy

Motivation

Unsupervised learning

Supervised learning

Learning goals and tasks

Choosing a model

Assessing model accuracy

## Non-linear regression

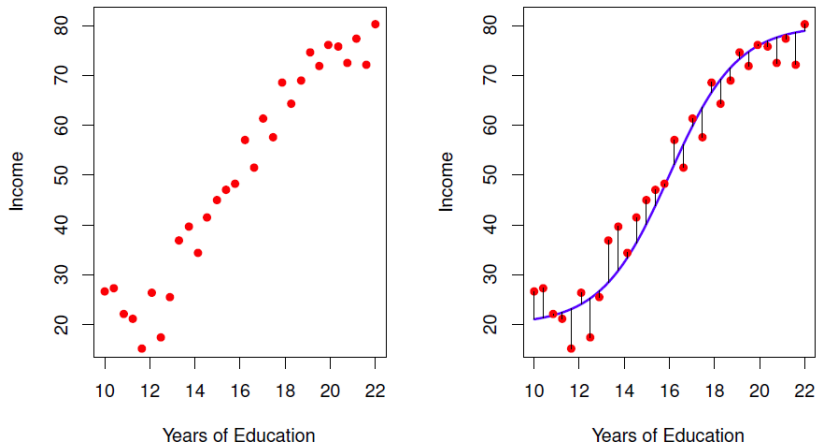


Figure 5: Image by James et al. (2021), based on the `Income` data set in R. The red dots are the observed values of `income` in tens of thousand dollars and `years of education` for 30 individuals.

## A three-dimensional plot

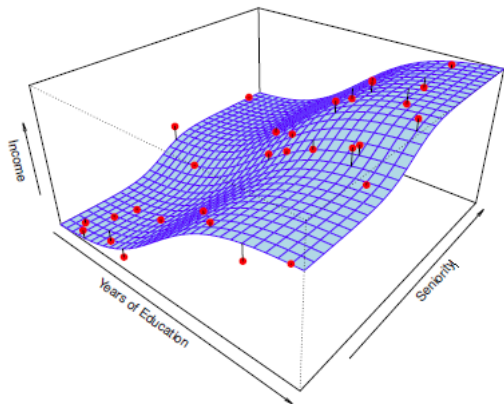


Figure 6: Image by James et al. (2021), based on the Income data set in R. The income is displayed as a function of years of education and seniority.

## Combined plots

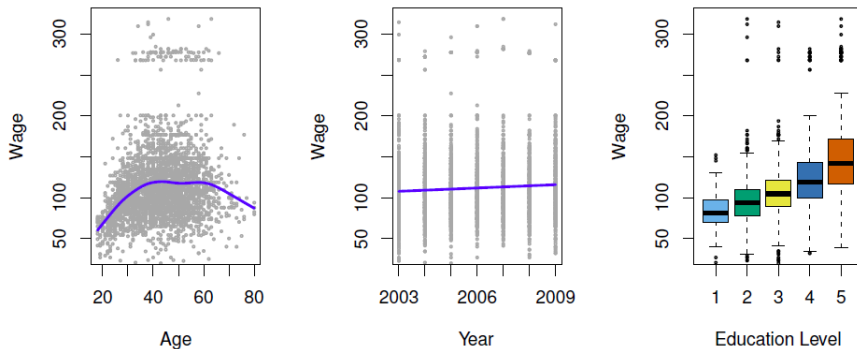


Figure 7: Image by James et al. (2021), based on the Wage data set in R. The wage is displayed as a function of age, year and education.

## 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  $X_1, \dots, X_p$ .

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

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

- $f$  denotes a fixed but unknown function of  $X_1, \dots, X_p$ .
- $\varepsilon$  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

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

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:

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

## Prediction – ii) Accuracy

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

$$\begin{aligned} E \left[ \left\{ Y - \hat{Y} \right\}^2 \right] &= E \left[ \left\{ f(X) - \hat{f}(X) + \varepsilon \right\}^2 \right] \\ &= E \left[ \left\{ f(X) - \hat{f}(X) \right\}^2 \right] + E \left[ \varepsilon^2 \right] + 2E \left[ \left\{ f(X) - \hat{f}(X) \right\} \varepsilon \right] \\ &= \underbrace{\left\{ f(X) - \hat{f}(X) \right\}^2}_{\text{reducible error}} + \underbrace{\text{Var}(\varepsilon)}_{\text{irreducible error}}. \end{aligned}$$

- A more appropriate learning technique might reduce the *reducible error*.
- irreducible error* comes entirely from the inherent observation noise  $\varepsilon$ ; 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  $\varepsilon$ , since  $Y = f(X) + \varepsilon$ .

