MATH 342W || Lecture #2 || Feb 3, 2021

Let's pretend:

There are 3 causal drivers:

z_1: has sufficient funds to pay back loan @ time it's due?

$z_1 \in \{0,1\}$

z_2: unforeseen emergency?

$z_2 \in \{0,1\}$

z_3: criminal intent?

$z_3 \in \{0,1\}$

$$y = t(z_1, z_2, z_3) = z_1(1-z_2)(1-z_3)$$

Problems in practice?

(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.

Q: What is the next best thing, since you have to make a decision now & you need a model that works now?

A: You obtain info. that approximates the info. in the z's & combine this info to approximate y.
we denote these proxies that do this approx. the x's and we denote p to be the # of such proxies: $x_1, x_2, \ldots, x_p$

For example:

X_1: Salary at time of loan application $\in \mathbb{R}$

X_2: missing payments previously $\in \{0,1\}$

X_3: criminal charge in the past $\in \{0,1\}$

$\Rightarrow p=3$

X_j's are called:

features, characteristics, attributes, variables, ind. variables, regressors, covariates.

Q: What is normally done in the real world?
A: You use the features that are available.

To learn from data, you measure the x_j's on subjects $i=1...n$.

Let $\vec{X}_i := [x_{i,1}, x_{i,2}, \dots, x_{i,p}] \in \mathcal{X}$, the input space

"Subjects" are also called:

observations, settings, records, objects, inputs

Types of Variables

	$X_2 \in \{0,1\}$	binary variable
	$X_1 \in \mathbb{R}$	continuous variable
	X_3	also binary

<u>But</u>, let's consider measuring x_3 differently:

$x_3 \in \{$ none, infraction, misdemeanor, felony $\}$

[this is an "<u>ordinal categorical variable</u>".]

Q: How do we make this a metric?

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

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

Downside: Coding is arbitrary.

(2) Binarize / dummify this 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 levels ($L=4$), but now I made $L-1=3$ variables.

<u>Why</u>? You capture the last category (the reference category) by setting all "dummies" / binary variables to zero.

<u>Note</u>: If the variable is "nominal categorical", meaning no inherent order, you must do #2 to be able to use it in a model e.g.

$x \in \{$ red, blue, green, yellow, purple, brown ... $\}$

/ /

Q: Can we say that $y = f(x_1, x_2, \ldots, x_p)$?
A: "No! It is only approximating it at best."
-Gabriel

$y = t(z_1, \ldots, z_t)$ when you don't know the z's.

$y \approx f(x_1, \ldots, x_p)$

OR $y = f(x_1, \ldots, x_p) + \delta$, where $\delta = t - f$

Q: What is δ?
A: It's an error.
It's error due to "ignorance" - ignorance of the true causal drivers. It's the error due to the fact that proxies aren't the real thing.
You are missing information.

Q: How do we decrease δ?
A: Increase p with more useful variables.

Q: How do we get f?
Note: There is no "analytical solution".

A: The approach we use is "learning from data".
This is an "empirical approach".
There are many flavors. We will concentrate on supervised learning from historical data.

This requires three ingredients:

(1) Training Data

$$\mathbb{D} = \{\langle \vec{x}_1, y_1 \rangle, \langle \vec{x}_2, y_2 \rangle, \ldots, \langle \vec{x}_n, y_n \rangle\}$$

(These are n historical examples of inputs/outputs.)

Alternate notation:

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

(2) $\mathcal{H}$:= a set of "candidate functions" w/ elements h that approximate f.

(We need this because the space of all functions is too large and too ill-defined to directly find the "best one". You need to limit this space!)

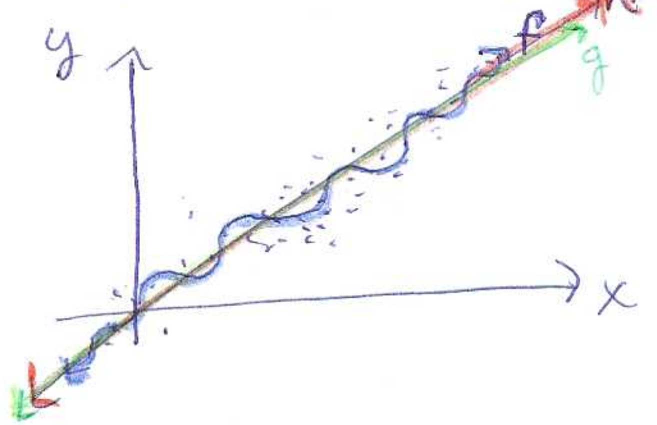
(3) We need $\mathcal{A}$:= the algorithm that takes in $\mathbb{D}$ and $\mathcal{H}$ and returns g, an approximation to f so that:

$$g = \mathcal{A}(\mathbb{D}, \mathcal{H}).$$

¿ Is it true that $f \in \mathcal{H}$? No! f is arbitrarily complicated and the set $\mathcal{H}$ contains usually simple functions that can be fit with $\mathcal{A}$.

However, there is a $h^* \in \mathcal{H}$ which most closely approximates f. Here is an example:

$p=1$, $x \in \mathbb{R}$, $y \in \mathbb{R}$



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

$\mathcal{H} = \{\text{all linear models}\}$

$= \{b_0 + b_1 x : b_0 \in \mathbb{R}, b_1 \in \mathbb{R}\}$

$g = \mathcal{A}(\mathbb{D}, \mathcal{H})$.

$$y = h^*(\vec{x}) + \mathcal{E}$$

$$= h^*(\vec{x}) + \underbrace{\underbrace{[f(\vec{x}) - h^*(\vec{x})]}_{\text{(model mis-specification error)}} + \underbrace{[t(\vec{z}) - f(\vec{x})]}_{\delta \text{ (ignorance error)}}}_{\mathcal{E}}$$

$$\boxed{y = \overbrace{g(\vec{x})}^{\text{model}} + \underbrace{e}_{\substack{\text{residual} \\ \text{("full error")}}}}$$

$$= g(\vec{x}) + \underbrace{\underbrace{[h^*(\vec{x}) - g(\vec{x})]}_{\substack{\text{estimation} \\ \text{error}}} + \underbrace{[f(\vec{x}) - h^*(\vec{x})] + \underbrace{[t(\vec{z}) - f(\vec{x})]}_{\delta}}_{\mathcal{E}}}_{e}$$

Q: What is the "null model" g_0, which is the model if you didn't have any x's whatsoever?

A: $g_0 = \text{Mode}[\vec{y}]$

Q: What is the simplest possible candidate space $\mathcal{H}$?

$$\mathcal{H} = \{\mathbb{1}_{x \geq \theta} : \theta \in \mathcal{X}\} \quad \text{e.g. } g(x) = \mathbb{1}_{x > 600}$$

Until Next Time...