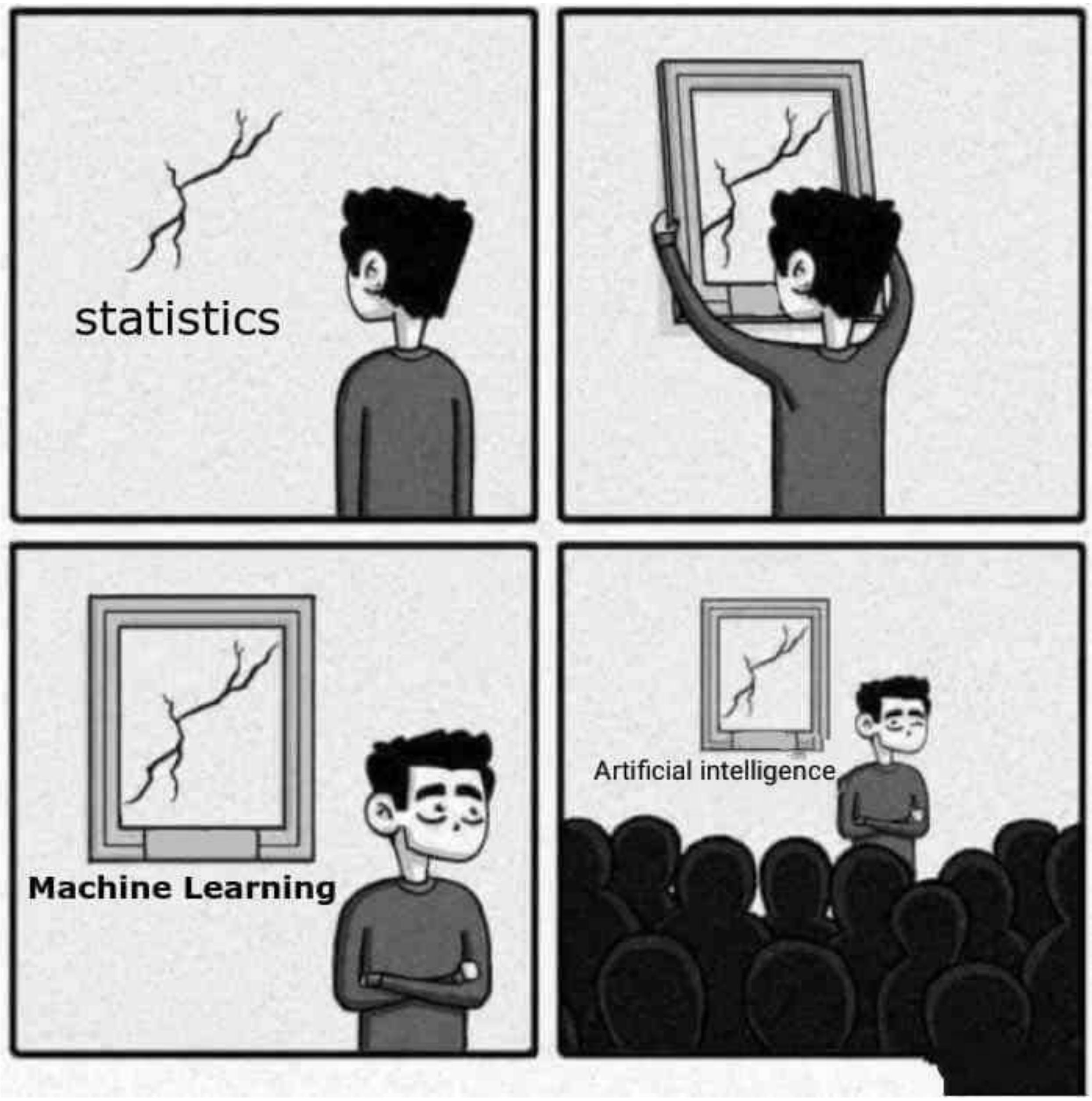
Chapter 2: Statistical Learning



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.

