

Ch 2.1: What is Statistical Learning?

Lecture 2 - CMSE 381

Prof. Lianzhang Bao

Michigan State University

::

Dept of Computational Mathematics, Science & Engineering

Weds, Jan 15 , 2025

Announcements

Last time:

- Discussed where to find everything
 - ▶ Course webpage
 - ▶ Slack
 - ▶ D2L
- Check out the syllabus!

Lec #		Date	Topic	Reading	HW	Pop Quizzes	Notes
1	M	1/13	Intro / Python Review	1			
2	W	1/15	What is statistical learning	2.1		Q1	
3	F	1/17	Assessing Model Accuracy	2.2.1, 2.2.2			
	M	1/20	MLK - No Class				
4	W	1/22	Linear Regression	3.1		Q2	
5	F	1/24	More Linear Regression	3.1	HW #1 Due Sun 1/26		
6	M	1/27	Multi-linear Regression	3.2			
7	W	1/29	Probably More Linear Regression	3.3		Q3	

Announcements:

- Get on slack!
 - ▶ +1 point on the first homework if you post a gif in the thread
- First homework due Sun Jan 26
- First office hours next week

Covered in this class

- Input/output variables
 - Prediction vs inference
 - Reduceable vs irreducible 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

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
5	4	151.5	41.3	58.5	18.5
6	5	180.8	10.8	58.4	12.9
7	6	8.7	48.9	75	7.2
8	7	57.5	32.8	23.5	11.8
9	8	120.2	19.6	11.6	13.2
10	9	8.6	2.1	1	4.8
11	10	199.8	2.6	21.2	10.6
12	11	66.1	5.8	24.2	8.6

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

Data available at msu-cmse-courses.github.io/CMSE381-S25/DataSets/DataSets.html

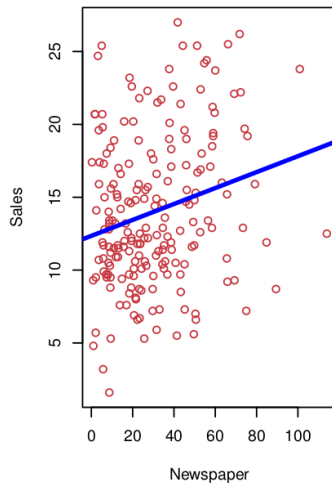
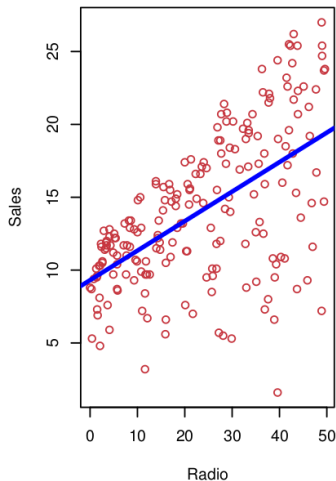
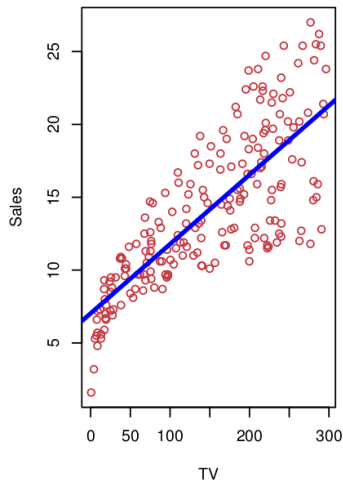
Notation and Big Assumption

Input variables: X_1, X_2, \dots, X_p

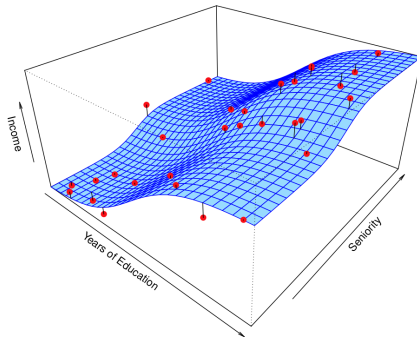
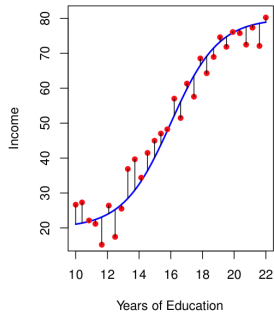
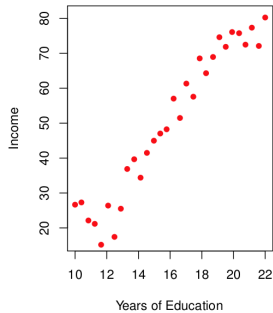
Output variable: Y

