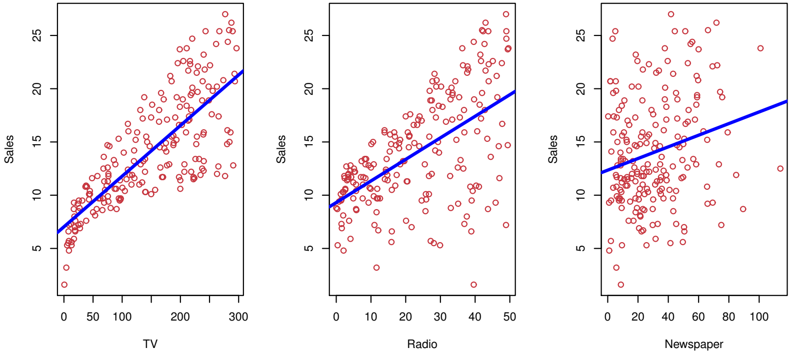
Basic Concepts

Damek Davis
School of ORIE, Cornell University
ORIE 4740 Lec 1.5 (Jan 25)

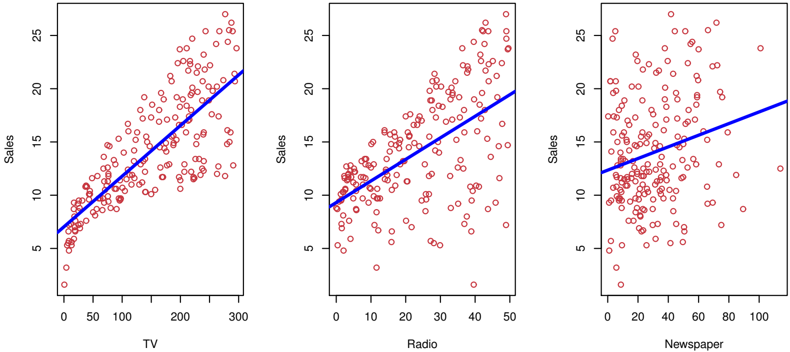
Reference: Chapter 2 of ISLR

Predicting Sales



- **Goal:** predict sales as a function of budget on TV, Radio, and Newspaper.

Predicting Sales



- **Goal:** predict sales as a function of budget on TV, Radio, and Newspaper.
- Independent predictions ignore relations between budgets, so may suggest using a more flexible model.

$$
\begin{aligned}
Y &= f(X) + \epsilon \\
&= \underbrace{f(X_1, X_2, X_3)}_{\text{model}} + \underbrace{\epsilon}_{\text{error}}
\end{aligned}
$$

Supervised or Unsupervised?

The sales prediction task is

Choose one:

A. Supervised learning

B. Unsupervised learning

Supervised Learning

- Learn a rule for predicting the value of a response variable based on the value of some set of predictor variables.

- Have a set of training dataset in which the predictors and outcome values are known for each data points.

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

Response or Predictor?

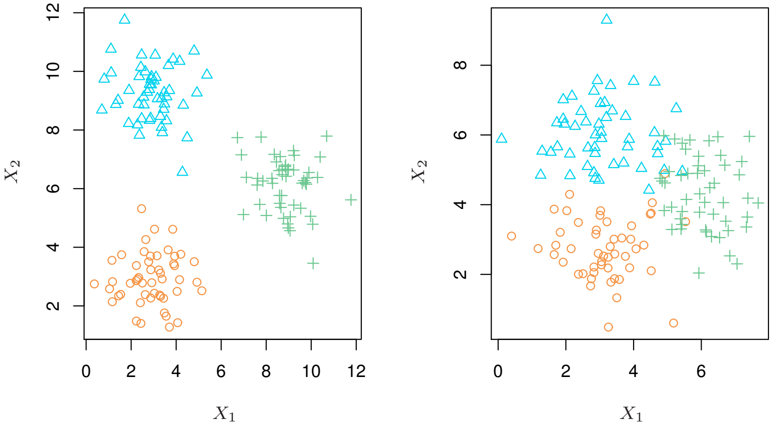
In the sales prediction task, the radio budget is a

Choose one:

A. Response variable

B. Predictor variable

Unsupervised learning



- **Goal:** Put data into “similar" groups, find a “good" or “compressed" “representation of data...."

