2/3/2021

Lets pretend there are three causal drivers:
Z_1: has sufficient funds to pay back loan at the time it's due?

$Z_1 \in \{0,1\}$

Z_2: unforeseen emergency?
$Z_2 \in \{0,1\}$

Z_3: criminal intent?
$Z_3 \in \{0,1\}$

$y = t\{Z_1, Z_2, Z_3\}$
$= Z_1(1-Z_2)(1-Z_3)$

Problem in patice?
(1) you don't know the z's because they are realized in the future
(2) you may not know the function t which can be very complicated

What is the next best thing since you have to make a decision now and you need a model that works now
You obtain information that approximates the information in the z's and combine this info the approximate y, we denote those proxies that do this approximation the xs and we denote p to be the number of such proxies: $x_1, x_2 \cdots x_p$ for example.
x_1: salary at the time of loan application OR
. missing payment previously $\in \{0,1\}$
x_2:
x_3: criminal charge in the past $\in \{0,1\}$
$\Rightarrow p=3$ X_j's are called features, characteristics, attributes, variables, regressors

If the variable is "nominal categorical" meaning no inherent order, you must do #2 to be able to model.
ex: $x \in \{red, blue, green, \dots\}$

Can we say that $y = f(x_1, x_2, \dots, x_p)$? No! Only approximating at best
$y = f(z_1, \dots z_t)$ or $y = f(x_1, \dots x_p) + \delta$, s.t $\delta = t \neq f$
$y \approx f(x_1, \dots x_p)$

δ is an error in the model, it's error due to ignorance of the true causal drivers, the fact that the proxies aren't the real thing. You are missing info

How do we decrease δ, Increase P with more useful variables
How do we get f? Note that there is no "analytical solution"
The approach we use is learning from data, "Empirical approach"
There are many flavors, We will concentrate on "supervised learning" from "historical data". This requires 3 ingredients:

(1) Training Data $D = \{\langle \vec{x}_1, y_1 \rangle, \langle \vec{x}_2, y_2 \rangle, \dots, \langle \vec{x}_n, y_n \rangle\}$
these are n historical examples of inputs/outputs

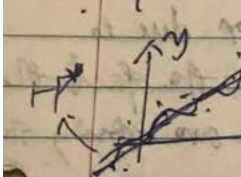
Alternate notation:
$D = \langle X, \vec{y} \rangle$ where $X = \begin{bmatrix} \leftarrow \vec{x}_1 \rightarrow \\ \leftarrow \vec{x}_2 \rightarrow \\ \vdots \\ \leftarrow \vec{x}_n \rightarrow \end{bmatrix}$, $\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$

(2) H = a set of candidate functions with elements h that approximate f. We need this because the space of all functions is too large and ill-defined to directly find the "best one"

$$f(x_1, x_2) = \frac{1}{1+e^{ax_1+bx_2}}$$

(3) We need $A :=$ the algorithm that takes in D, H and returns g, an approximation to f. $g = A(D, H)$.
Is it true that $f \in H$? NO, f is arbitrarily complicated and unknown and the set curly-H contains mindy simple functions that can be fit with curly-A.
However, there is a $h^* \in H$ which is the candidate model that most closely approximates f.

Ex. $p=1$, $x \in R$ $y \in R$



$f(x) = x_1 + 0.1\sin(x)$

$H = \{$ all linear models $\}$ (give more models)

$= b_0 + b_1 x : b_0 \in R, b_1 \in R$

$g = A(D, H)$

model residual (The fall between error/predicted observed)

$y = g(\vec{x}) + e$

$= g(\vec{x}) + [h^*(\vec{x}) - g(\vec{x})] + [f(\vec{x}) - h^*(\vec{x})] + [t(\vec{z}) - f(\vec{x})]$

estimation error ② ③ e

$y = h^*(\vec{x}) + \varepsilon$ (model misspecification error)

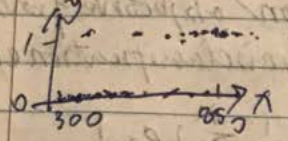
$= h^*(\vec{x}) + (f(\vec{x}) - h^*(\vec{x})) + (t(\vec{z}) - f(\vec{x}))$

(ignorance error)

* How do we decrease model misspecification?
expand the set of candidate functions H to be more complicated and thus more expressive of complex relationships
* How do we decrease estimation error?
increase sample size n (more historical examples). The more

Back to the loan example where $y \in \{0,1\}$
Lets say we have $P=1$ feature, the credit score: $x \in [300, 850]$. So your training data looks like
$D = \langle x, y \rangle = \langle \begin{bmatrix} 613 \\ \vdots \\ 650 \end{bmatrix} \begin{bmatrix} \vec{y} \\ 0 \\ \vdots \end{bmatrix} \rangle$ n

Let's plot the data. what is the "null model" g_0.
which is the model if you didn't have any x's whatsoever.

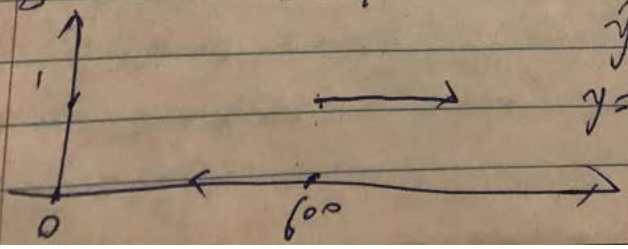


$g_0 = \text{Mode}[\vec{y}]$
what is the simplest possible candidate space H?
$H = \{\mathbb{1}_{x \geq \theta} : \theta \in \mathbb{R}\}$ e.g. $g(x) = \mathbb{1}_{x \geq 600}$

2/8/2021
null model $g_0 = \text{Mode}[\vec{y}]$ if you don't have any x's
$g_0 = \text{Mode}[\vec{y}]$

The simplest possible candidate space H?
$H = \{\mathbb{1}_{x \geq \theta} : \theta \in \mathbb{R}\}$ e.g. $g(x) = \mathbb{1}_{x \geq 600}$

$H = \{\mathbb{1}_{x \geq \theta} : \theta \in \Theta\}$
model parameter
parameter space.



$\hat{y} = g(\vec{x})$
$y = g(\vec{x}) + \epsilon = \hat{y} + \epsilon$
$= \hat{y} + (y - \hat{y})$

Covariates

What is normally done in the real world? you use the features that are available.

To learn from data, you measure the x_{ij}'s on subjects $i = 1 \dots n$.

Let $\vec{x}_i = [x_{i1}, x_{i2}, \dots, x_{ip}] \in \mathcal{X}$, input space.

Subjects are also called: observations, settings, records, objects, inputs

$x_2 \in \{0, 1\}$ ← binary variable } types/names of variables

$x_1 \in \mathbb{R}$ ← continuous variable

Let's consider measuring x_3 differently: this is an ordinal categorical variable

$x_3 \in$ { none, infraction, misdemeanor, felony } 违规 轻罪 重罪

How do we make this a metric?

Method (1) Code it in order of severity spacing by 1:

$x_3 \in \{0, 1, 2, 3\}$.

Downside: coding is arbitrary.

Method (2) Binarize / Dummify the categorical variable.

$x_{3a} \in \{0, 1\}$ infraction or not?

$x_{3b} \in \{0, 1\}$ misdemeanor or not?

$x_{3c} \in \{0, 1\}$ felony or not?

One variable became 3 variables $\Rightarrow p = 5$

I had 4 ~~variables~~ levels, but now only 3.

You can capture the last category by setting all dummy (binary) variables to zero.