Supervised Learning

Today we are learning about concepts that will be used throughout this course:

▶ Types of variables
▶ Estimating a function
▶ Prediction and inference
▶ Parametric and nonparametric estimation
▶ Interpretability
▶ More R!
What is supervised learning? Let’s start with an example.

- Let’s look at the Advertising data set (not in the course data package!!!)
- In the data:
  - sales for each of 200 products
  - advertising budgets for each product in TV, radio and newspaper

If you have a new product, how should you spend your advertising money to generate the most sales? How much? In which media?

How can we use data to answer these questions?
If you have a new product, how should you spend your advertising money to generate the most sales?

Idea: predict the level of sales from TV, radio and newspaper

Input variables (covariates, independent variables, predictors, features):

- TV
- radio
- newspaper

Output variable (response, dependent variable):

- sales
Supervised Learning: Advertising

If you have a new product, how should you spend your advertising money to generate the most sales?

[Graphs showing scatter plots with lines of best fit for TV, Radio, and Newspaper advertising against sales.]
Supervised Learning: Advertising

Formula:

\[ \text{sales} = f(\text{TV}, \text{radio}, \text{newspaper}) + \text{noise} \]

We want to find \( f \)!

In general,

\[ X = (X_1, X_2, \ldots, X_p)^T \]
\[ Y = f(X) + \epsilon \]
Supervised Learning: Regression and Classification

\[ X = (X_1, X_2, \ldots, X_p)^T \quad \text{inputs} \]
\[ Y \quad \text{output} \]
\[ Y = f(X) + \epsilon \quad \text{relationship} \]

We are interested in studying \( f \) in two settings: regression and classification.

**Regression:** \( Y \) has continuous values, like $81,200 or 72.

**Classification:** \( Y \) has categorical values, like low/medium/high or red/green/blue
Supervised Learning: Income

Another example: we want to predict the annual income (income) of an individual based on years of education (years of education).
Supervised Learning: Income

Let’s include seniority (seniority) as well.
Why estimate \( f \)?

- prediction
- inference

Prediction:

- We have a new product with a set advertising budget TV, radio and newspaper). What will its sales be?
- Alice has 16 years of education and 0 years of seniority. What will her income be?

Goal: accurately estimate output for new inputs.
In general, prediction is a two-step process:

1. Use data \((X_1, Y_1), \ldots, (X_n, Y_n)\) to estimate \(f\) with \(\hat{f}\)
2. Feed new input \(X\) through \(\hat{f}\) to get estimated output:

\[
\hat{Y} = \hat{f}(X)
\]

How accurate can we make \(\hat{Y}\)?

\[
Y - \hat{Y} = \epsilon + f(X) - \hat{f}(X)
\]

- irreducible error: \(\epsilon\)
- reducible error: \(f(X) - \hat{f}(X)\)
Generally, we measure our success by the expected mean squared error (MSE):

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

Fix both $X$ and $\hat{f}$. What are the reducible and irreducible errors with the MSE?

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

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

$$= (f(X) - \hat{f}(X))^2 + \text{Var}(\epsilon)$$
Inference

If you have a new product, how should you spend your advertising money to generate the most sales?

Which media contribute to sales? Which gives the biggest boost? If I spend more on advertising, how much should sales increase?
In **inference**, we want to learn about relationships between inputs and outputs:

- how will increasing one input affect the output?
- is a specific combination of inputs associated with an increase in the output?

Let’s load the housing data.

Can you think of some inference questions? Prediction questions? What is the difference between the two?
Fitting \( f \)

Suppose we have \( n \) observations, \((x_1, y_1), \ldots, (x_n, y_n)\), where \( x_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]^T \).

OK, I now have:

- a question that I want to answer (prediction or inference)
- data

To answer my question, I need to estimate the relationship

\[ Y = f(X) + \epsilon. \]

How do I find \( \hat{f} \) using \((x_1, y_1), \ldots, (x_n, y_n)\)?
Fitting $f$

How do I find $\hat{f}$ using $(x_1, y_1), \ldots, (x_n, y_n)$?

1. select a statistical model
2. select the model parameters using the data

What types of statistical models are there?

- parametric: described by a finite number of parameters
- non-parametric: not described by a finite number of parameters
A **parametric model** is a statistical model described by a finite number of parameters. Examples include:

- a Gaussian distribution (parameters are $\mu$ and $\sigma^2$)
- a Bernoulli distribution (parameter is $\pi$)
- a linear model

$$ Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon $$

income $\approx \beta_0 + \beta_1 \times \text{years of education} + \beta_2 \times \text{seniority}$

What are the parameters of a linear model?
income \approx \beta_0 + \beta_1 \times \text{years of education} + \beta_2 \times \text{seniority}

What does this model say about the structure of $f$?
income \approx \beta_0 + \beta_1 \times \text{years of education} + \beta_2 \times \text{seniority}

Is this model good for prediction? What can it tell us for inference?
Parametric models:

- few parameters (when is this good?)
- well-described interactions between inputs, parameters and output (when is this useful?)
- limited flexibility (desirable or undesirable?)
Nonparametric models are not described by a finite number of parameters.

