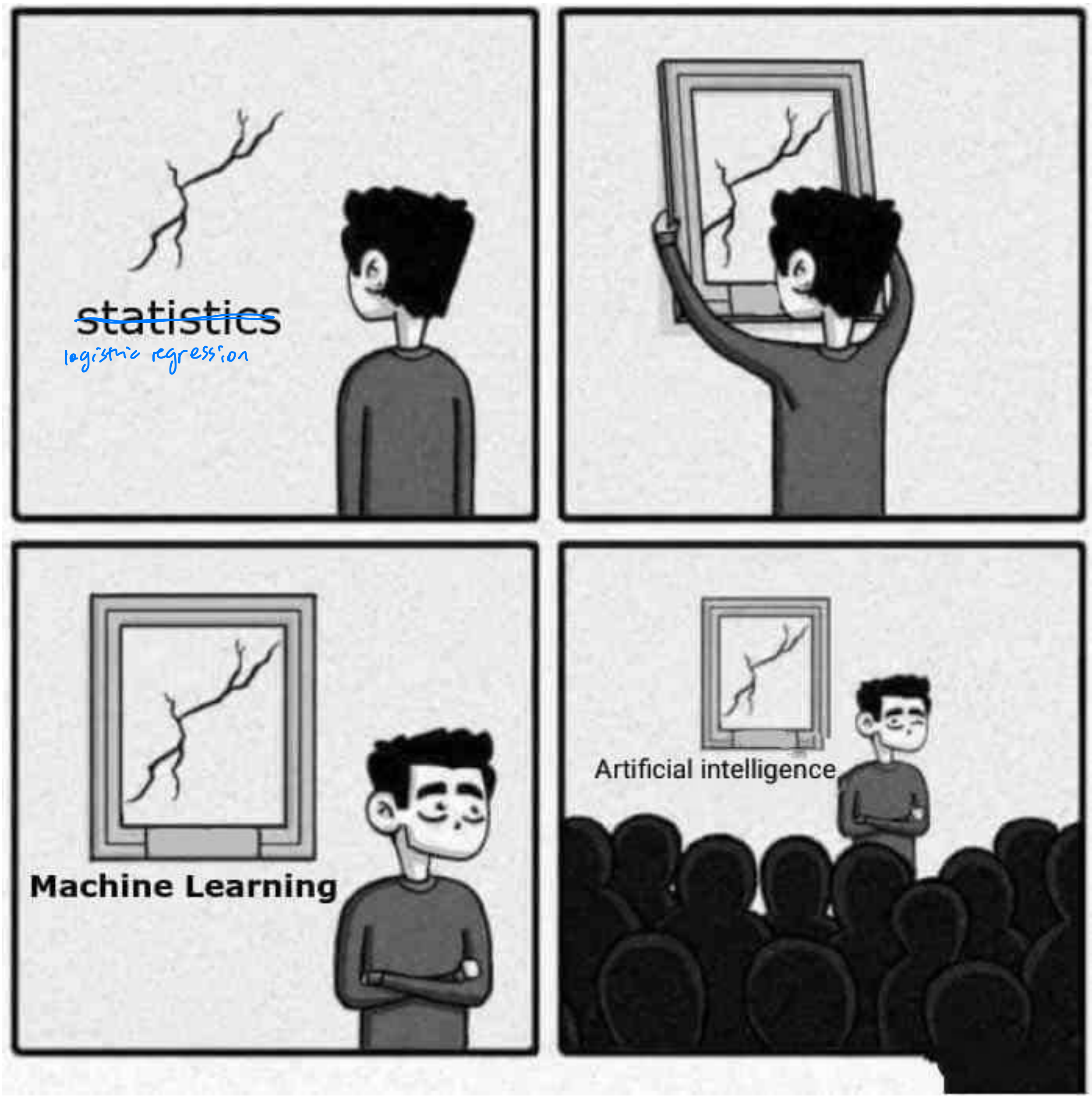
Chapter 2: Statistical Learning



Credit: https://www.instagram.com/sandserifcomics/

statistical machine learning is more than just statistics and more than just machine learning.

We choose methods based on data AND our goals.