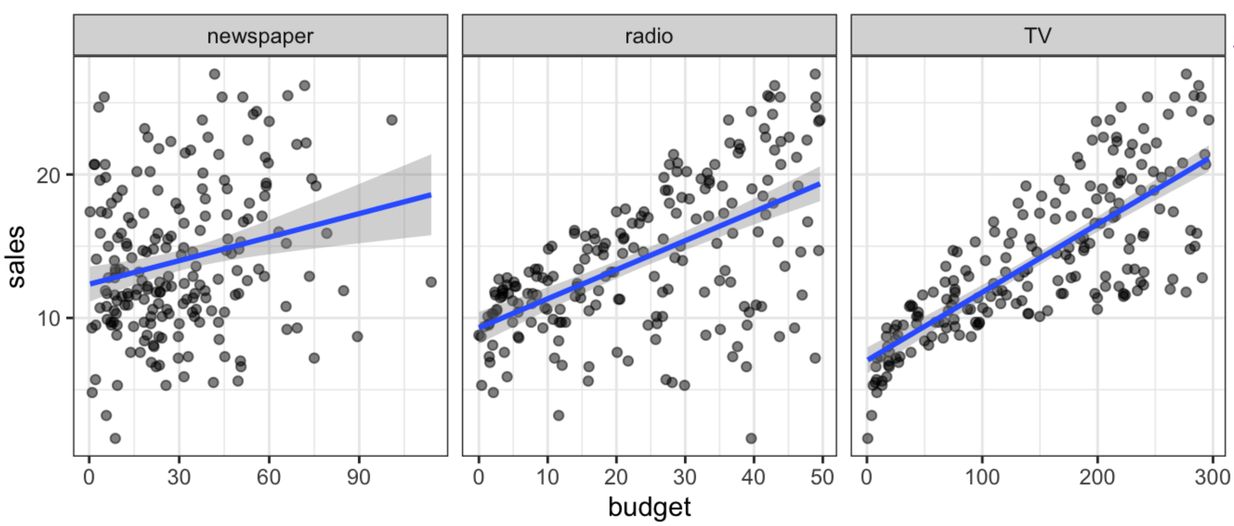
1 What is Statistical Learning?

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

X_1	X_2	X_3	Y
TV	**radio**	**newspaper**	**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 an association between ads and sales we can tell client how to advertise to increase sales.
$\Rightarrow$ develop an accurate model to predict sales based on ad budgets.

input variables "predictors", "indepependent variables", "features"

advertising budgets.
X_1 – TV
X_2 – radio
X_3 – newspaper

output variable "response", "dependent variable"

Y – sales

More generally – Observe quantitative variable Y and p predictors $X_1, \dots, X_p$.
Assuming there is some relationship between predictors ($X_1, \dots, X_p$) and response (Y).

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 1 input 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 → $$\hat{Y} = \hat{f}(X).$$ (remember error averages to 0).

← estimate of f

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 very 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. we can reduce error by using an appropriate statistical learning method to estimate it.

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

Why? e contains unmeasured variables that could be useful for predicting Y, (also measurement error)

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

expected value of squared difference b/w predicted and actual Y. →

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

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 $X_1, \dots, X_p$ 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 $X_1, \dots, X_p$

$\Rightarrow$ $\hat{f}$ 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 fraction are substantially associated w/ response $\Rightarrow$ identifying important predictors can be useful.

2. What is the relationship btw/ response and each predictor?
   some predictors may have a positive (or negative) relationship w/ Y.

3. Can the relationship w/ Y and each predictor be adequately summarized by a linear relationship or is it more complicated?

To return to our advertising data,

Inference questions:

- Which media contribute to sales?
- Which media generate the biggest boost in sales?
- How much increase in sales is associate w/ a given increase in TV ads?

prediction question:

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

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

E.g. linear models allow for simple and interpretable inference but may not yield most accurate predictions.

highly nonlinear models can provide accurate predictions, but much less interpretable.