Unsupervised Learning

- The response variable is unknown for the training data

Unsupervised Learning

- The response variable is unknown for the training data
- E.g.: Clustering. Categorizing customers into groups with similar purchasing habits

Unsupervised Learning

- The response variable is unknown for the training data
- E.g.: Clustering. Categorizing customers into groups with similar purchasing habits
  - The training data is the purchasing data of a set of customers

Unsupervised Learning

- The response variable is unknown for the training data
- E.g.: Clustering. Categorizing customers into groups with similar purchasing habits
  - The training data is the purchasing data of a set of customers
  - The groups are not pre-defined.

Unsupervised Learning

- The response variable is unknown for the training data
- E.g.: Clustering. Categorizing customers into groups with similar purchasing habits
  - The training data is the purchasing data of a set of customers
  - The groups are not pre-defined.
- Challenge: Unclear how “well" your algorithm works!

Unsupervised Learning

- The response variable is unknown for the training data
- E.g.: Clustering. Categorizing customers into groups with similar purchasing habits
  - The training data is the purchasing data of a set of customers
  - The groups are not pre-defined.
- Challenge: Unclear how “well" your algorithm works!
- Sometimes used as preprocessing step **before** supervised learning.

Unsupervised Learning

- The response variable is unknown for the training data
- E.g.: Clustering. Categorizing customers into groups with similar purchasing habits
  - The training data is the purchasing data of a set of customers
  - The groups are not pre-defined.
- Challenge: Unclear how “well" your algorithm works!
- Sometimes used as preprocessing step **before** supervised learning.
  - “Labeling data" is costly (human intervention)

Unsupervised Learning

- The response variable is unknown for the training data
- E.g.: Clustering. Categorizing customers into groups with similar purchasing habits
  - The training data is the purchasing data of a set of customers
  - The groups are not pre-defined.
- Challenge: Unclear how “well" your algorithm works!
- Sometimes used as preprocessing step **before** supervised learning.
  - “Labeling data" is costly (human intervention)
  - Find small set of “representative data" samples, try to find labels for those samples.

Supervised Learning

- **This class:** mostly about supervised learning.

Supervised Learning

- **This class:** mostly about supervised learning.
- We have a training dataset: $(x_i, y_i), i = 1, 2, \ldots, n$
- We assume data follows relationship

$$\boxed{Y = f(X) + \epsilon}$$

  - The model is never perfect, so expect nonzero error.

Supervised Learning

- **This class:** mostly about supervised learning.
- We have a training dataset: $(x_i, y_i), i = 1, 2, \ldots, n$
- We assume data follows relationship

$$\boxed{Y = f(X) + \epsilon}$$

  - The model is never perfect, so expect nonzero error.
- Core Question:

How to estimate f?

  - Many tradeoffs to consider.

Tradeoffs

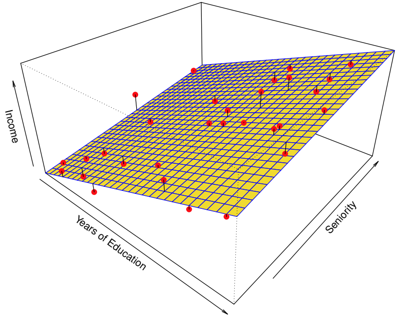
- linear vs. nonlinear methods
- regression vs. classification
- evaluation: MSE vs. classification error
- evaluation: training error vs. test error
- model selection: flexibility vs. interpretability

Linear vs. Nonlinear

How does response depend on predictors?

Linear vs. Nonlinear

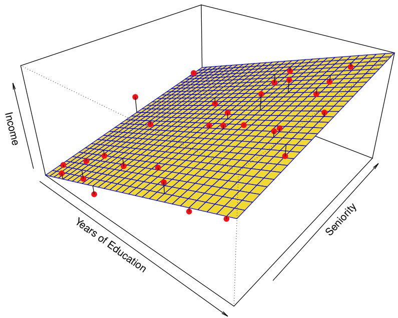
How does response depend on predictors?

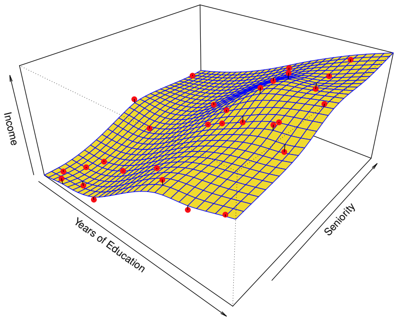


$$Y \approx \underbrace{\hat{f}(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2}_{\text{a linear function of } X}$$

Linear vs. Nonlinear

How does response depend on predictors?



