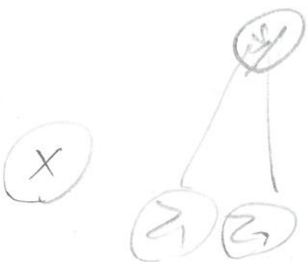
Math 340 Lec 13 5/7/19

"causal inputs"

Basic causality... reminder $y = f(z_1, \ldots, z_8)$

We don't in general see the z's. We see x's. What are x's? Could be anything. corr vs. causation

(A)



Bad model (not cause, not corr)

does it matter? No!

(B)

corr but not causal)

incidental effect of a true cause. ex umbrella sales vs. car accidents

does this matter for pred? No

(C)

(corr & causal)

Can you be causal w/o corr? NO.

Could the situation in (A) have a non-trivial sample corr?

Yes! Chance capitalism. "Spurious corr." X is divorced from causal chain. It is only due to chance... (demo). When you get more n, it disappears.

Could the situation in (B) or (C) not exhibit a corr? Yes! It may be too weak to detect but with more n you can detect it. It will still likely be a low R^2 model.

So if X is corr, is it causative? Yes or no! And you have no idea (without further work)

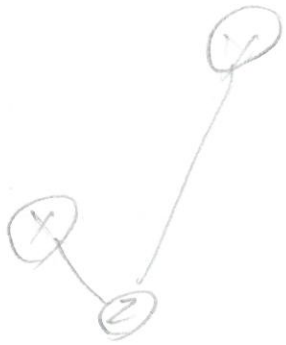
(A) No. Not an corr.

(B) No. Corr but is incident with a real cause (the "lurking variable" ⟹ If non-spurious corr ⟹ ∃ causality somewhere.

(C) Yes < weak / strong

What is an intuitive def. of causality?

If X is manipulated ⟹ Y changes

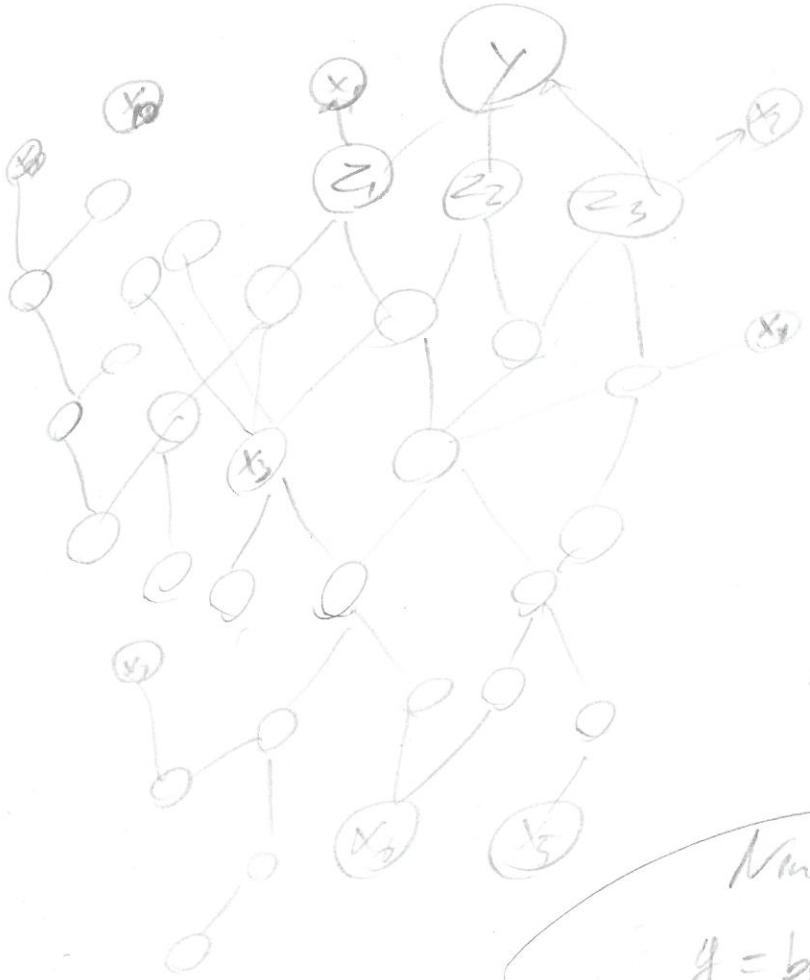


If $X \to X+1 \Rightarrow y$ doesn't change
Umbrellas & accidents

If $Z \to Z+1 \Rightarrow y$ changes

---

the real world dataset $x_1, \ldots, x_p$. How does this look?



$X_1, X_2, X_{10} \Rightarrow$ corr but not causal
$X_3, X_5, X_6 \Rightarrow$ causal
$X_9, X_{10} \Rightarrow$ not corr

You like fit a model
$y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$ & $b_1 = 3.26$
How to interpret b_1? like how the 3.26 was?
Everyone, bungles this up! EVER BE!