(inference is often challenging or impossible).

Office hours in STAT 006 (basement)

10am - 12pm (or by appointment).

1.2 How do we estimate f?

"training data" "training"

We have observed n differnt data points & want to estimate f w/ $\hat{f}$

Goal:

apply a statistical learning method to training data to estimate unknown f.

"training"

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., $f(X) = \underbrace{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}_{\text{parameters}}$

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

   e.g. estimate $\beta_0, \beta_1, \dots, \beta_p$ w/ 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 it is usually easier to estimate a set of parameters than to fit an arbitrary function.

Disadvantage:

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

then the estimate (and predictions) will be poor.

We can try a more flexible model, but this means more parameters and can lead to overfitting $\Rightarrow$ fitting errors in training data too closely!

Non-parametric

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

Why?

Advantage:

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

Disadvantage:

- they don't reduce the problem!

=> need a lot of data.

eg. splines (ch. 7).

1.3 Prediction Accuracy and Interpretability

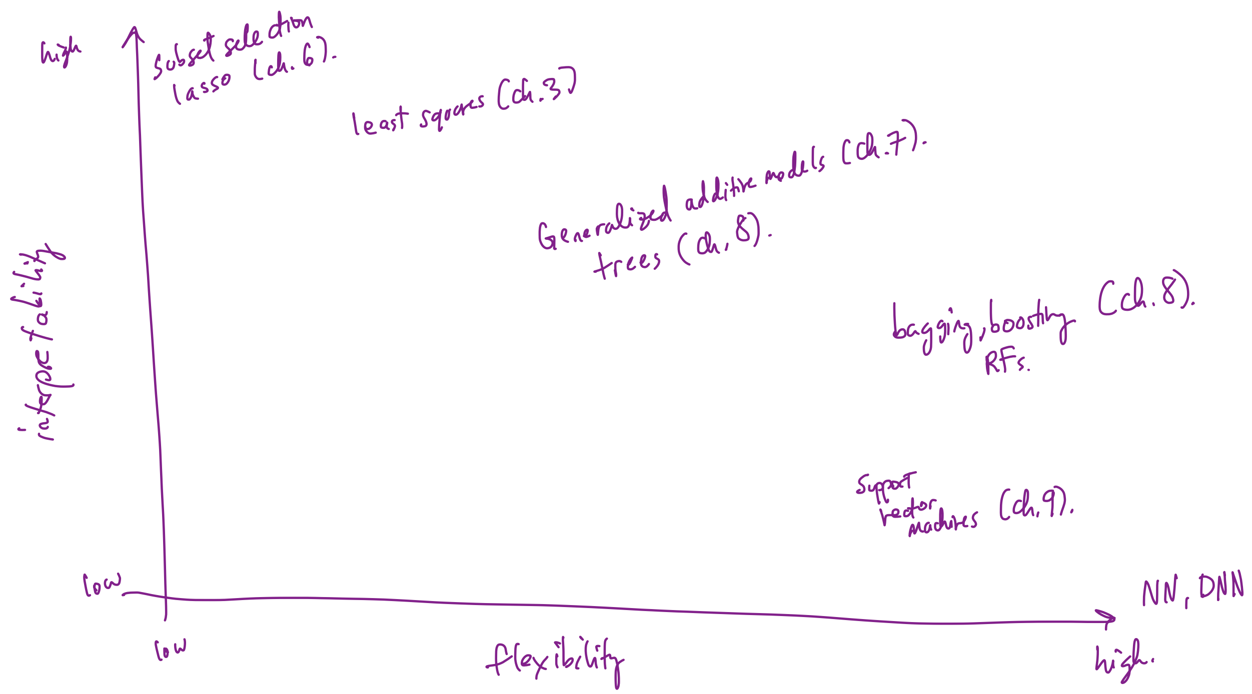
Of the many methods we talk about in this class, some are less flexible – they produce a small range of shapes to estimate f.

e.g. linear regression vs. splines.

