Ch 2.1: What is Statistical Learning?

Lecture 2 - CMSE 381

Prof. Elizabeth Munch

Michigan State University
::
Dept of Computational Mathematics, Science & Engineering

Weds, Aug 30, 2023

Announcements

Last time:

- Discussed where to find everything
  - Github
  - Slack
  - D2L
- Check out the syllabus!

Lec #	Date			Reading
1	Mon	Aug 28	Intro / First day stuff / Python Review Pt 1	1
2	Wed	Aug 30	What is statistical learning?	2.1
	Fri	Sep 1	Assessing Model Accuracy	2.2.1, 2.2.2
3	Mon	Sep 4	No class - Labor day	
4	Wed	Sep 6	Linear Regression	3.1
5	Fri	Sep 8	More Linear Regression	3.1/3.2

Announcements:

- Get on slack!
  - +1 point on the first homework if you post a gif in the thread
- First homework due Fri Sep 1
- First office hours next week

Covered in this class

- Input/output variables
- Prediction vs inference
- Reduceable vs irreduceable error
- Overfitting
- Classification vs regression
- Supervised vs Unsupervised learning

- Please note: no jupyter notebook for today's class, slides only

An example data set: `Advertising`

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
6	8.7	48.9	75	7.2
7	57.5	32.8	23.5	11.8
8	120.2	19.6	11.6	13.2
9	8.6	2.1	1	4.8
10	199.8	2.6	21.2	10.6
11	66.1	5.8	24.2	8.6

- Sales of a product in 200 markets, along with amount spent on three differnt types of advertising
- Goal:
- Input variables:
- Output variable:

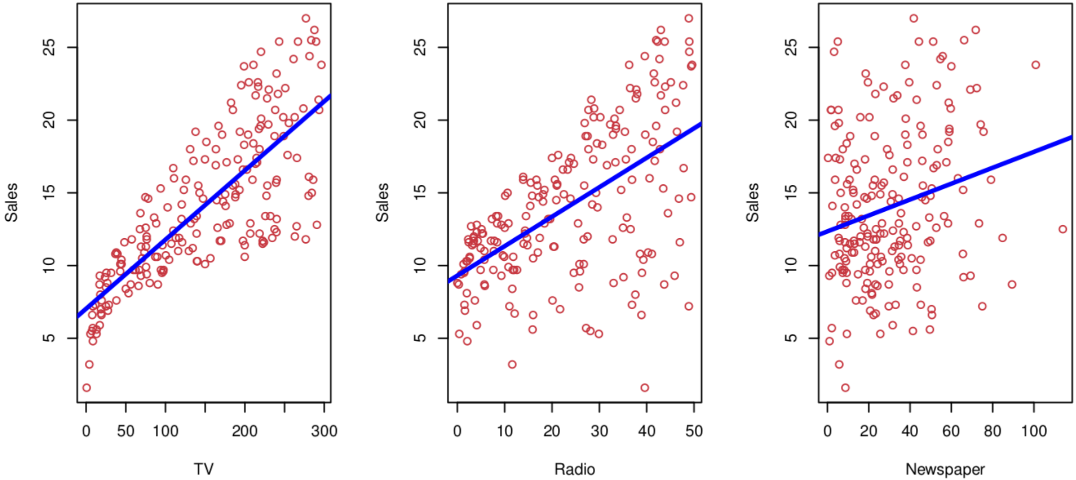
Data available at
https://github.com/nguyen-toan/ISLR/blob/master/dataset/Advertising.csv

Notation and Big Assumption

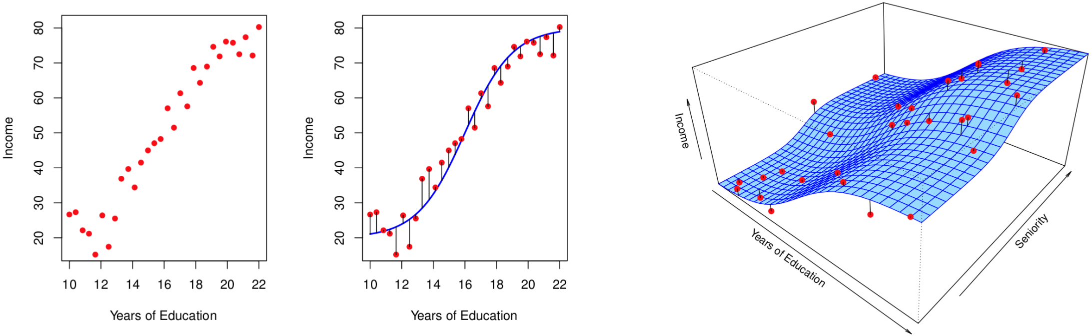
Input variables: $X_1, X_2, \cdots, X_p$
Output variable: Y

$$Y = f(X) + \varepsilon$$

Advertising Example



More examples



Section 1

Prediction vs Inference

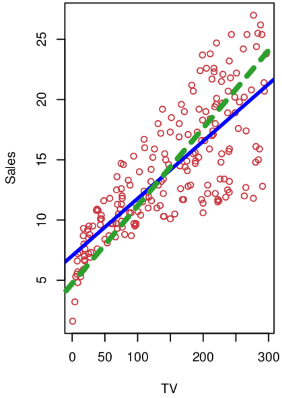
Prediction

Given a value X, try to provide an estimate for $f(X)$.