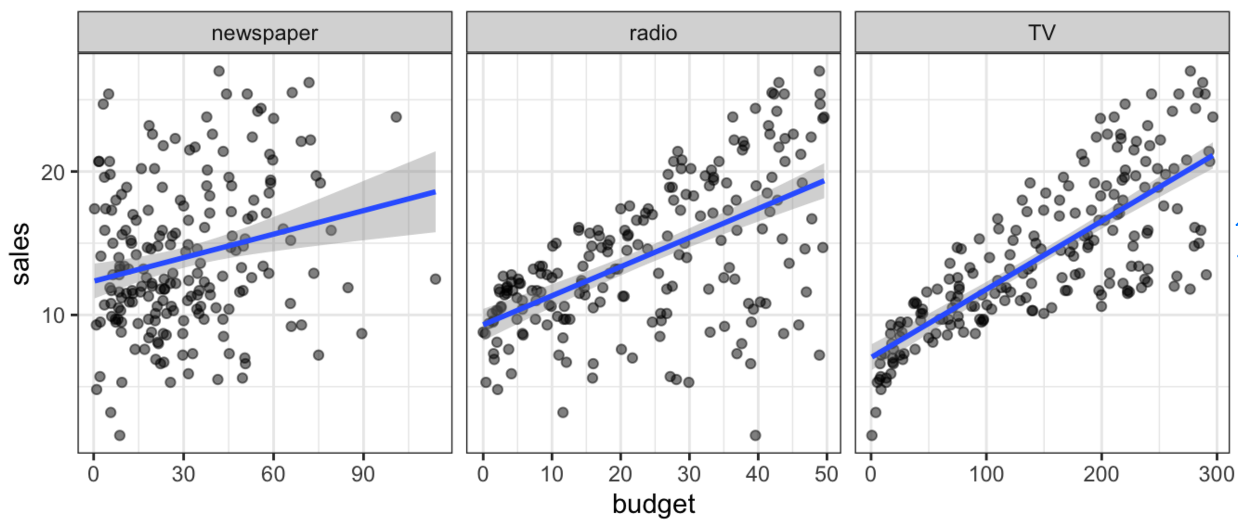
1 What is Statistical Learning?

A scenario: We are consultants hired by a client to provide advice on how to improve sales of a product.

X_1 TV	X_2 radio	X_3 newspaper	Y sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5

⋮ $n=200$

We have the advertising budgets for that product in 200 markets and the sales in those markets. It is not possible to increase sales directly, but the client can change how they budget for advertising. **How should we advise our client?**



If there is a relationship between ads and sales we can tell the client how to advertise to increase sales.

⇒ develop an accurate model to predict sales on 3 media budgets.

input variables "predictors", "independent variables", "features"

advertising budgets

X_1 - TV

X_2 - radio

X_3 - newspaper

output variable "response", "dependent variable"

Y - sales

More generally – observe quantitative response Y and p predictors $X_1, \ldots, X_p$

Assume there is some relationship between response and predictors.

fixed but unknown

random error term, mean 0 and independent of X

$$Y = f(X) + e.$$

systematic information that X provides about Y.

f can involve more than one variable (e.g. TV, radio, newspaper).

Essentially, *statistical learning* is a set of approaches for estimating f.

1.1 Why estimate f?

There are two main reasons we may wish to estimate f.

our goals for an analysis.

Prediction

In many cases, inputs X are readily available, but the output Y cannot be readily obtained (or is expensive to obtain). In this case, we can predict Y using

prediction for Y $\rightarrow$
$$\hat{Y} = \hat{f}(X)$$
estimate of f

remember error averages to 0.

In this case, $\hat{f}$ is often treated as a "black box", i.e. we don't care much about it as long as it yields accurate predictions for Y.

exact form not as important

The accuracy of $\hat{Y}$ in predicting Y depends on two quantities, *reducible* and *irreducible* error.

reducible: $\hat{f}$ is not a perfect estimate for f, but we can reduce error by using an appropriate statistical learning method to estimate it.

irreducible: Even if $\hat{f}$ was estimated perfectly we would still have some error because $\hat{Y} = \hat{f}(X)$ but Y is a function of e! We cannot reduce this no matter how well we estimate f.

why? e contains unmeasured variables that might be useful in predicting Y, or measurement error.

Consider an estimate $\hat{f}$ and predictor X (fixed).

expected value of squared difference between predicted and actual Y.

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

$$= \underbrace{E\left[\left(f(X) - \hat{f}(X)\right)^2\right]}_{\text{reducible}} + \underbrace{\text{Var}(e)}_{\text{irreducible}}$$

variance of error term.

We will focus on techniques to estimate f with the aim of reducing the reducible error. It is important to remember that the irreducible error will always be there and gives an upper bound on our accuracy. almost always unknown in practice.