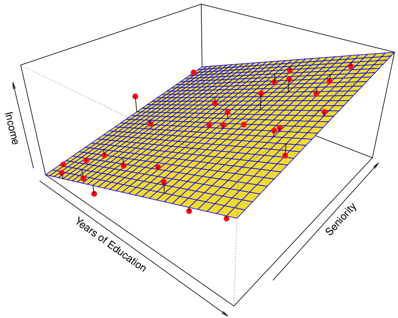


$$Y \approx \underbrace{\hat{f}(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2}_{\text{a linear function of } X}$$

$$Y \approx \underbrace{\hat{f}(X)}_{\text{a nonlinear function of } X}$$

Regression vs. Classification

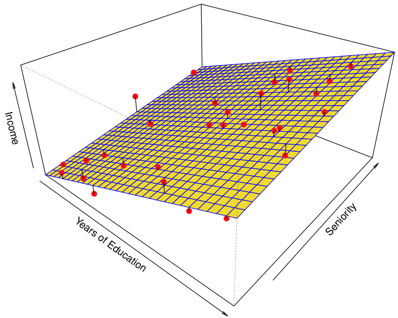
Is response variable continuous or discrete?



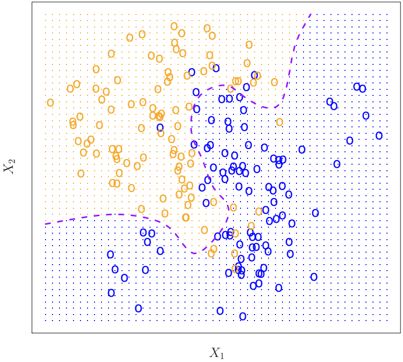
$Y \in \mathbb{R}$

Regression vs. Classification

Is response variable continuous or discrete?



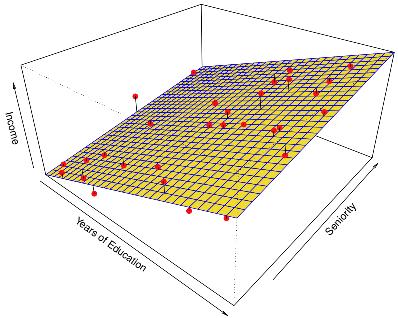
$Y \in \mathbb{R}$



$Y \in \{-1, +1\}$

Evaluation: MSE vs. Classification Error

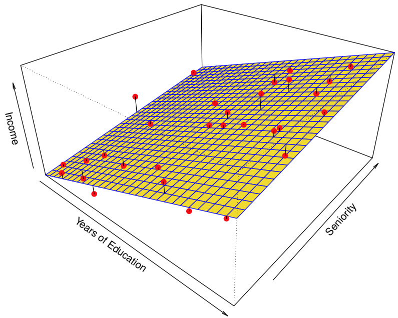
Do we measure error continuously or discretely?

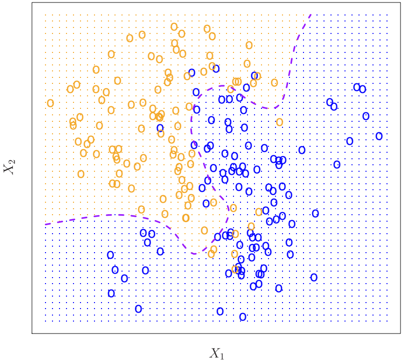


$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Evaluation: MSE vs. Classification Error

Do we measure error continuously or discretely?





$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i))$$

Evaluation: Training error vs Test Error

Do we care about training error or test error?

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \qquad\qquad \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i))$$

- Computed on the training dataset $(x_i, y_i), i = 1, 2, \ldots, n$.

Evaluation: Training error vs Test Error

Do we care about training error or test error?

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \qquad\qquad \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i))$$

- Computed on the training dataset $(x_i, y_i), i = 1, 2, \ldots, n$.
- Important: errors on the test dataset $(\tilde{x}_i, \tilde{y}_i), i = 1, 2, \ldots, m$.
    - previously unseen data
    - not used to build $\hat{f}$.

Evaluation: Training error vs Test Error

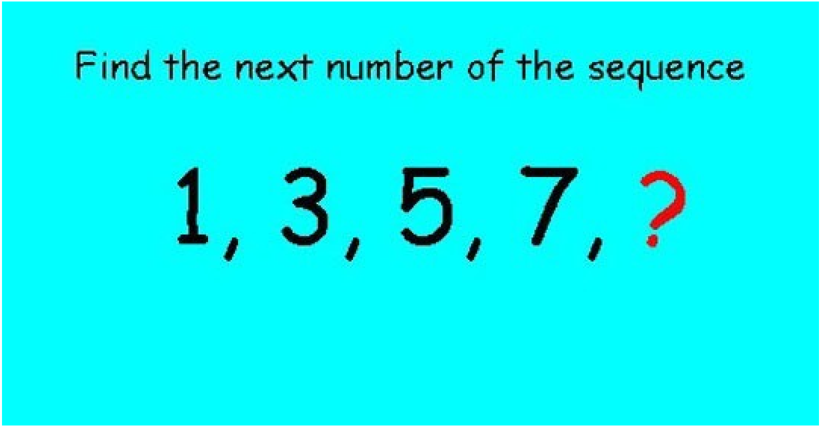
Do we care about training error or test error?

