1

Math 390.4 Lec 2 1/31/18

Creditworthiness Model

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

we had a true system (not a model but sometimes called the "the model")

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

z_1: has sufficient funds

z_2: unforeseen emergency

z_3: criminal intentions

Problem: $\{z_1, z_2, z_3\}$ unobserved is impossible to obtain. What to do?

Next best thing: ~~try to~~ define & collect information "related" to $\{z_1, z_2, z_3\}$ How $\{z_1, z_2, z_3\}$

active process: you need to think about this information, how to collect it, and then actually collect it - may be expensive! take time!

Can't always do this!!

Business as usual: use what you have (or what is easily available)

Better

Let's pretend we've got the resources to "define & collect".

x_1: salary. How to measure: historical avg. salary

x_2: previous loan repayments. " " " ": did they ever miss a previous loan payment?

x_3: historical criminal record: " " " ": previous crime type?

2

Do $\{X_1, X_2, X_3\}$ contain the same info as $\{Z_1, Z_2, Z_3\}$? No.

But ... X_1, X_2, X_3 are possible to observe and Z_1, Z_2, Z_3 are impossible to observe

$\{X_1, X_2, X_3\} \underset{\text{the same info as}}{\overset{\text{have}}{\approx}} \{Z_1, Z_2, Z_3\}$

Note: row vector

Let $\vec{x} = [X_1, X_2, X_3] \in \mathcal{X}$ the "input space" the "covariate space"

an "observation", an "record", an "object", an "input" "indep. var"

features, attributes, characteristics, regressors, covariates

$\dim[\vec{x}] = $ "p" standard notation or "d"

inputs, information, variables, dep. variable

What does the set $\mathcal{X}$ look like?

Valid in a mathematical model? Yes!

$X_1 \in \mathbb{R}$ why negative? Could have debt. X_1 is called a "continuous variable"

$X_2 \in \{$missed a payment, did ~~not~~ miss a payment$\}$

Valid in a mathematical model? No!

$\Rightarrow X_2 \in \{0, 1\}$ X_2 is called a "binary variable" or "dummy variable"

0: did not miss payment; 1: did miss payment

$X_3 \in \{$none, infraction, misdemeanor, felony$\}$

Valid in a mathematical model? No! X_3 is called a categorical variable.

What to do?

Two approaches for categorical variables

a) Code it with an inherent order. This is known as an "ordinal factor/categorical variable"

$$X_3 \in \{\underset{\substack{\uparrow \\ \text{none}}}{0}, \underset{\substack{\uparrow \\ \text{infraction}}}{1}, \underset{\substack{\uparrow \\ \text{misdemeanor}}}{2}, \underset{\substack{\uparrow \\ \text{felony}}}{3}\}$$

Downsides: (a) who picks the numerical encoding? Arbitrary!
(b) Is it ~~truly~~ ordinal?

b) Code it without an inherent order. This is known as a "nominal factor/categorical variable"

How??

$X_{3a} \in \{0, 1\}$ — 0: not none, 1: none

$X_{3b} \in \{0, 1\}$ — 0: not infraction, 1: infraction

$X_{3c} \in \{0, 1\}$ — 0: not misdemeanor, 1: misdemeanor

$X_{3d} \in \{0, 1\}$ — 0: not felony, 1: felony

X_3 is now 4 different dummy variables.

$$p = 3 \to 6$$

This may be good or bad... we will see why later.

Once again... we are trying to find a model for y, creditworthiness for Bob.
The true model is

$y = t(z_1, z_2, z_3)$ but we cannot observe z's for Bob

But we do have

X_1, X_2, X_3 that we observe for Bob.

Since $\{x_1, x_2, x_3\}$ has a lot of the information contained in $\{z_1, z_2, z_3\}$

Can we say $y = t(x_1, x_2, x_3)$? No... not the same!

Can we say

$$y = f(x_1, x_2, x_3) \; ?$$

No... we cannot use imperfect information that does not exactly capture the phenomenon to explain the phenomenon precisely

Instead...

$$y \approx f(x_1, x_2, x_3)$$

$$\Rightarrow y = f(x_1, x_2, x_3) + \delta$$

some difference between approx & truth.

You cannot model y exactly!

What is f? f is the "best" functional relationship we have.

How do we get f?

Can we solve it analytically? e.g. $h(x) = x^2$ find $\min\{h(x)\}$.

the analytical sol is to take deriv, set $= 0$.

There is no analytical sol since there is no governing theory we can use to logically deduce the answer.

We can use an "empirical solution" using "historical data".

This is called "learning from data". Many flavors...

the first (most common) is "supervised learning".

"Supervised Learning" requires 3 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 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 " " " " " $y_1 = 0$ " " ditto " " "

⋮

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

$$\mathbb{D} = \langle X, \vec{y} \rangle \quad \text{where} \quad \begin{aligned} X &\in \mathcal{X}^n \text{ of dim } n \times p \\ \vec{y} &\in \mathcal{Y}^n = \{0,1\}^n \end{aligned}$$

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