Math 390.03 Lec 24 5/9/18

L1

Variance term

$$MSE = \sigma^2 + \left(\rho \, Var[g_t] + \frac{1-\rho}{T} Var[g_t]\right) + \underbrace{Bias[g_t]^2}_{\text{low}}$$

Under model averaging with same alg. & bootstrapped data $\mathbb{D}_{(i)}$

$\ll Var[g_t]$ itself since $\rho \neq 1$

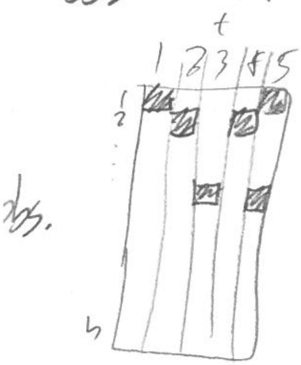
$\xrightarrow{\text{as } T \to \infty} \rho \, Var[g_t]$

this is why bagging works so well

Validation for bagged models (not only trees)

Usually $\mathbb{D} = \mathbb{D}_{train} \cup \mathbb{D}_{test}$, build on train, validate on test.

Here, each tree has its own $\mathbb{D}_{train,t}$, $\mathbb{D}_{test,t}$ "out of bag (oob)". Why? Bootstrap sample. This means that each tree can validate itself by predicting on $\mathbb{D}_{test,t}$. Now let each tree do so and average by observation.



Validating the first obs is done on trees 1 & 5 and we average the $\hat{y}$'s. Validation for the second obs. is done on trees 2 & 4 and we avg. the $\hat{y}$'s.

If T is sufficiently large, all n obs. will be validated. We then compute e_i's and then an error metric e.g. R^2_{oos}, $s_{e,oos}$, etc.

$\Rightarrow$ Out of bag error est (oob error). It works! Everyone does it.

Theoretically it's $\approx$ K-fold CV with $K=2$, but I don't know much about it.

Why does bagging trees work so well?

① No need to specify a model (or do model selection) since trees figure out the model for you.

② No need to specify hyperparameters. Pick a low N_0 (you want the overfitting!! since it leads to low bias).

③ You get validation for free. No need to do K-fold CV.

Can we do better?

$$MSE = \sigma^2 + \underbrace{\rho \, Var(g_t) + \frac{1-\rho}{T} Var(g_t)} + Bias(g_t)^2$$

If we can make ρ as low as possible, then the variance term can be minimized.

ρ is the avg. correlation between the trees' $\hat{y}$'s. How can we "decorrelate" the trees? Use the classification/regression tree alg but

Instead of trying all splits... change this one step to be...

How about at each split, pick a random subset of features

$\{1, 2, \ldots, p_{try}\} \subset \{1, 2, \ldots, p\}$

and then minimize SSE as usual.

What does this do? Each g_t now is a more different than previously. Downside? $Bias(g_t) \uparrow$ but not too much.

regional students

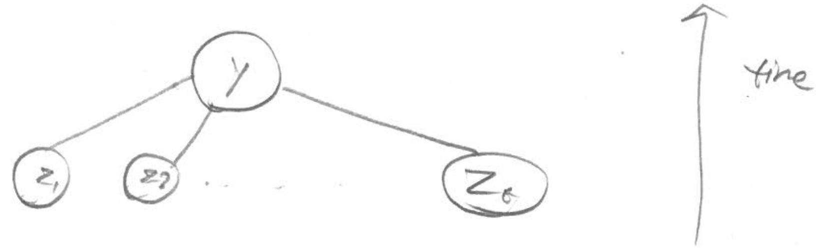
Each one is "random". Together, it's called a "random forest."®

(Breiman, 2001)

For regression

$y = t(z_1, \dots, z_t)$

An inexact illustration



$z_1, \dots, z_t$ are said to be immediate causal factors of the phenomenon y. They occur before y.

This is not so clear! Let's think about income. What are the z's?

Well with one bank acct, it is the program that increases the # in the db!

We don't in general see the z's in our data. We see x's. What are they?

An x could be disconnected

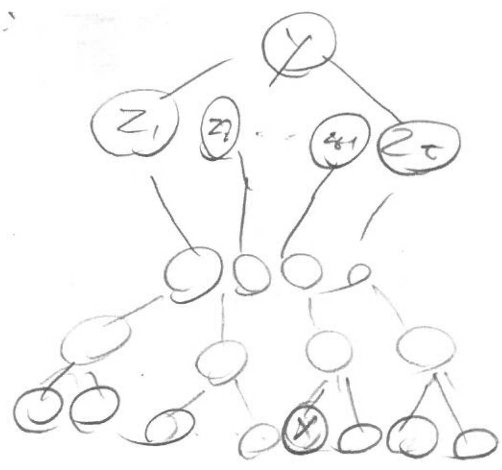
(A)

Ex?

Very bad model!

It could be:

(B)



An x can be indeed causal... but far back in the chain.

Ex?

It could be

(C)

Instead of a cause, it can be an incidental effect of an actual cause

Let's assume n is very high

Correlation w/ y for (A)? No

Corr ... (B)? Yes

" " " (C)? Yes

Let's assume n is moderate

Corr ... (A)? Maybe → it could be by random chance

" " (B)? Maybe → it could be too weak to detect

" " (C)? Maybe → " "

Corr. in (A) is called a "spurious corr". Why? X is divorced from causal chain. It only appears corr w/ Y due to random chance. (Noise) But what happens when you get more data? It disappears! What happens when you get more data for (B) or (C)? You will see the correlation!

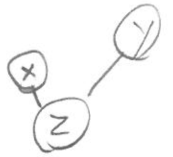
Is corr causation?

(A) No! It's not even a real corr!

(B) Yes! It is remote ... and removed from the immediate effects but it is real

(C) No! It itself is an effect

What must be present to call a variable "causal". An ability to manipulate it and see a change in Y. Below... X & Y correlated.



If you manipulate X here, does Y change?

e.g. # umbrellas sold (X) and car accidents (Y) are correlated. How? Through a common cause (Z) rainfall amount.