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \qquad\qquad \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i))$$

- Computed on the training dataset $(x_i, y_i), i = 1, 2, \ldots, n$.
- Important: errors on the test dataset $(\tilde{x}_i, \tilde{y}_i), i = 1, 2, \ldots, m$.
  - previously unseen data
  - not used to build $\hat{f}$.
  - $\Longrightarrow$ measure testing MSE or testing Classification error.

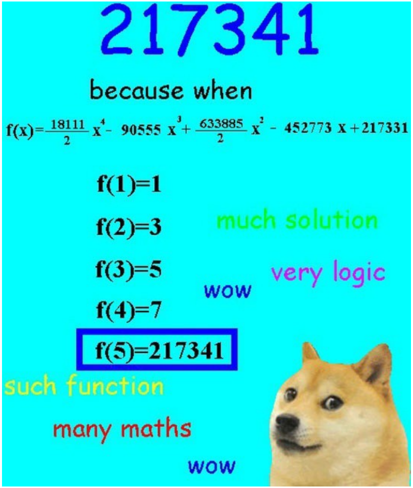
$$\boxed{\text{MSE} = \frac{1}{m}\sum_{i=1}^{n}(\tilde{y}_i - \hat{f}(\tilde{x}_i))^2 \qquad\qquad \frac{1}{n}\sum_{i=1}^{n} I(\tilde{y}_i \neq \hat{f}(\tilde{x}_i))}$$

Training vs Testing Error



- training data = $\{(1, 1), (2, 3), (3, 5), (5, 7)\}$,
- $\hat{f}(i) = i$th odd number. Perfect training MSE

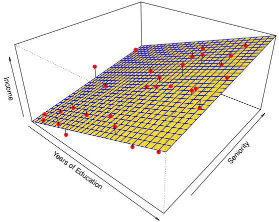
Training vs Testing Error



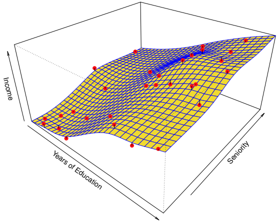
- testing data = $\{(5, 27341)\}$
- Testing MSE = $(9 - 27341)^2 = 747,038,224$

Model Selection: Flexibility vs. Interpretability

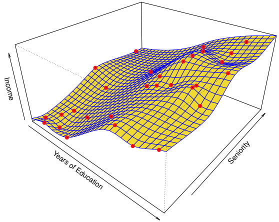
Which model will perform best on test data?



