Math 390.4 - Lecture 2 01/31/19~Thursd

Last Time: Philosophy of Modeling

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

it is not the model, it is reality.

But this is impossible because you don't know the z's.

Next Best Thing:

obtain $x_1, \ldots, x_p$ which hopefully captures much of the information in the z's.

$$\underset{\text{BOB}}{\vec{x}_i} = [x_{i1}, x_{i2}, \ldots, x_{ip}] \in \mathcal{X}$$

observation
setting
subject
record
object
input

variables, features
attributes, characteristics
Regressors
Covariates Measurements

"input space"
covariate
space

x_1: credit score $\in \mathbb{R}$

continuous variable

x_2: criminality - Many Metrics

$x_2 \in$ {has past criminal history, does not have}
$\quad\quad\quad\quad\quad\quad\quad\quad \downarrow 1 \quad\quad\quad\quad \downarrow 0$

indicator variable
binary variable
dummy variable

- or -

$x_2 \in$ {none, infraction, misdemeanor, felony}

factor variable, categorical variable
with (L = 4) levels.

Levels are the number of possible states of a factor variable.

Two strategies is use factor var's is med models :

a.) original encoding

$X_2 \in <0, 1, 2, 3>$

ordinal factor varitable

- Major downfall : Encoding is arbitrage

b.) Nominal Encoding

$X_{2a} \in \{0, 1\}$ ~ infraction or not

$X_{2b} \in \{0, 1\}$ ~ misdemeanor or not

$X_{2c} \in \{0, 1\}$ ~ felony or not

$X_{2a} = X_{2b} = X_{2c} = 0 \Rightarrow$ "NONE"

- Downside : $p = 3 \to p = 5$, $L - 1 = 3$

↳ More Regressors

fav. color, states, Make of a car } Things that arent nominal by design.

¿Can you say : $Y = f(x_1, \ldots, x_p)$? NO

if x has an appropimate of z then you can't put the x's together to exactly make y.

$$Y = f(x_1, \ldots, x_p) + \underbrace{\delta}_{t(\vec{z}) - f(\vec{x})}$$ error due to ignorance.

How to minimize δ :

increase the number of relevant variables.

Find f The approach we use is called "learning f from da
an empirical approach.

based on measurements, data, observations.

The type of data we will employ is
"supervised learning".

historical data oversees the

Supervised learning needs 3 ingredients:

① "Training data", "Historical data"

$D = \{ < \vec{X_1}, y_1 >; < \vec{X_2}, y_2 >; \dots < \vec{X_n}, y_n > \}$

n: # of historical examples (sample size)

$\vec{X_1}$ is Bob's measurements $y_1 = 1$ (he repayed)
$\vec{X_2}$ is Jill's measurements $y_2 = 1$ (she repayed)
$\vec{X_3}$ is Bill's measurements $y_3 = 0$ (he did not repay)

$$X = \begin{bmatrix} \vec{X_1} \\ \vec{X_2} \\ \vdots \\ \vec{X_n} \end{bmatrix} \text{ by dimensions } n \times p, \qquad \vec{y} \in \mathcal{Y}^n, \quad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y \end{bmatrix}$$

$x \in \mathcal{X}^n$

$$\mathbb{D} = [X, \vec{y}]$$

Ex: $f : \mathbb{R}^p \to \mathbb{R}$ f is an arbitrary and unknown relationship between $\mathbb{R}^p \to \mathbb{R}$

② $\mathcal{H} :=$ a set of candidate functions h that
Assumption: can approximate f.

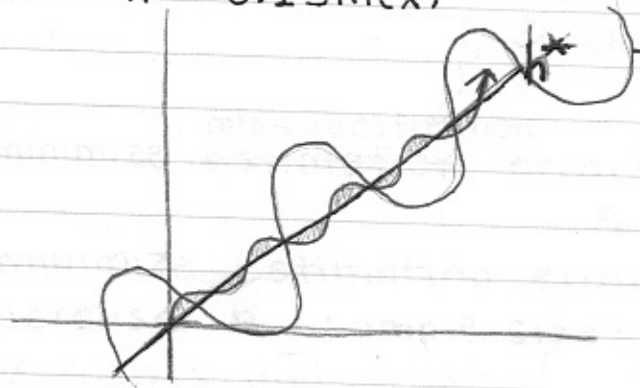
③ $\mathcal{A}$ = an algorithm that takes $\mathcal{H}$ and $\mathbb{D}$ and
provides $g \in \mathcal{H}$ as the best approximatio
of f, which is h^*.

Is $f \in \mathcal{H}$? Generally speaking ... <u>no</u>.
However, $\exists\ h^* \in \mathcal{H}$ that is the best approximation of

$$y = h^*(x_1, \dots, x_p) + \underbrace{f(\vec{x}) - h^*(\vec{x})}_{\text{Misspecification Error}} + \underbrace{t(\vec{z}) - f(\vec{x})}_{\text{Error due to ignorance}}$$

What did I misspecify?

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



your misspecification error just

If you make this assumption then you make lines and are off everywhere its not a line.

Misspecification Error = you didn't include enough possible factors in h.

$\mathcal{H} = \{$all linear functions of $x\}$

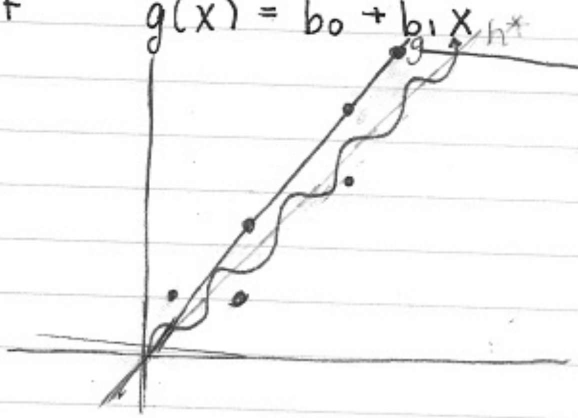
$= \{ \beta_0 + \beta_1 X : \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R} \}$

0 Misspecification Error ← Simple

$\{ \beta_0 + \beta_1 X + \beta_2 \sin X \}, \quad \beta_0, \beta_1, \beta_2 \in \mathbb{R}$

- you have f inside h.
- your h is complex enough, has enough degrees of freedom.

"you want" $g(x) = b_0 + b_1 X$



Why didnt I get H^*?

getting f isn't possible cause youre limited to lines.

getting H^*

$$Y = \underbrace{g(\vec{X})}_{\text{model}} + \underbrace{\underbrace{h^*(\vec{X}) - g(\vec{X})}_{\text{Estimation Error}} + \underbrace{f(\vec{X}) - h^*(\vec{X})}_{\text{Misspecification Error}} + \underbrace{t(\vec{Z}) - f(\vec{X})}_{\text{Error due to ignorance}}}_{e \text{ Residual}}$$

How to predict? For a new object $\vec{X}_*$, how to predict y? $\hat{Y} = g(\vec{X}_*)$

$$Y = h^*(\vec{X}) + \underbrace{f(\vec{X}) - h^*(\vec{X}) + t(\vec{Z}) - f(\vec{X})}_{\varepsilon}$$

choose

How to minimize ~~estimation~~ misspecification error? Make $\mathcal{H}$ richer.
A better?

How to minimize estimation error?
Increase n (sample size)
X matrix and go like $\begin{bmatrix} \downarrow & \downarrow \end{bmatrix}$ not $\begin{bmatrix} \rightarrow \\ \rightarrow \end{bmatrix}$

get more observations.

- all of the errors appear random so you don't know where the error is coming from.