Data Science Lab Lecture 1

Note: For all in a Thursday 3:30 - 6:30 Lab section, please come to EER 1.512 @ 5 pm (only for Jan 18)

This class: predictive modelling

Example from supervised pred. mod.

Q: Will loan default?

	Age	Income	#accts		
B1	59	100	4		1
B2					
⋮					

answer (label)

Find mapping from data: (age, income, #accts) to label: 1/0 default/not

Jargon: Supervised learning = we have the label.

Unsupervised: No label

Ex: clustering

What is a good "mapping" from data → label?

Would like:

ⓐ A good mapping should mostly agree with most of our data.

ⓑ should agree with future data.

But how to test this?

Empirical Risk — How our mapping does on training data

aka Training Error

vs

True Risk - How we do on future data.

aka Generalization Error

Ex: Nanochip Data set

	H	W	failed y	$\hat{y}$
chip 1	0.8	0.8	1	
2	0.3	0.75	0	
3	0.2	0.8	0	

Loss Function

$$L(y, \hat{y}) = \begin{cases} 1 & \text{if } y \neq \hat{y} \\ 0 & \text{if } y = \hat{y} \end{cases}$$

Empirical Risk of a mapping: $h(\underbrace{H, W}) \to \hat{y}$

$$x_i = (H_i, W_i)$$

so $X_2 = (0.3, 0.25)$

$$h(x_i) = \hat{y}_i$$

Empirical Risk

aka Training Error

$$L_S(h) = \frac{1}{n} \sum_{i=1}^{n} L(h(x_i), y_i)$$

Mapping h:

$$h(x) = h(H, W)$$

$$= \begin{cases} \text{If } x = (0.8, 0.8) \to 1 \\ x = (0.3, 0.25) \to 0 \\ x = (0.2, 0.8) \to 0 \\ \text{o.w.} \quad \to 1 \end{cases}$$

Useless.