underfit



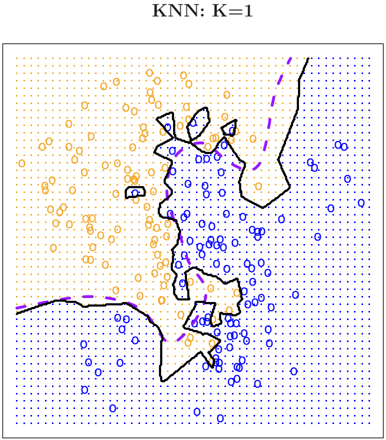
good fit

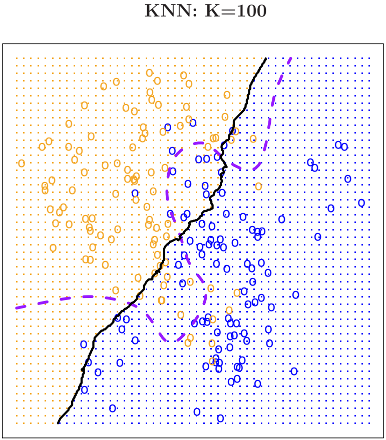


overfit

Model Selection: Flexibility vs. Interpretability

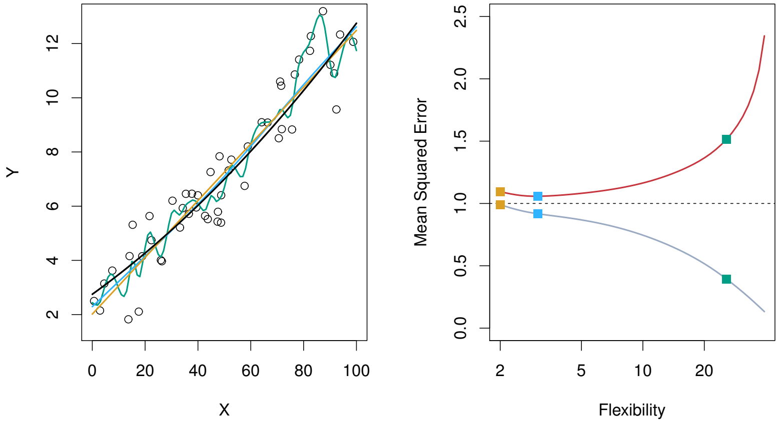
Which model will perform best on test data?



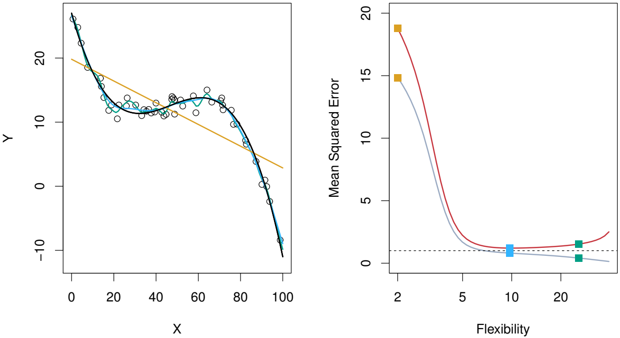


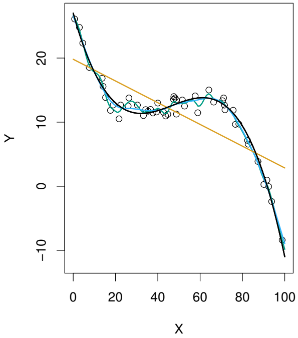
Model Selection: Flexibility vs. Interpretability

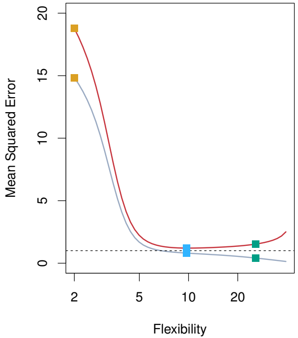
Which model will perform best on test data?

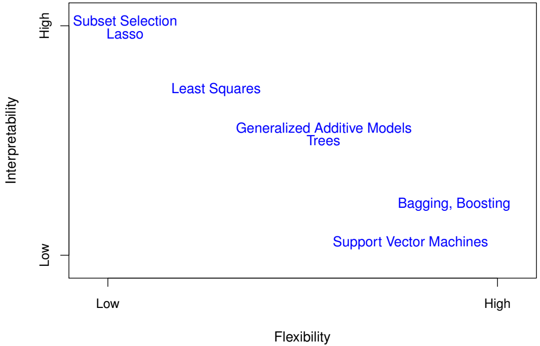


(training data, test error)


Y
X
0
20
40
60
80
100
−10
0
10
20
Mean Squared Error
Flexibility
2
5
10
20
0
5
10
15
20


Y
X
0
20
40
60
80
100
−10
0
10
20


Mean Squared Error
Flexibility
2
5
10
20
0
5
10
15
20

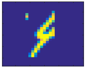

Interpretability
High
Low
Flexibility
Subset Selection
Lasso
Least Squares
Generalized Additive Models
Trees
Bagging, Boosting
Support Vector Machines

Statistical Learning vs Machine Learning

- **A lot of overlap:** Supervised/Unsupervised Learning.
- **Differences:**
  - **ML:** massive data, prediction-focused, algorithm-centric
  - **SL:** big/small data, holistic view on statistical aspects of model

Regression or Classification?

You want to build a model that determines whether the following images are fours or eights:



This is a Choose one:

A. Regression Task

B. Classification Task

What to try next?

You have a training set with one predictor variable and one response variable. You fit a model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

and get perfect training error. On the the other hand, you are astonished to find out that your model performs really poorly on the test data. Which model should you try next?

Choose one:

A. $Y = \beta_0 + \beta_1 X + \beta_2 X^2$

B. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4$

Some of the figures in this presentation are taken from “An Introduction to Statistical Learning, with applications in R” (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani
The digit images were taken from the MNist dataset.
Slides based on Yudong Chen’s slides.
Some images due to Machine learning @ Berkeley Group