Inference

Sometimes we are interested in understanding the way Y is affected as $X_1, \ldots, X_p$ change. We want to estimate f, but our goal isn't to necessarily predict Y. Instead we want to understand the relationship between X and Y.

i.e. how Y changes as a function of $X_1, \ldots, X_p$

$\Rightarrow \hat{f}$ no longer a black box! We need to know its form.

We may be interested in the following questions:

1. Which predictors are associated w/ the response?

   often only a small fractions are substantially associated w/ Y $\Rightarrow$ identifying important predictors can be useful

2. What is the relationship between response and predictor?

   positive? negative? linear? etc.

3. Can the relationship between Y and each predictor be adequately summarized by a linear equation or is it more complex?

To return to our advertising data,

inference questions

- Which media contribute to sales?
- Which media generate the biggest boost in sales?
- How much of an increase in sales is associated w/ a given increase in TV budget?

prediction question.

- What can I expect sales to be if we spend \$200k on TV ads and \$0 on newspaper and radio?

Depending on our goals, different statistical learning methods may be more attractive.

e.g. linear models allow interpretable inference but may not give the most accurate predictions.

highly nonlinear approaches can provide accurate predictions but much less interpretable (inference is challenging or impossible).

1.2 How do we estimate f?

We have observed n different data points ← "training data" & we want to estimate "train" f w/ $\hat{f}$

Goal:

apply a statistical learning method to the training data in order to estimate our unknown function f.

In other words, find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X, Y). We can characterize this task as either *parametric* or *non-parametric*

Parametric

1. Make an assumption about the shape of f.

   e.g. $f(\underline{X}) = \underline{\beta_0} + \underline{\beta_1} X_1 + \dots + \underline{\beta_p} X_p$ ← "f is linear in $\underline{X}$"

   parameters.

2. Use the training data to fit or "train" the model.

   e.g. estimate $\beta_0, \beta_1, \dots, \beta_p$ using ordinary least squares (one of many choices).

This approach reduced the problem of estimating f down to estimating a set of *parameters*.

Why?

This simplifies the problem of estimating f.

Disadvantage:

What if the model we choose is very different from the shape of f?

Then the estimate (and any predictions) will be poor.

We can try a more flexible model ⇒ more parameters and can lead to overfitting

⇓

fit errors in training data

Non-parametric

shape

Non-parametric methods do not make explicit assumptions about the functional form of f. Instead we seek an estimate of f that is as close to the data as possible without being too wiggly.

Why?

Advantage:

- fit a wider range of possible shapes for f.
- no restrictions on shape so can't assume wrong shape for f!

e.g. splines. (ch. 7).

Disadvantage

- They don't reduce the problem!
  $\Rightarrow$ need a lot of data.

1.3 Prediction Accuracy and Interpretability

Of the many methods we talk about in this class, some are less flexible – they produce a small range of shapes to estimate f.