# Inference

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

- 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  $X_1, \dots, X_p$  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  $(X_1, \dots, X_p)$  maximizes  $Y$ ?
  - How much is  $Y$  affected by each predictor  $X_i$ ?

E.g., in Figure 1 we might have

  - \* 60% of  $Var(\text{sales})$  can be explained by TV budget,
  - \* 30% of  $Var(\text{sales})$  can be explained by Radio budget,
  - \* 8% of  $Var(\text{sales})$  can be explained by Newspaper budget,
  - \* remaining 2% can be explained by  $X_4, X_5, \dots, X_p$
  - These questions can be difficult to answer if  $f$  is highly non-linear!

## 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  $X_1, \dots, X_p$  are quantitative or qualitative is less important.

## Example of classification – $K$ -nearest neighbors

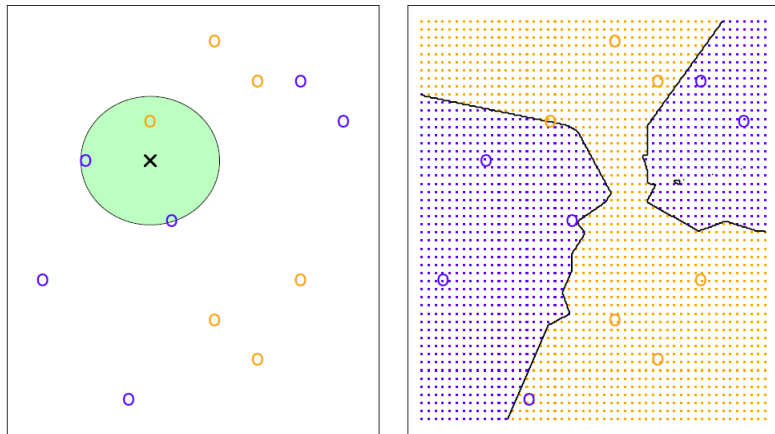


Figure 8: Image by James et al. (2021). Classification using the  $K$ -nearest neighbor approach with  $K = 3$ . **Left:** We assign a new observation (this is the "x") to the class for which most of three neighbors of "x" belong to. **Right:** A decision line/region for how we would assign a new element for  $K = 3$  if they would fall in a certain indicated region.

## Example of regression – $K$ -nearest neighbors

Same idea can be extended to regression:

$$\hat{f}_K(x_0) = \begin{cases} \textit{mode} \text{ of observed responses of } K \text{ nearest neighbors of } x_0, & \text{if classification,} \\ \textit{mean} \text{ of observed responses of } K \text{ nearest neighbors of } x_0, & \text{if regression.} \end{cases}$$

$K$ -NN is useful to have as a quick example that you can pull out at any time.

Motivation

Unsupervised learning

Supervised learning

Learning goals and tasks

Choosing a model

Assessing model accuracy

## Estimate $f$ : goal and notation

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

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

for scalar  $y_i$  and  $p$ -dimensional vector  $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^\top \in \mathbb{R}^p$ .

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

## Choosing a model

A statistical learning method will use the observed 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 of models

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

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

## Choosing a model

Why not allow  $\hat{f}$  to be any possible function?

- Advantage: A larger class of possible  $\hat{f} \Rightarrow$  more likely that  $\hat{f}$  better estimates  $f$ .
- There exists a function that is continuous nowhere, e.g., the **Dirichlet function**  $1_{\mathbb{Q}}(\cdot)$ .  
Therefore we must place some sort of smoothness condition on what  $\hat{f}$  is allowed to be.
- Okay...so maybe just allow  $\hat{f}$  to be any (piecewise) continuous function? (No.)

## Choosing a model: too much flexibility

**Interpolation theorem:** Any  $n$  bivariate data points  $(x_i, y_i)$  (where  $x_i$ s are distinct) can be interpolated by a polynomial of order at most  $n - 1$ .