So, what does that mean?

Let’s learn about an example: \( k \)-nearest neighbors (kNN)
Idea: average the values of the $k$ closest observations

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

where $N_k(x)$ is the set of observations with the $k$ smallest distances to the query point $x$

- *Classification*: pick the majority label
- *Regression*: average the values

Why is this a nonparametric model? Does it have parameters?
I have data on my past running. I recorded the temperature and whether the run was fun:

- Temperature (degrees F)
- Wind Speed (mph)
- Fun (yes, no)

It is now 65 degrees and the wind is 9 mph. Will my run be fun?
**$k$-Nearest Neighbors: Classification**

The diagram illustrates the concept of $k$-Nearest Neighbors (kNN) for classification. It shows a scatter plot with two dimensions: Temperature (x-axis) and Wind (y-axis). The points are labeled as Fun (red circles) and Not Fun (white circles). The query point is marked with a green dot and a circle, indicating that its nearest neighbors are classified as Fun, leading to the prediction that the user will have fun.

**Will I have fun?**

1 Nearest Neighbor: Yes

---

Lecture 4  
Data Mining  
September 12, 2013  
24 / 38
$k$-Nearest Neighbors: Classification

![Graph showing wind vs. temperature with points marked as 'Fun' or 'Not Fun'.]

Will I have fun?
1nn: Yes
2nn: Yes or No

Temp

Wind

0 5 10 15 20 25 30
20 40 60 80

2 Nearest Neighbor
$k$-Nearest Neighbors: Classification

![Graph showing wind vs temp with nearest neighbors classification for fun and not fun decisions.](image-url)
**k-Nearest Neighbors: Classification**

Will I have fun?
1nn: No

1 Nearest Neighbor
$k$-Nearest Neighbors: Classification

Will I have fun?
1nn: No
2nn: No

2 Nearest Neighbor
$k$-Nearest Neighbors: Classification

Will I have fun?
1nn: No
2nn: No
3nn: No

3 Nearest Neighbor

Fun
Not Fun
1-Nearest Neighbor: Congressional Approval

Date
1980 1990 2000 2010
Approval
$k$-Nearest Neighbors: Regression

5-Nearest Neighbors: Congressional Approval
$k$-Nearest Neighbors: Regression

10-Nearest Neighbors: Congressional Approval

Date

Approval

1980 1990 2000 2010

20 40 60 80
**k-Nearest Neighbors: Regression**

50-Nearest Neighbors: Congressional Approval

![Graph showing 50-Nearest Neighbors for Congressional Approval over dates from 1980 to 2010.](image)
Nonparametric vs Parametric Models

Let’s go back to Advertising. Suppose that we fit it with a linear model and kNN.
► which model will produce a more accurate prediction? with a lot of data? with a little data?
► which model will tell us about the which media will produce the best return?
► which model will be faster to evaluate when we have a lot of data ($n$ is large)?
Nonparametric models are generally more flexible than parametric models. Why would we ever want a more restrictive model?

- prevent overfitting
- data compression
- **interpretability**: model parameters often mean something

Often, more interpretable models are less flexible and vice versa.
**Loops** iterate through a predetermined set of values
- a way to do operations over and over on different sets of data
- can loop through indices of lists or arrays

Let's load `Auto.csv` and find the average miles per gallon... with a loop.

```r
Auto <- read.csv("Auto.csv",header=T,na.strings="?")
Auto <- na.omit(Auto) # Remove NAs
attach(Auto)
total.mpg <- 0
for (i in 1:length(mpg)){
  total.mpg <- total.mpg + mpg[i]
}
mean.mpg <- total.mpg/length(mpg)
mean.mpg
mean(mpg)
```
You can write **functions** as well. Some examples of functions are `mean()`, `length()`, and `ls()`. They are useful for storing subroutines that you wish to call multiple times. Let’s write a function that computes the mean of a vector using a loop.

```r
my.mean <- function(vec){
  # my.mean computes and returns the mean of vec
  # Inputs:
  #   vec : vector of numerics
  # Outputs:
  #   mu : mean of vec
  mu <- 0
  for (i in 1:length(vec)){
    mu <- mu + vec[i]/length(vec)
  }
  return(mu)
}
```
OK, let’s try another function and let’s do a bit of debugging. This is more of an art than a science.