Build a model:

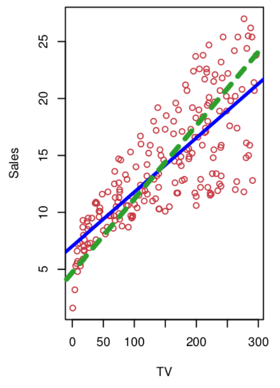
$$\hat{Y} = \hat{f}(X)$$

Example: If we spend \$150 on TV advertising, what do we predict we will we make in sales?



Group question:

1		TV	Radio	Newspaper	Sales
2	1	230.1	37.8	69.2	22.1
3	2	44.5	39.3	45.1	10.4
4	3	17.2	45.9	69.3	9.3



The blue solid line is f. The green dashed line is $\hat{f}$.

- What is the predicted sales for the first three data points using the green dashed line $\hat{f}$ shown in the graph?
- Using the dashed green line as the predicted model $\hat{f}$, what is the error in each of the three predicions?

Reduceable vs irreducable error

All models are wrong, some are useful.

$$Y - \hat{Y}$$

Reducible Error

Irreducible Error

More on error

- Given estimate $\hat{f}$ (fixed)
- Set of predictors X (fixed)
- Prediction $\hat{Y} = \hat{f}(X)$

$$E(Y - \hat{Y})^2 =$$

Inference

Want f, but not for prediction (or possibly combined with prediction)

- Which predictors are associated with the response?
- What is the relationship between the response and each predictor?
- Can the relationship between Y and each predictor be adequately summarized using a linear equation? Is it more complicated?

Determine whether each scenario is prediction, inference, or both.

Application	Prediction	Inference
Predict effectiveness of vaccine		
Determine the address written on the image of an envelope.		
Identify risk factors for getting long covid.		
Transcribe an audio file of a person talking.		
Predict stock prices.		

Section 2

How to estimate f?

Input: Training data

- n data points observed
- x_{ij} is the jth predictor for observation i
- y_i is the response variable for the ith observation
- Training data:
  - $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$
  - $x_i = (x_i, x_{i2}, \cdots, x_{ip})^T$

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
6	8.7	48.9	75	7.2
7	57.5	32.8	23.5	11.8
8	120.2	19.6	11.6	13.2
9	8.6	2.1	1	4.8
10	199.8	2.6	21.2	10.6
11	66.1	5.8	24.2	8.6

Step 1: Select a model

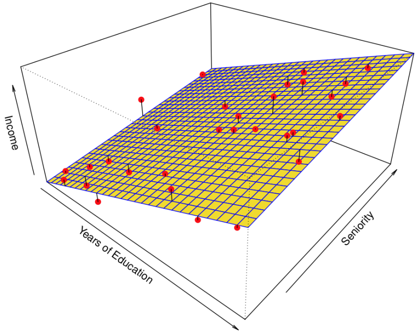
Example:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Step 2: Train the model

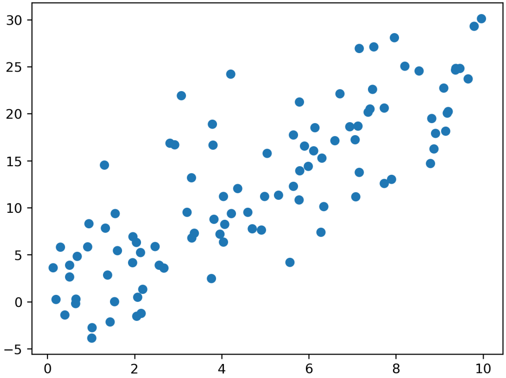
Example:
Find $\beta_i' s$ so that

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$



How do you decide on the coefficients?

