Math 390.4 Lec 3 2/5/16

"Supervised Learning" requires 3 ingredients:

① "Training data"

$$\mathbb{D} := \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_n, y_n)\}$$

these are historical input-output examples

$\vec{x}_1$ may be Bill's characteristics where $y_1 = 1$ (i.e. he paid his loan)

$\vec{x}_2$ may be Jill's " " $y_1 = 1$ " "

$\vec{x}_3$ " " Tony's " " " $y_1 = 0$ " " didn't " "

⋮

there are n examples. Sometimes $\mathbb{D}$ is denoted using vector and matrix notation:

$$\mathbb{D} = (X, \vec{y}) \quad \text{where} \quad X \in \mathcal{X}^n \text{ of dim } n \times p$$
$$\vec{y} \in \mathcal{Y}^n = \{0,1\}^n$$

② $\mathcal{H} = \{$ all candidate functions h that approximate $f\}$

Why needed? f may be a very complicated function you will never be able to learn. So pick a large set of candidate functions that can approximate f.

③ A: an algorithm that takes in $\mathbb{D}, \mathcal{H}$ and selects one best candidate function, g $\Rightarrow g = A(\mathbb{D}, \mathcal{H})$

Recall:

$$y = t(z_1, \ldots, z_t)$$

- y: the phenomenon you wish to explain so you can predict in the future
- t: the true relationship that produces the phenomenon
- $z_1, \ldots, z_t$: the causal attributes about the object that will produce the phenomenon

Farmer Narration

- Θ: the target of statistical inference
- β: distinguished for coefficients in a linear model
- w: "weights" used by computer scientists

$z_1, \ldots, z_t$ are unobservable

but $x_1, \ldots, x_p$ are observable

$$y = f(x_1, \ldots, x_p) + \delta, \quad \delta = t(\vec{z}) - f(\vec{x})$$

- f: the best possible approximate relationship given attributes you can actually measure
- δ: i.e. the difference between the true relationship and the "best we can do" i.e. error due to ignorance

Goal: estimate f

Now... you happen to have historical data $\mathcal{D}$ consisting of n prior cases

You have a simplified function space $\mathcal{H}$ of functions to approx f and an algorithm A. You now use A to produce g. If $f \in \mathcal{H}$,

$$y = g(x_1, x_2, \ldots, x_p) + \varepsilon \quad \text{where} \quad \varepsilon = t(\vec{z}) - g(\vec{x})$$

$$= \underbrace{\left(t(\vec{z}) - f(\vec{x})\right)}_{\text{error due to ignorance}} + \underbrace{\left(f(\vec{x}) - g(\vec{x})\right)}_{\text{parameter estimation error}}$$

The second case $f \notin \mathcal{H}$, this means there is a "closest" function $h^* \in \mathcal{H}$ but due to random chance, we pick g instead.

$\Rightarrow y = g(\vec{x}) + \varepsilon$ where $\varepsilon = t(\vec{x}) - g(\vec{x})$

$= \underbrace{(t(\vec{x}) - f(\vec{x}))}_{\text{error due to ignorance}} + \underbrace{(f(\vec{x}) - h^*(\vec{x}))}_{\text{model misspecification } f \notin \mathcal{H}} + \underbrace{(h^*(\vec{x}) - g(\vec{x}))}_{\text{parameter estimation error } g \neq h^*}$

Thought experiment
if $\rho \uparrow$ what goes down?
if $n \uparrow$ what goes down?

LEC END

"three sources of error"

How to predict for new person? $\vec{x}^*$ ... let $\hat{y} = g(\vec{x}^*)$. Hopefully $\hat{y} \approx y$ most of the time

Back to our problem...

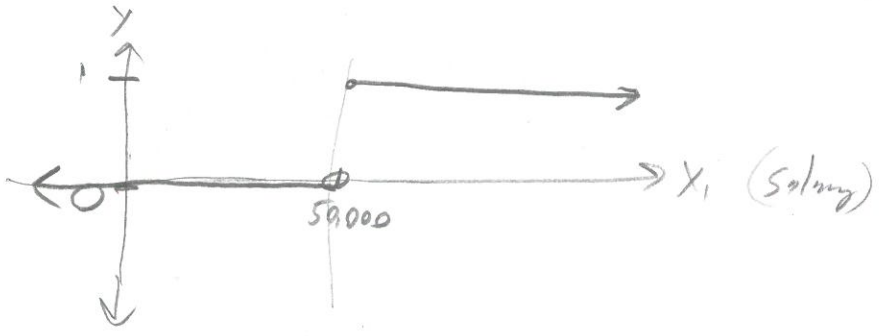
$y \in \{0,1\}$, $\vec{x} \in X$ where $p = 3$. To make it even simpler, let's say we only have x_1, salary.

let $\mathcal{H} = \{ \mathbb{1}\, x \geq x^* \}$

Indicator function is $\mathbb{1}\, x \geq x^* = \begin{cases} 1 & \text{if } x \geq x^* \\ 0 & \text{if not} \end{cases}$

What would a g look like?

Threhold model



Parameter notation

θ, β, w, many other symbols used (unfortunately)

What do all possible choices of $\mathcal{H}$ look like? Elements in $\mathcal{H}$ are indexed by x^*. Every single value $\{x_1 : x^*\}$, the threshold creates a new model. x^* is called a "parameter".