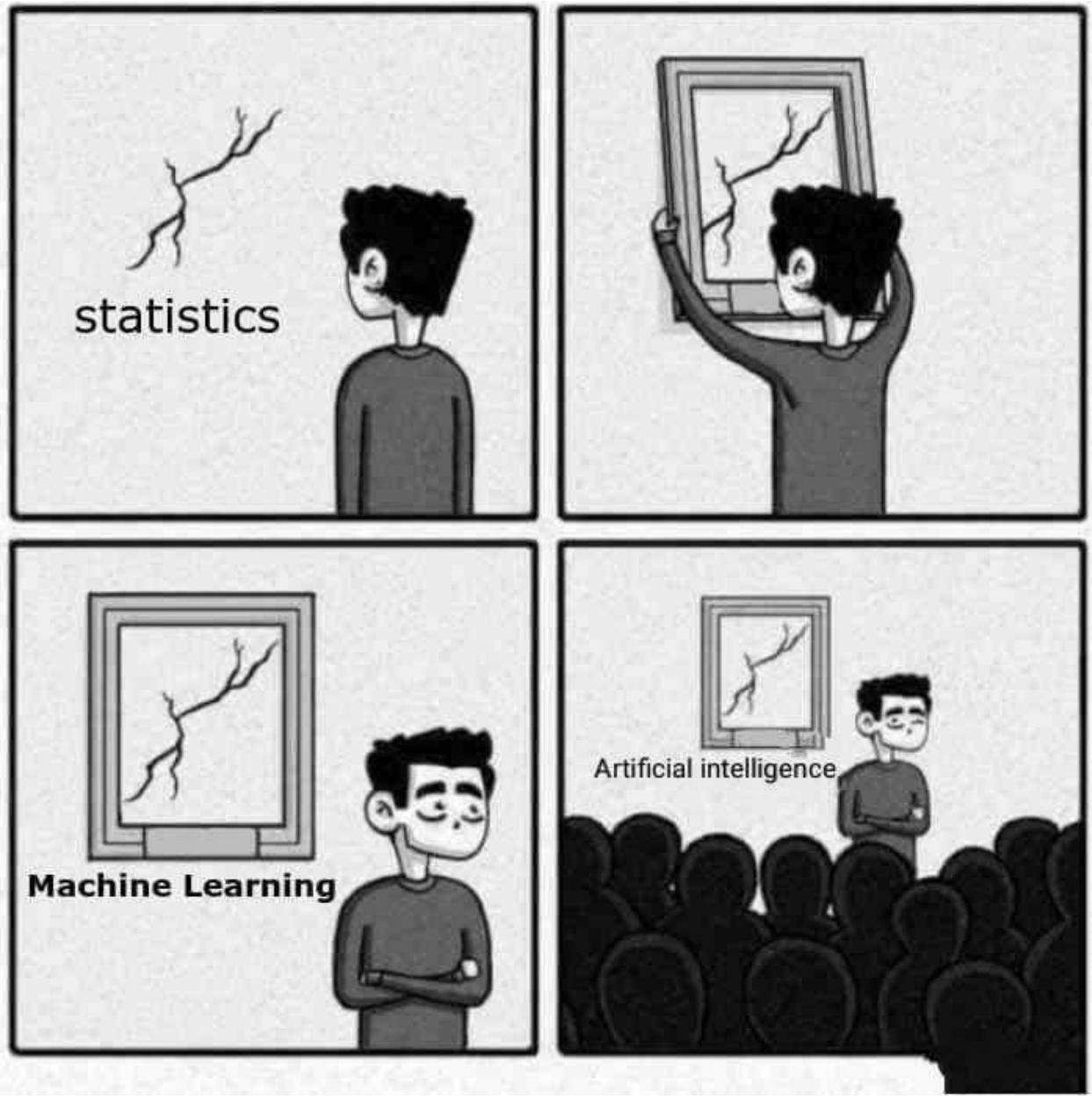
Chapter 2: Statistical Learning



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

- Statistical machine learning is more than just statistics and it is more than just machine learning.
- We choose methods based on data AND our goals.