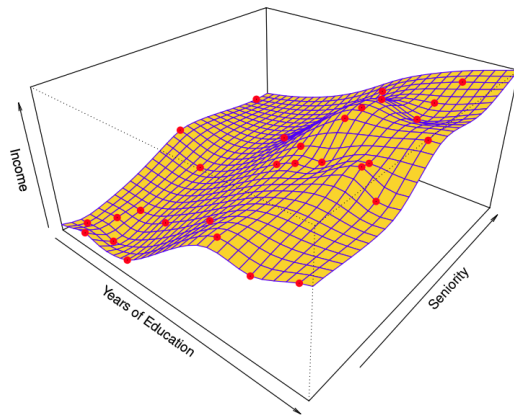
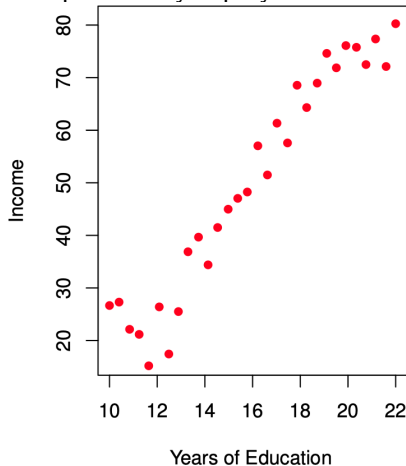


Figure 9: Image by James et al. (2021). Right: “A rough thin-plate spline fit to the Income data from Figure 2.3. *This fit makes zero errors on the training data.*”

## 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.

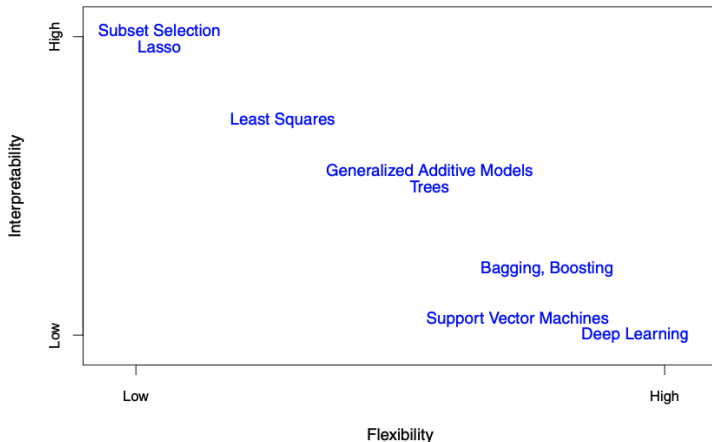


Figure 10: Image by James et al. (2021). No method dominates *all* other methods.

## Choosing a model: what do we mean by “flexibility”?

In the previous slide, a model’s “flexibility” was quantified by a single scalar number.

- Sometimes this is well-defined. For example:
  - {linear model with no interaction}  $\subseteq$  {linear model with two-way interactions}
  
  - {continuously differentiable function}  $\subseteq$  {continuous function}
  
- But we do not always have nested model classes, in which case it is more difficult to quantify model flexibility.

## Choosing a model: final comments

Language: “make an *assumption* about the functional form of  $f$ ”

- Does not necessarily mean that we believe it to be true.
- Start simple to fully (and more quickly) understand something. Then can generalize.

Motivation

Unsupervised learning

Supervised learning

Learning goals and tasks

Choosing a model

Assessing model accuracy

## How do we assess model accuracy?

Naïve workflow:

1. Compute/learn/fit/train various estimators  $\hat{f}_1, \hat{f}_2, \hat{f}_3, \dots$  using all available data.
2. Choose the fitted estimator that produces the smallest error on this data.

Issue: an estimator  $\hat{f}$  can achieve *zero error* on this data by overfitting to the data, but such an estimator will not generalize well to unseen data.

Better workflow:

1. Split the available data into two sets: *training data* and *test data*.
2. Compute/learn/fit/train various estimators  $\hat{f}_1, \hat{f}_2, \hat{f}_3, \dots$  using the training data.
3. Assess the generalization ability of each estimator using the test data.