Wrong interpretation

If I/you/somebody increases x by one unit or is set to any given observation, y will increase by 3.26.

Why is this wrong? May not be causal, and if causal, may not be linear

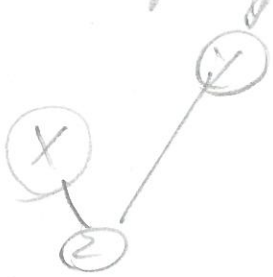
Best I could think of:

When comparing two "randomly observed" observations (A) and (B) sampled in the same way as obs's in (X) where (A) has an x_1 value one unit larger than the x_1 value of (B) but shares the same values of x_2, x_3, then (A) is predicted to have a response y that differs by b_1 units on avg. than the response of (B) assuming the linear model is true.

This is very limited!!

How can you claim the "wrong interpretation"? Need randomized controlled experiment. Even then it is hard!

the being of regression ...



$y = b_0 + b_x x$

b_x is positive and significant

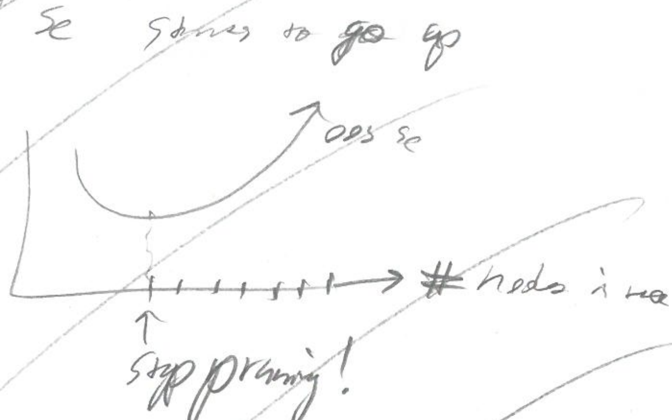
$y = b_0 + b_x x + b_z z$

$\Rightarrow b_z$ is positive and significant & b_x is ≈ 0. Why? Look at regression!

b_x ∴ If z stays the same... (done)

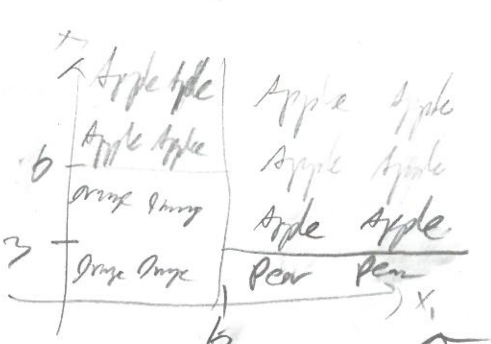
1:17

How to prune a tree: Let $N_0 = 1$ and then "prune". Collapse tree back to internal node one-by-one by looking at minimal reduction in OOS se. Continue until se starts to go up



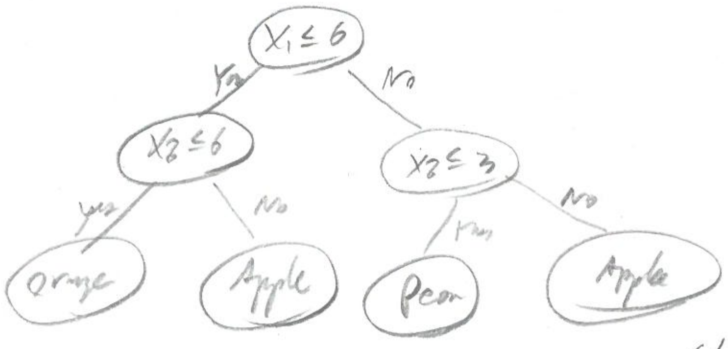
Kind of like backwards stepwise regression. DEMO

If $y = \{1, 2, \dots, L\}$ multilabel classification. If $L=2$, binary classification.



If Binary classification, we discussed perceptron, SVM with logistic objective. Both are linear models and thus suffer from same problem of model selection. Also they can't build models like so

Tree:



Written as a linear model:

$$\hat{y} = g(x) = \text{Orange}\, \mathbb{1}_{x_1 \le 6} \mathbb{1}_{x_2 \le 6} + \text{Apple}\, \mathbb{1}_{x_1 \le 6} \mathbb{1}_{x_2 > 6} + \text{Pear}\, \mathbb{1}_{x_1 > 6} \mathbb{1}_{x_2 \le 3} + \text{Apple}\, \mathbb{1}_{x_1 > 6} \mathbb{1}_{x_2 > 3}$$

Classification Trees (1984)

Algorithm to generate this: Same thing as the regression algorithm except (1) in step 2, the cost becomes

$$\text{Gini} := \sum_{l=1}^{L} \hat{p}_l (1 - \hat{p}_l) \quad \text{where} \quad \hat{p}_l = \frac{\# y_i \text{ in label } l}{\# \text{ in node}}$$