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

Advertising Example



More examples



Section 1

Prediction vs Inference

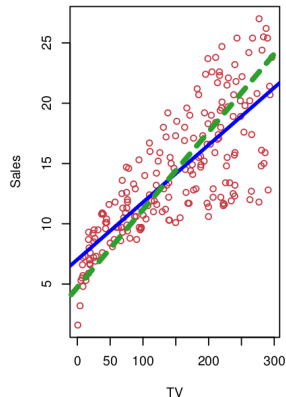
Prediction

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

Build a model:

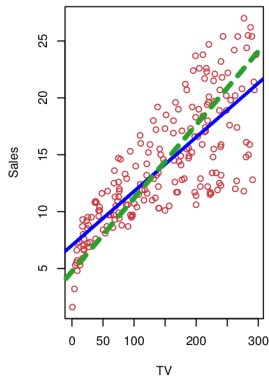
$$\hat{Y} = \hat{f}(X)$$

Example: If we spend \$250 on TV advertising, what do we predict we will we make in sales?



Group question:

		TV	Radio	Newspaper	Sales
1					
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?
roughly 17, 7, 6
- Using the dashed green line as the predicted model \hat{f} , what is the error in each of the three predictions?
roughly 5, 3, 3

Reduceable vs irreducible error

All models are wrong, some are useful.

$$Y - \hat{Y}$$

Reducible Error

Irreducible Error

\hat{f}

will not be a perfect estimate for f

$Y = f(x) + \epsilon$

true <--

$$f = \alpha + \beta x$$

estimate <--

$$\hat{f} = \hat{\alpha} + \hat{\beta} x$$

More on error

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

$$E(Y - \hat{Y})^2 = E[f(x) + \varepsilon - \hat{f}(x)]^2$$
$$[f(x) - \hat{f}(x)]^2 + 0 + \text{Var}(\varepsilon)$$

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	p	
Determine the address written on the image of an envelope.	i	
Identify risk factors for getting long covid.	i	
Transcribe an audio file of a person talking.	i	
Predict stock prices.	p	

Section 2

How to estimate f ?

Input: Training data

- n data points observed
- x_{ij} is the j th predictor for observation i
- y_i is the response variable for the i th observation
- Training data:
 - ▶ $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$
 - ▶ $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$

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
5	4	151.5	41.3	58.5	18.5
6	5	180.8	10.8	58.4	12.9
7	6	8.7	48.9	75	7.2
8	7	57.5	32.8	23.5	11.8
9	8	120.2	19.6	11.6	13.2
10	9	8.6	2.1	1	4.8
11	10	199.8	2.6	21.2	10.6
12	11	66.1	5.8	24.2	8.6

Parametric methods

Step 1: Select a model

Example:

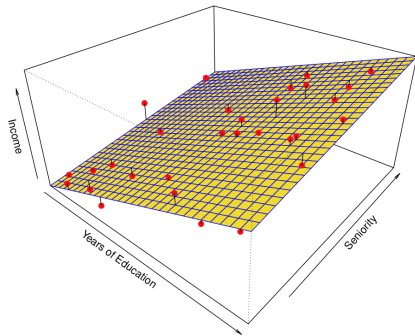
$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Step 2: Train the model

Example:

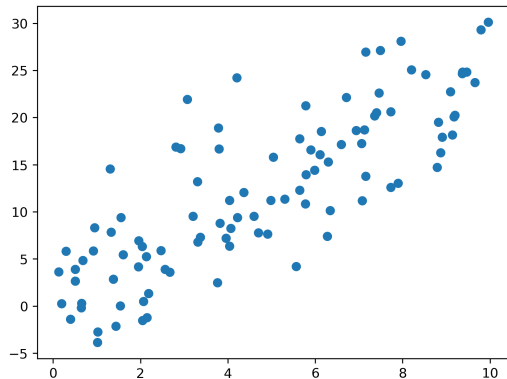
Find β_i 's so that

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



How do you decide on the coefficients?