1 What is Statistical Learning?

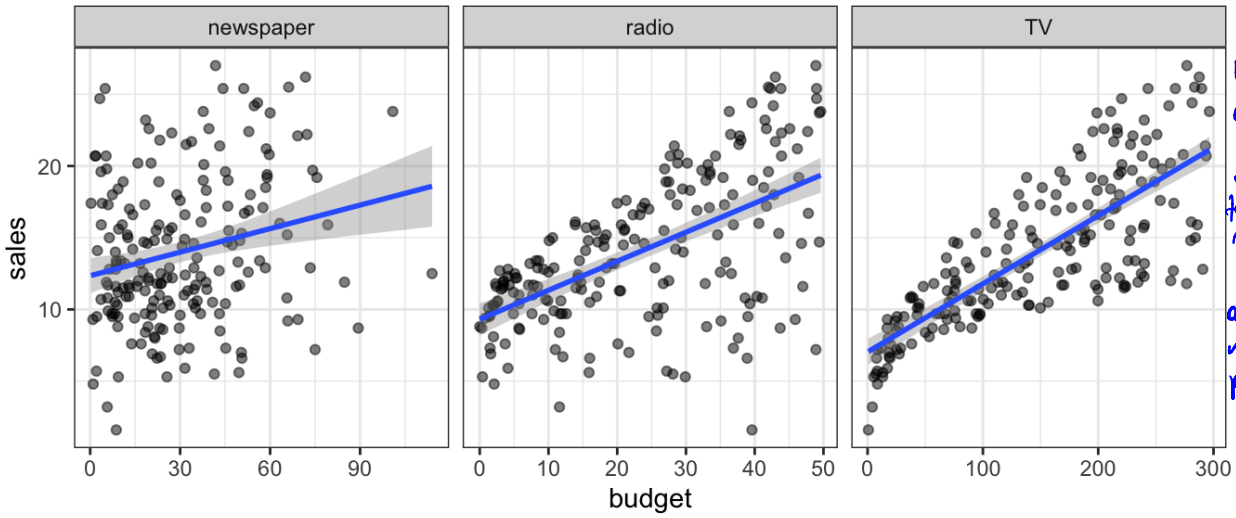
A scenario: We are consultants hired by a client to provide advice on how to improve sales of a product.

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

4 markets

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 sales and advertising, maybe we tell our client how to advertise to increase sales ⇒ develop an accurate model to predict sales based on 3 budgets.

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

advertising budgets

X_1 - TV

X_2 - radio

X_3 - Newspaper

y output variable "response", "dependent variable"

sales

More generally – observe quantitative variable Y and p predictors $X_1, X_2, \ldots, X_p$

assume there is some relationship between predictors and Y.

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

f: fixed but unknown; systematic information that X provides about Y.

e: random error term, mean 0 and independent of X

f can involve more than one input variable (e.g. TV, radio, and 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.

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

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

prediction for Y; $\hat{f}$: estimate of f

remember errors e averages out 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 we estimated f perfectly (with $\hat{f}$) 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 would be useful for prediction of Y

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 does Y change 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 with the response?

   often small fraction of predictors are substantially associated with Y => identify important predictors (ch. 6).

2. What is the relationship between the response and each predictor?

   Some predictors may have positive (or negative) relationship with Y

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

To return to our advertising data,

inferential questions
- Which media contribute to sales?
- Which media generate the biggest boost in sales?
- How much increase in sales is associated w/ a given increase in TV advertising?

prediction question
- What can I expect sales to be if I 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 for simple and interpretable inference but may not yield the most accurate predictions.

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

1.2 How do we estimate f?

"sample" / "training data"

We have observed n different data points & want to estimate (train) f w/ $\hat{f}$

Goal:

apply a (statistical) learning method to training data in order to estimate the 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*

("approximately equal")

Parametric

1. Make an assumption about the shape of f

   e.g. $f(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$

   ($\beta_0, \beta_1, \ldots, \beta_p$: parameters)

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

   e.g. estimate $\beta_0, \beta_1, \ldots, \beta_p$ using 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 because it is usually easier to estimate a set of parameters than to fit an arbitrary function f.

Disadvantage:

What if the model we choose is very different from the shape of f?
Then the estimates (and potentially any predictions) will be 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 (shape) of f. 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 so can't assume the wrong shape for f

e.g. splines (Ch. 7)

disadvantages

- They don't reduce the problem!
  ⇒ need a lot of data
- can't incorporate domain knowledge easily