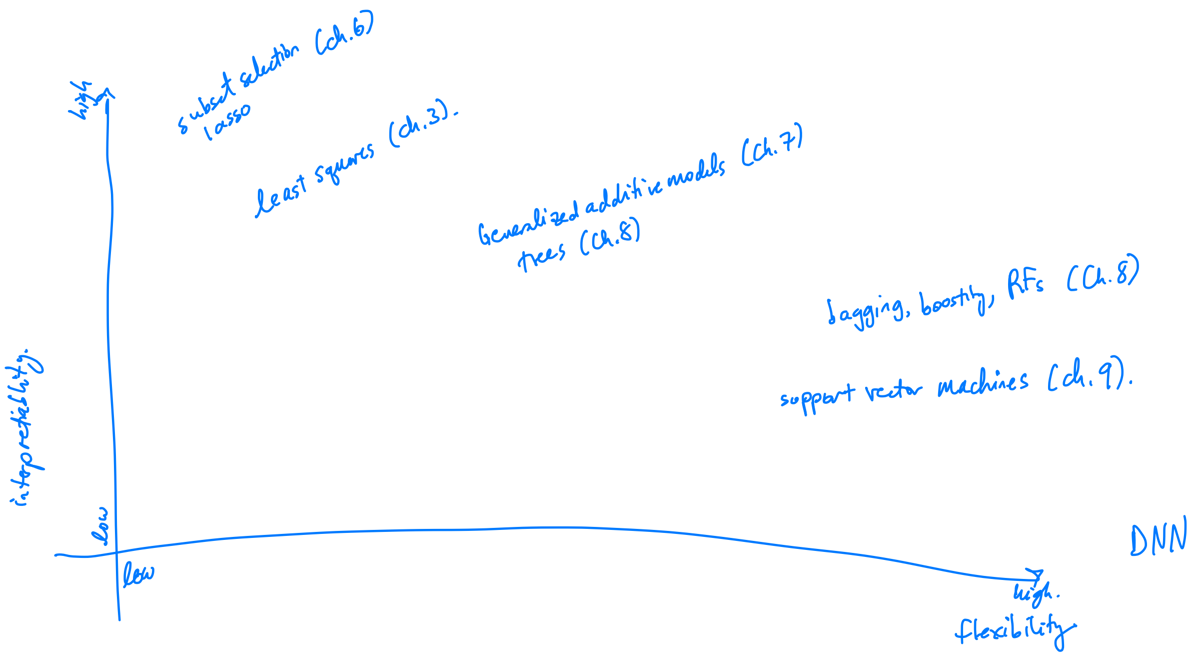
e.g. linear regression vs. splines

Why would we choose a less flexible model over a more flexible one?

will talk about how to choose in Ch.5.

- If you are interested in inference, restrictive models are more interpretable.
- Flexible methods can lead to complicated estimates of f so that it is difficult to understand how any individual predictor is related to the response.

in some setting we care about prediction $\Rightarrow$ flexible model may be preferred.

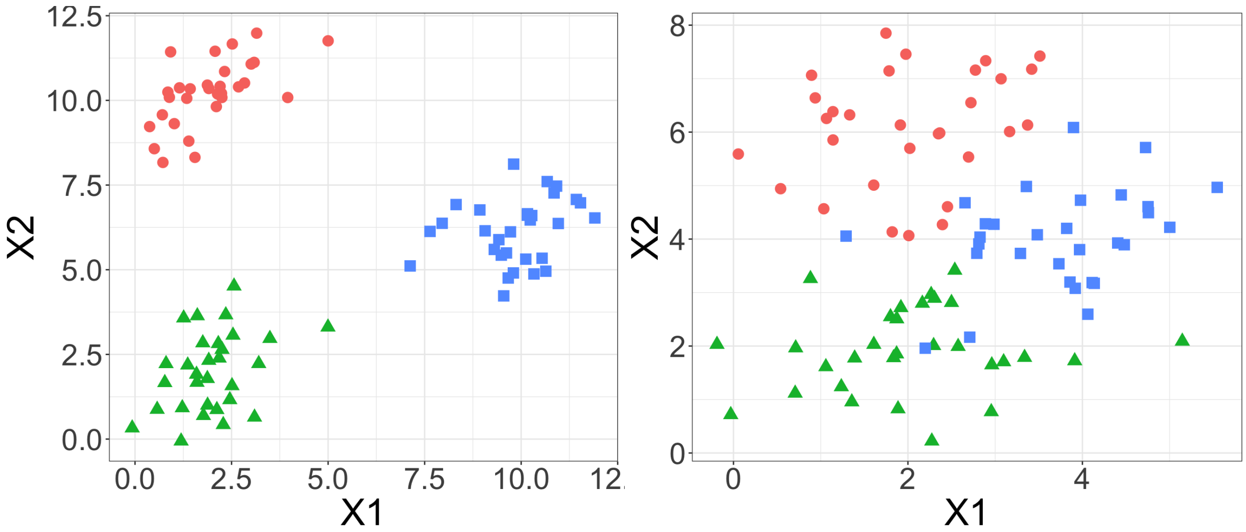


2 Supervised vs. Unsupervised Learning

Most statistical learning problems are either *supervised* or *unsupervised* –

What's possible when we don't have a response variable?

- We can seek to understand the relatopnships between the variables, or
- We can seek to understand the relationships between the observations.



Sometimes it is not so clear whether we are in a supervised or unsupervised problem. For example, we may have $m < n$ observations with a response measurement and $n - m$ observations with no response. Why?

In this case, we want a method that can incorporate all the information we have.

3 Regression vs. Classification

Variables can be either quantitative or categorical.

Examples –

Age

Height

Income

Price of stock

Brand of product purchased

Cancer diagnosis

Color of cat

We tend to select statistical learning methods for supervised problems based on whether the response is quantitative or categorical.

However, when the predictors are quantitative or categorical is less important for this choice.