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

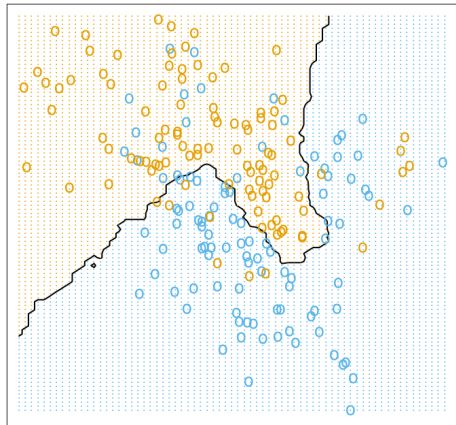


Desmos toy: <https://www.desmos.com/calculator/skvt8c73l7>

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

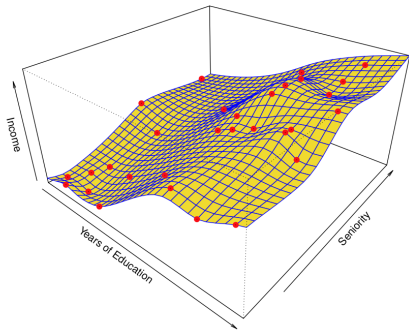
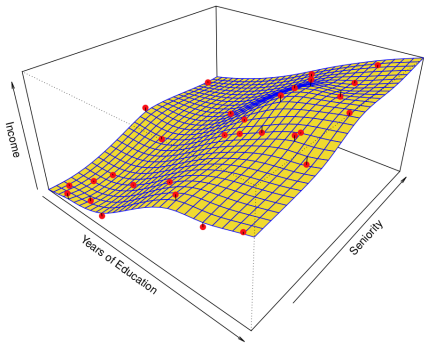
overfitting

Cons

possible to use wrong model (?)

underfitting

Overfitting

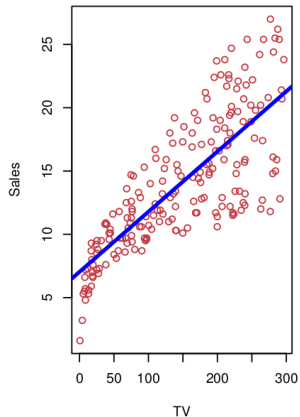


Prediction Accuracy vs Model Interpretability



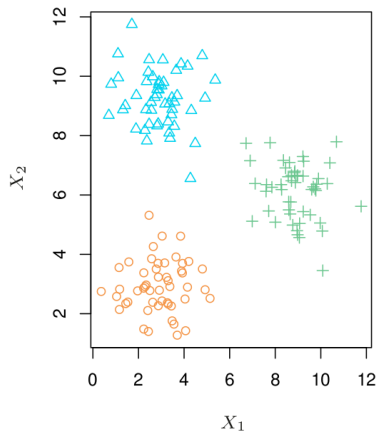
Supervised learning:

Training data has response variable y for every input x



Unsupervised Learning:

Training data does not have 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.

- r • Is this classification or regression?
- i • Do we want inference or prediction?
- 500 • What is n , the number of data points?
- 4 • 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.

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

Wrap up

Next time:

- Friday 1/17
 - ▶ Bring Laptop!
 - ▶ First homework due Sun Jan 26
 - ▶ There will be a quiz this Friday

Lec #	Date	Topic	Reading	HW	Pop Quizzes	Notes
1	M 1/13	Intro / Python Review	1			
2	W 1/15	What is statistical learning	2.1		Q1	
3	F 1/17	Assessing Model Accuracy	2.2.1, 2.2.2			
	M 1/20	MLK - No Class				
4	W 1/22	Linear Regression	3.1		Q2	
5	F 1/24	More Linear Regression	3.1	HW #1 Due Sun 1/26		
6	M 1/27	Multi-linear Regression	3.2			
7	W 1/29	Probably More Linear Regression	3.3		Q3	
8	F 1/31	Last of the Linear Regression				

Announcements:

- Get on slack!
 - ▶ +1 point on the first homework if you post a gif in the thread
- Office hours!
 - ▶ Dr. Bao: Tu-W 9am-10am
 - ▶ Christy Lu: TBD