1.3 Prediction Accuracy and Interpretability

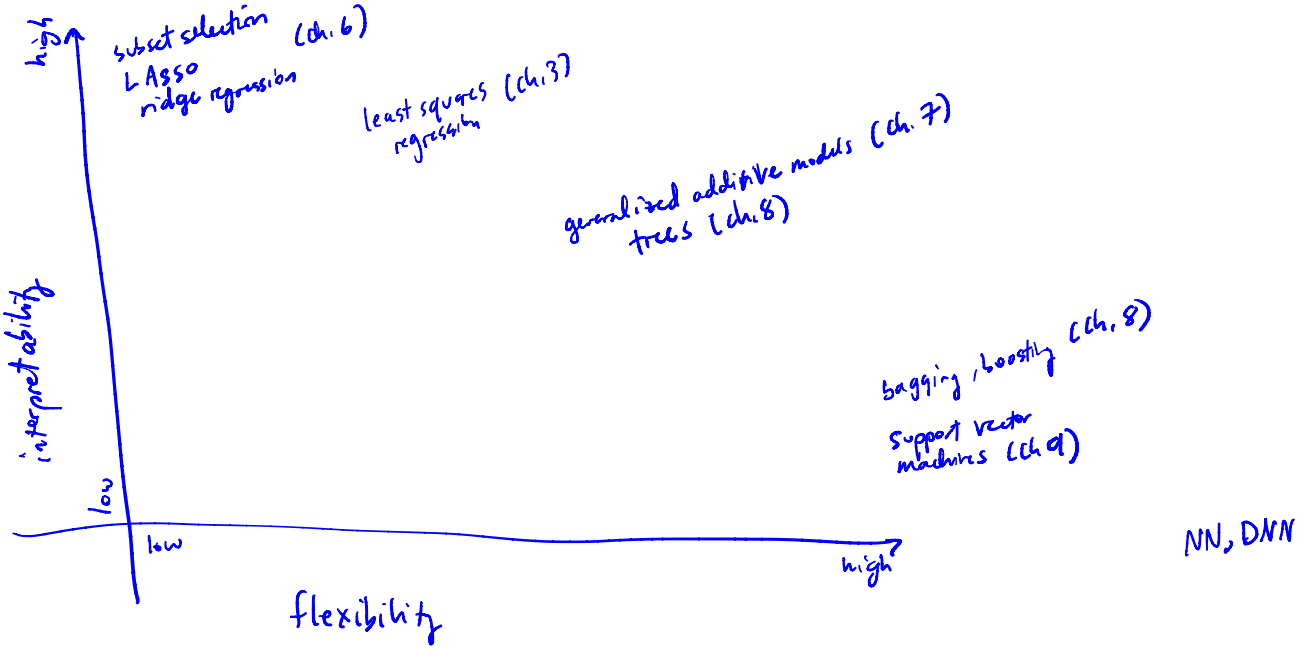
Of the many methods we talk about in this class, some are less flexible – they produce a small range of shapes to estimate f.

e.g. linear regression vs. splines

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

- If we are interested in inference, restrictive models are more interpretable.
- flexible models can lead to complicated estimates of f that are difficult to understand how any individual predictor is associated with the response.

in some settings we only care about prediction => more flexible model may be preferred.



2 Supervised vs. Unsupervised Learning

Most statistical learning problems are either *supervised* or *unsupervised* – (semi-supervised)

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, GAM, boosting, SVM, etc.

Unsupervised

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

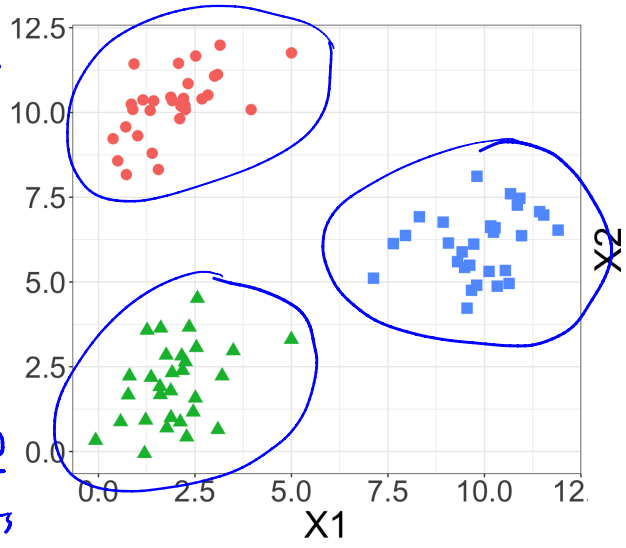
"cluster analysis"

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

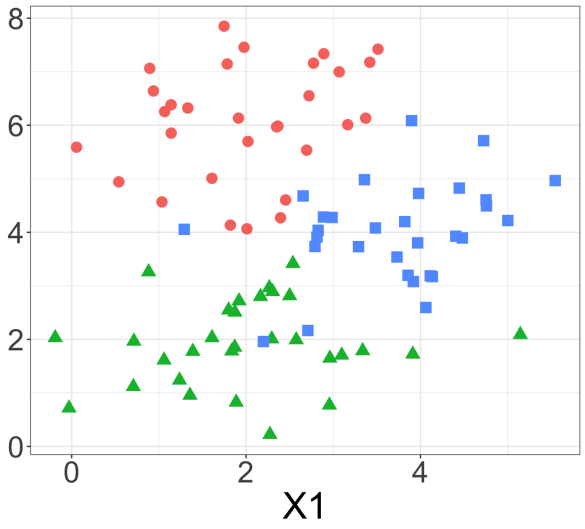
n=90 observations of p=2 variables

easy to plot

3 groups



well separated groups, would be easy to cluster



groups overlapping. will be harder to cluster

when $p > 2$ leads to $\frac{p(p-1)}{2}$ distinct scatter plots to look at

may need more automated approach (or way to project onto lower dimensions) ch. 10.

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 its 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.

3 Regression vs. Classification (supervised)

Variables can be either quantitative or categorical.

↙ numeric values → one of K different 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 – either
name of color – "Categorical"
RGB scale – 3 numeric (quantitative) variables.

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

↓ "regression" ↘ "classification"

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

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