R Language 1/30

- Heine-Borel Theorem: [a,b] is compact in $\mathbb{R}$.

Ctrl + Enter is line by line

- Ctrl + Shift + Enter is whole 'code'
  - rarely used, use it just like you would in Engl's

No commas in R

log(2, base=2)

arccos: acos()

e: exp()

1L will give you integer

- ladder function "1:3" will give you all integers between 1 & 3. Switching to make "3:1" will flip the numbers. Or you could use seq(1,3). seq(1,9, by = 2) will have the #'s jump by 2.
- rep(1, 10) will give you 10 1's.
  rep(c(1,2,3), 10) will give you 10 "1,2,3"'s
- X = 1 assigns the value 1 to X.
  Don't use T or F for variables or Boolean operators.
- one-based indexing for vectors, i.e 1,..., n.
  X = sample(rep(c(0,1), 50)) gives 50 0's, 50 1's in random order.

D.S

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

(z_1, z_2, z_3: true causal drivers; t: true function; y: phenomenon)

pay back mortgage

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

z_1: has the money (1)
z_2: unforseen emergency (1)
z_3: criminal intent (1)

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

Not practical!

Next best thing: find features that approximate the "information" in z_1, z_2, z_3.
There are 3 that are directly related:
x_1: salary at time of application (continuous)
x_2: miss previous payment (binary)
x_3: do they have a record? (binary)

Def: x_j's are called features / characteristics / attributes / variable / independent variables / regressors / covariance

Let p denote the # of features.
Let $\vec{x}_i := [x_{i,1}, \ldots, x_{ip}] \in \mathcal{X}$ "input space"
$\vec{x}_i$ is called the i^{th} observation / record / input / unit / subject / setting.

$\mathcal{X}$ in this example $\in \mathbb{R}^p$.

Can we measure x_3 better?

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

Categorical or factor variable w/ $L = 4$,

Mathematical models require numerical values, what do we do?

- 2 options:

1. Note: this is a factor variable w/ a monotonic order. Code this variable via
$x_3 \in \{0, 1, 2, 3\}$
Downside: the coding is arbitrary, e.g.

$x_3 \in \{0, 1, 5, 100\}$, to put more weight on each.

2. Create multiple features:

$x_{3a} \in \{0, 1\}$ is infraction?
$x_{3b} \in \{0, 1\}$ is misdimeanor?
$x_{3c} \in \{0, 1\}$ is felony?

"None" is captured by $x_{3a} = x_{3b} = x_{3c} = 0$.

→ $L-1$ binary variables

ex. Consider $x_j \in \{red, green, blue\}$

$\Rightarrow$ must use option 2.

S is an <u>unordered factor</u>,

- $y = t(z_1, z_2, z_3) = f(x_1, \ldots, x_p) + \delta$

$\delta := t - f$ is the <u>error due to ignorance</u>,

f is the best possible way of combining $x_1, \ldots, x_p$ to minimize δ.

To minimize δ, need to add as many relevant variables to f as possible. Don't use irrelevant information.

- How to get f

Analytical solution? No.

* The approach we use is learning from data. Use data to get an estimate of f.

This procedure is also termed "supervised learning."

3 ingredients

1. Training data

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

This is n historical examples (subject w/ response) i.e it happened already.

(n: sample size)

$\vec{x}_1$: Bob's features
$y_1 = 1$ (paid back loan)
$\vec{x}_2$: Bill's features
$y_2 = 0$ (did not pay back loan)
⋮

If n is large, then better estimate.

Standard Notation

$$X = \begin{bmatrix} \leftarrow \vec{x}_1 \rightarrow \\ \leftarrow \vec{x}_2 \rightarrow \\ \vdots \\ \leftarrow \vec{x}_n \rightarrow \end{bmatrix}$$ $n \times p$ matrix

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$ col. vector length n

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

2. $H :=$ a set of candidate functions for f.

Recall: $f: \mathcal{X} \rightarrow Y$, e.g $f: \mathbb{R}^p \rightarrow \mathbb{R}$

You need to simplify H.

3. A is an algorithm s.t

"learn"

$g = A(\mathbb{D}, H)$, a way to (select) a model

$g \in H$ using $\mathbb{D}$.