Lecture 02

Let's pretend there are 3 casual drivers

* For a loan

z_1: has sufficient funds to pay back loan at the time it's due? $z_1 \in \{0,1\}$

z_2: Unforseen emergency? $z_2 \in \{0,1\}$

z_3: criminal intent? $z_3 \in \{0,1\}$

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

Problems in practice?

(1) You don't know z's because they are realized in the future

(2) You may not know the function f which can be very complicated

what is the next best thing since you have to make a decision now and you need a model that works now?

You obtain information that approximates the information in the z's and combine this information to approximate y. We denote these proxies that do this approximation the x's and we denote p to be the number of such proxies: $x_1, x_2, \ldots, x_p$ For example:

x_1: Salary at the time of loan application $\in \mathbb{R}$

x_2: missing payments previously $\in \{0,1\}$

x_2: criminal charge in the past $\in \{0,1\}$ ≤ 7 $p=3$

x_j's are called features, characteristics, attributes, variables, independent variables, covariates, regressors.

What is normally done in the real world?
You use the features that are available

To learn from data you measure x_j's
on subjects $i = 1 \ldots n$

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

Subjects are also called observations, settings, records, objects, inputs.

$x_2 \in \{0,1\}$ binary variable
$x_1 \in \mathbb{R}$ continous variable
x_3 is a binary variable
} types of variables

Let's consider measuring $x-3$ differently

$x_3 \in \{$none, infractions, misdemeanor, felony$\}$
this is an ordinal category variable

How do we make this a metric?
(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. Why?

You can capture the last category (called the refrence category) by setting all "dummies" (binary variables to zero

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, ... $\}$

Can we say that $y = f(x_1, x_2, \ldots, x_p)$?

"No It's only approximating at best"
- Gabriel

$y = +(z_1, \ldots, z_t)$ where you don't know + or the z's

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

or

$y = f(x_1, \ldots, x_p) + \delta$ s.t $\delta = +\text{-} f$

What is delta? It's an error. This error due to... Ignorance. Ignorance of the true causal drivers. This errors due to the fact that the process aren't the real thing. You're missing information.

How do we decrease delta? Increase p with more useful variables

How do we get f? Note that there is no "analytical solution". 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

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

these are "historical" examples of inputs/outputs

$D = (X, \vec{y})$ 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. H := Set of candidate functions with 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 the space

3. We need A := the algorithm that takes in D, H and returns g, an approximation to f, $g = A(D, H)$

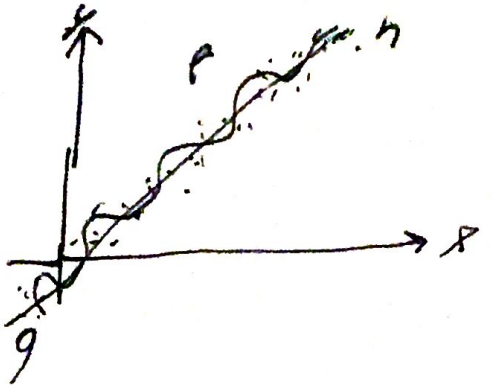
Is it true that $f \in H$?

No f is arbitrarily complicated and unknown and the set curvy-H contains usually simple functions that can be fit

with curly - A

However there is a $h^* \in H$ which is the candidate model that mostly closely approximates f. Here is an example:

$p=1, X \in \mathbb{R}, Y \in \mathbb{R}$



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

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

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

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

$g = h^*(\vec{x}) + \varepsilon$

$= h^*(\vec{x}) + [f(\vec{x}) - h^*(\vec{x})] + (+(\vec{x}) - f(\vec{x}))$

ε

model misspecification error

irreducible error

model residual
(the "full error" the difference between predicted and observed)

$$y = g(\vec{x}) + \varepsilon$$

$$= g(\vec{x}) + h^*(\vec{x}) - g(\vec{x}) \quad \text{estimation error}$$

$$\begin{cases} + f(\vec{x}) - h^*(\vec{x}) \\ + t(\vec{z}) - f(\vec{x}) \end{cases} \quad ; \quad \sigma$$

$$\varepsilon$$

How do we decrease model misspecification error?

Expand the set of candidate functions H to be more complicated and thus more expressive of complex relationships

How do we decrease estimation error?
Increase sample size n (more historical examples). The rows in $\mathcal{D}$

Back to the loan example where y in $\{0, 1\}$

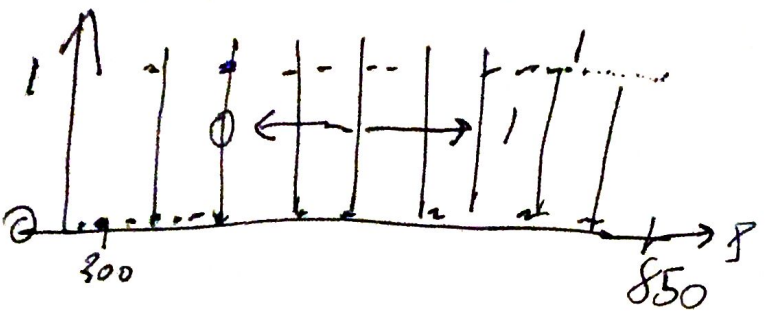
Let's say we have $p=1$ feature, the credit score:

X in [300, 850]. So our training data looks like

$$D = \langle X, \vec{y} \rangle = \left(\begin{bmatrix} 810 \\ 390 \\ 750 \\ \vdots \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \\ \vdots \end{bmatrix} \right) \uparrow n$$

(X: $\begin{bmatrix} 810 \\ 390 \\ 750 \\ \vdots \end{bmatrix}$, $\vec{y}$: $\begin{bmatrix} 1 \\ 0 \\ 1 \\ \vdots \end{bmatrix}$)

Let's plot the training data



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

$$g_0 = \text{mode}[\vec{y}]$$

What is the simplest candidate space H?

$$H = \{ \mathbb{1}_{x \geq \theta} : \theta \in X \}$$

e.g. $g(x) = \mathbb{1}_{x > 600}$