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

## Regression: idea

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

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

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

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

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

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

## Regression: a sketch

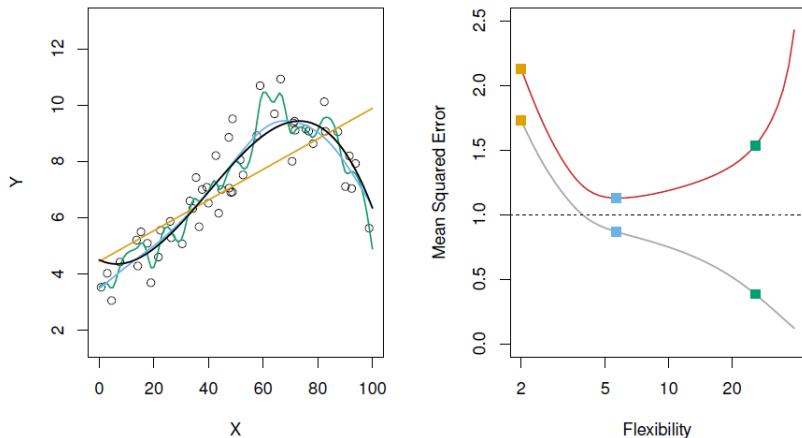


Figure 11: Image by James et al. (2021). **Left:** Data (black circles) simulated from  $f$  (black curve) with noise. Three estimates for  $f$ : linear regression (orange line), and two smoothing splines (blue and green curves). **Right:** Training MSE (in grey), test MSE (in red), and minimum possible 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 estimator yields a small training MSE but a large test MSE, we say that the estimator is *overfitting* the training data.

## Regression: another sketch

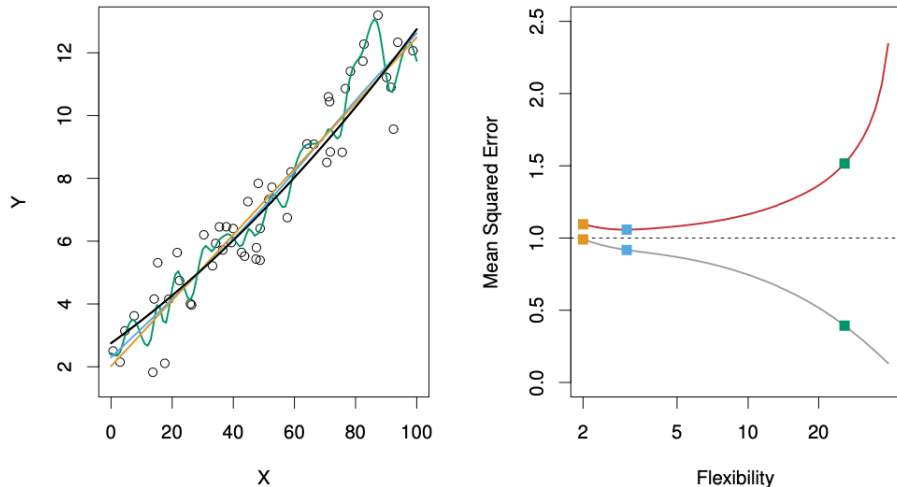


Figure 12: Image by James et al. (2021). Details are same as in previous figure, but using a different true  $f$  that is much closer to linear. In this setting, linear regression provides a very good fit to the data. U-shape in the test-MSE curve is barely noticeable.

## 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:

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

- $\text{Var} \left( \hat{f}(x) \right)$  is the amount  $\hat{f}(x)$  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  $f$ ).
- $\text{Var}(\varepsilon)$  is the variance of the error term  $\varepsilon$  (this is a property of the data, not of  $\hat{f}$ ).
  - \* The expected test MSE can never be below  $\text{Var}(\varepsilon)$ . (Dotted horizontal line in previous 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.

## Another exposition on bias-variance trade-off

- StatQuest: Bias and Variance (6:35)
- qmd file on Canvas

## Classification: training and test error rate

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

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

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

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

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

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

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

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

where  $(x_{n+1}, y_{n+1}), \dots, (x_{n+m}, y_{n+m})$ , 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

$$\arg \max_j P(Y = j|X = x). \quad (8)$$

where  $P(Y = j|X = x)$  is the conditional probability that a point whose predictor value is  $x$  belongs to class  $j$ .

### Predicting the penguin species given some measurements

We want to predict the species (Adelie, Chinstrap, or Gentoo) of a *new* penguin with

$$x^* = (\text{island}=\text{Biscoe}, \text{bill\_len} = 38.7, \text{bill\_dep} = 19.2, \text{body\_mass} = 3800, \text{sex}=\text{male}).$$

Because a penguin must belong to one of these three species, we have that

$$1 = P(Y = \text{Adelie}|X = x^*) + P(Y = \text{Chinstrap}|X = x^*) + P(Y = \text{Gentoo}|X = x^*).$$

The Bayes classifier will then predict that a penguin with measurements  $x^*$  belongs to the species  $j \in \{\text{Adelie}, \text{Chinstrap}, \text{Gentoo}\}$  that produces the largest value of  $P(Y = j|X = x^*)$ .

