Lecture 2:

$y \in Y = \{0, 1\}$

$Y = t(z_1, z_2, z_3)$ — The function.

z_1, z_2, z_3 — true Causal drivers.

$\downarrow$

Phenomenon

Payback mortgage (1 or 0)

z_1: has the money (1)

z_2: Unforces emeregency (1)

z_3: is criminal (1)

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

Next best thing find features that approximate "information" in z_1, z_2, z_3.

Here are 3 that are directly ~~reated~~ related

X_1: Salary at the time of application (continuous) $\in R$

X_2: Miss Previous Payment (binary) $\in (0,1)$

X_3: Do they have a record? (binary) $\in (0,1)$

X_j's are called "features", "characteristic", "attributes", "Variable", "independent Variables", "regressors", "covariates"

$\vec{X_i}$ is called the ith "observation", "setting", "record", "input", "unit", "subject"

Let P denote the # of features

Let $\vec{X_i} := [X_{i1}, X_{i2}, \cdots, X_{ip}] \in \mathcal{X}$ (e.g., R^p)

input space

$Y = t(z_1, z_2, z_3) \neq f(x_1, \ldots x_p)$

Can I measure x_3 better?

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

Categorical or factor Variable with

$L = 4$ labels

Mathematical models require numerical Values?

What do we do?

Two option:-

① Note: This is a factor Valided with a monotonic order

Code this Variables in

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

0 ↓ none

3 ↓ felony

Downside: the coding is arbitary

e.g. why not $x_3 \in \{0, 1, 5, 100\}$

② Create multiple features:-

$X_{3a} \in \{0,1\}$ is infraction?

$X_{3b} \in \{0,1\}$ is misdemeanor?

$X_{3c} \in \{0,1\}$ is felony?

none is captured by $X_{3a} = X_{3b} = X_{3c} = 0$

L-1 binary variables

Consider $X_3 \in \{$Red, Green, blue$\}$

unordered features

$\Rightarrow$ must use option 2.

$y = t(z_1, z_2, z_3) \neq f(X_1, \dots, X_p)$

$y = t(z_1, z_2, z_3) = f(X_1, \dots, X_p) + \delta$

$\delta := t - f$ is the error due to ignorance

f: is the "best" possible way of combining

$X_1, \ldots, X_p$ to minimize δ.

How to get f?

Analytical Solution? NO

The approach we will use is "learning from data"

Use data to get an estimate of f.

This procedure is also termed "supervised learning"

There are three ingredients?

① Training Data (Data)

Notation: $\mathbb{D}$

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

n: ~~this is to~~ historical examples.
(subject with response)
i.e. it happened already

Sample size

$\vec{X}_1$: Bob's feature Y_1:- 1 (he paid it back)

$\vec{X}_2$: Bells feature Y_2:- 0 (he didn't pay back)

⋮

Standard notation:-

$$X = \begin{bmatrix} \leftarrow \vec{X}_1 \rightarrow \\ \leftarrow X_2 \rightarrow \\ \vdots \\ \leftarrow X_n \rightarrow \end{bmatrix} \qquad \vec{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

$n \times p$ matrix

col. Vector length n.

$$D = \{X, \vec{Y}\}$$

② H := a set of candidate functions for f

Recall f:- $X \rightarrow y$ e.g: $f: R^p \rightarrow R$

You need to simply the set of possible function.

③ A: an Algorithm s.t $g = A(D, H)$

away to select/learn a model,

$g \in \mathcal{F}$ using D.