$$Y \approx \beta_0 + \beta_1 X_1$$

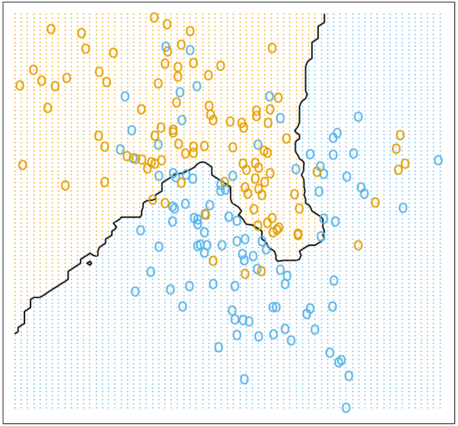


Desmos toy: https://www.desmos.com/calculator/skvt8c7317

Example Non-parametric method: Nearest Neighbors

$N_k(x) =$ Set of k nearest neighbors of x

$$\hat{f}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$



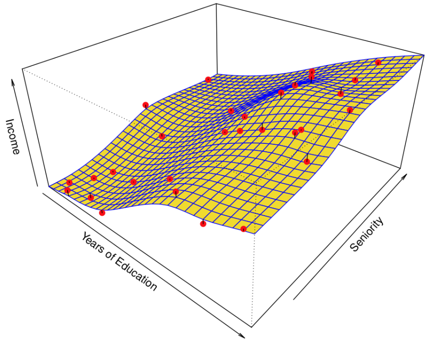
$k = 15$

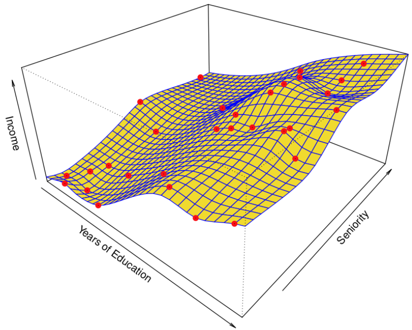
Parametric methods: Pros and Cons

Pros

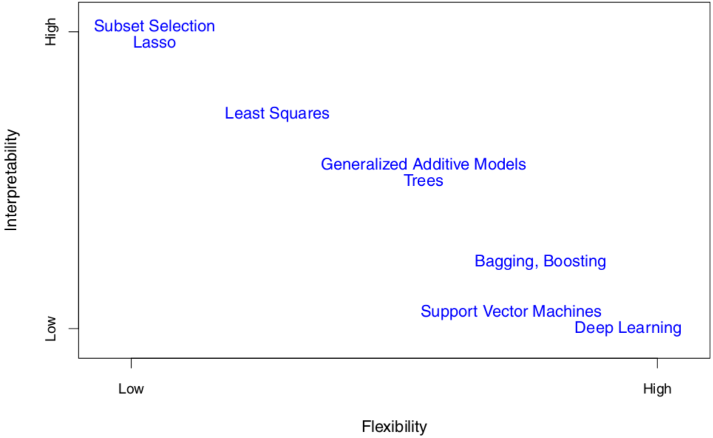
Cons

Overfitting

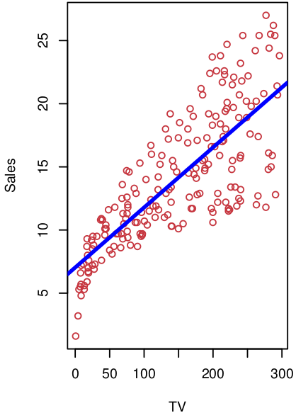




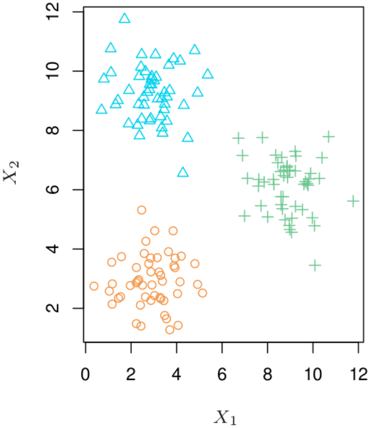
Prediction Accuracy vs Model Interpretability



Supervised learning:
Training data has response variable y for every input x



Unsupervised Learning:
Training data has response variable y for every input x



Regression vs Classification

Types of variables:

- Quantitative

- Qualitative / Categorical

Section 3

Group work

(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

- Is this classification or regression?
- Do we want inference or prediction?
- What is n, the number of data points?
- What is p, the number of variables?

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

- Is this classification or regression?
- Do we want inference or prediction?
- What is n, the number of data points?
- What is p, the number of variables?

TL;DR

Wrap up

Next time:

- Friday:
  - Bring Laptop!
  - First homework due Fri Sep 1
  - There will be a quiz this week
- Monday:
  - No class: Labor day!

Lec #	Date			Reading
1	Mon	Aug 28	Intro / First day stuff / Python Review Pt 1	1
2	Wed	Aug 30	What is statistical learning?	2.1
	Fri	Sep 1	Assessing Model Accuracy	2.2.1, 2.2.2
3	Mon	Sep 4	No class - Labor day	
4	Wed	Sep 6	Linear Regression	3.1
5	Fri	Sep 8	More Linear Regression	3.1/3.2
6	Mon	Sep 11	Even more linear regression	3.2.2
7	Wed	Sep 13	Probably more linear regression	3.3
8	Fri	Sep 15	Intro to classification, Logisitic Regression	2.2.3, 4.1, 4.2, 4.3

Announcements:

- Get on slack!
  - +1 point on the first homework if you post a gif in the thread
- Office hours!
  - Dr. Munch: Weds and Thurs 11-12
  - Rachel Roca: Tues 3:30 - 5pm and Fri 11:30 - 1pm