## Classification: Bayes classifier as a (theoretical) benchmark

The Bayes 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.
- Cannot compute the Bayes classifier if we don't know the conditional distribution of  $Y|X$ .
- Approximate the Bayes classifier by estimating the conditional probabilities  $P(Y=j|X=x)$ .

### Predicting the penguin species given some measurements

For the previous penguin, if we estimate that

$$P(\text{Adelie}|X = x^*) \approx 0.2; \quad P(\text{Chinstrap}|X = x^*) \approx 0.45; \quad P(\text{Gentoo}|X = x^*) \approx 0.35,$$

then an approximate Bayes classifier will predict that this penguin belongs to Chinstrap.

## Classification: KNN – i) Idea

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

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

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

- Eq. (9) is the proportion of  $x$ 's neighbors in the training data that belong to class  $j$ .
- Substituting Eq. (9) into Eq. (8) produces a classifier that assigns  $x$  to the *majority* class in  $\mathcal{N}_K(x)$ ; this classifier is exactly the KNN classifier discussed earlier.

## Classification: KNN – ii) Different $K$

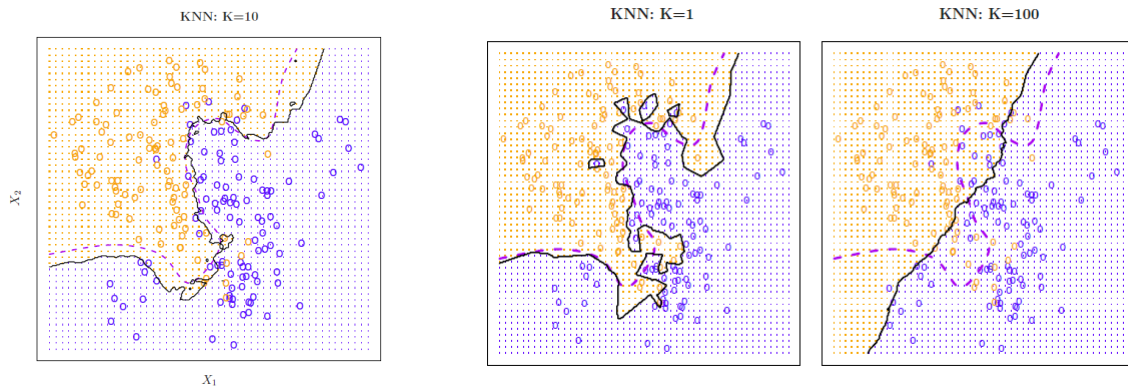


Figure 13: Image by James et al. (2021). KNN decision boundaries (black solid curve) for  $K = 10$ ,  $K = 1$ , and  $K = 100$ , and the Bayes decision boundary (purple dashed curve). Of the three decision boundaries, the  $K = 10$  boundary most closely matches the Bayes decision boundary. The  $K = 1$  decision boundary is too flexible, while the  $K = 100$  boundary is not flexible enough.

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

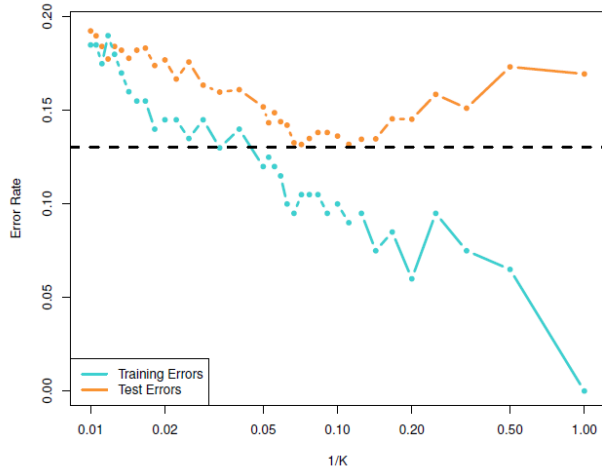


Figure 14: Image by James et al. (2021). The KNN training error rate (blue) and test error rate (orange) and the Bayes error rate (black horizontal dashed line). The jumpiness of the curves is due to the small size of the training data set.

## General comments

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

- True regardless of regression or classification.