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

- If you are interested in inference, restrictive models can be more interpretable.
- flexible methods can lead to complicated estimates of f so that it is difficult for us to understand how individual predictors are associated w/ response.

in some settings we only care about prediction $\Rightarrow$ more flexible model may be prefered.



2 Supervised vs. Unsupervised Learning

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

Supervised

for each observation of predictors x_i, $i=1,\dots,n$ there is an associated response y_i

goal: fit model that relates response to predictors.

maybe for prediction or inference.

methods: linear regression, logistic regression, GAMs, boosting, bagging, RFs (random forests), SVM, etc.

Unsupervised:

for every observation $i=1,\dots,n$ we have a vector of measurements x_i but no response y_i

e.g. cancer example from ch. 1.

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

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

"cluster analysis"

goal: based on observations $x_1, \dots, x_n$ discern if they fall into distinct groups.

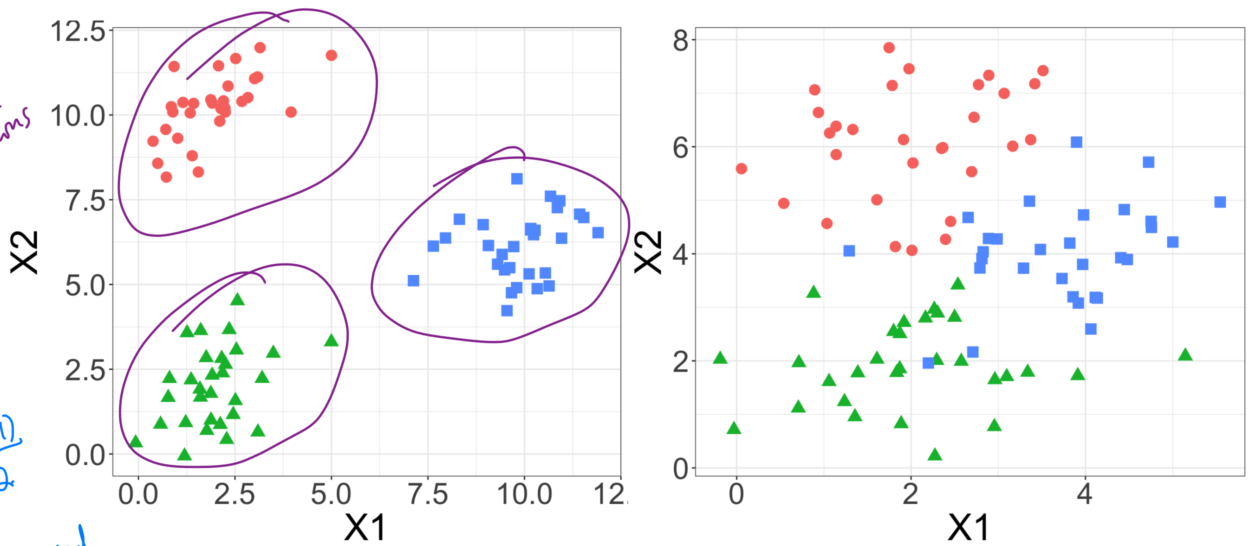
n=90 observations

p = 2

3 groups.

easy to plot

when p > 2 leads to $\frac{p(p-1)}{2}$ distinct scatter plots ⟹ may need more automated approach (ch. 10).



well separated groups would be easy to cluster.

groups overlapping, will be harder to cluster.

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?

Maybe it is expensive to collect y but not x.

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

"semi-supervised" methods

outside the scope of this class.

3 Regression vs. Classification

Variables can be either quantitative or categorical.

numeric values

→ one of K distinct classes or categories.

Examples –

Age

quantitative

Height

quantitative

Income

quantitative

Price of stock

quantitative

Brand of product purchased

Categorical

Cancer diagnosis

Categorical

Color of cat

Categorical. (unless RGB scale).

* 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.

Most methods in this course can use quant. or cat. predictors.