Lec 23 Math 390.3 5/7/18

We proved last time...

$$MSE = \sigma^2 + E_{\mathcal{X}}[Var[\hat{g}(\vec{x})]] + E_{\mathcal{X}}[Bias[g(\vec{x})]^2]$$

done

You've seen this before!

$$Y|\vec{X}=\vec{x} = g(\vec{x}) + \underbrace{\left(h^*(\vec{x}) - g(\vec{x})\right)}_{\text{constant in } \Delta\text{'s}} + \underbrace{\left(f(\vec{x}) - h^*(\vec{x})\right)}_{\text{constant in } \Delta\text{'s}} + \Delta$$

$$E\left[(Y - g(\vec{x}))^2 \,|\, \vec{X}\right] = E\left[\left(h^*(\vec{x}) - g(\vec{x}) + (f(\vec{x}) - h^*(\vec{x})) + \Delta\right)^2\right]$$

$$MSE = E\left[\underbrace{(h^* - g)^2}_{V} + \underbrace{(f - h^*)^2}_{B} + \Delta^2 + 2\underbrace{(h^* - g)}_{V}\Delta + 2\underbrace{(f - h^*)}_{B}\Delta + 2\underbrace{(h^* - g)}_{V}\underbrace{(f - h^*)}_{B}\right]$$

$$= E[V^2] + B^2 + \sigma^2 + 2B\,E[V] = Var[V] + B^2 + \sigma^2 + 2B\,E(V) + E(V)^2$$

If algorithm is "unbiased" in its estimate of g then $E(V) = 0$. E.g. OLS is unbiased. (Proof in ECON 387.)

If unbiased...

this should read Bias[g]^2

$$MSE = E_{\mathcal{X}}\left[Var\left[g - h^*\right]\right] + E_{\mathcal{X}}\left[Bias\left[h^*\right]^2\right] + \sigma^2$$

this should read Var[g] = E[(g - h*)^2]

estimation error component	misspecification error component	irreducible error component

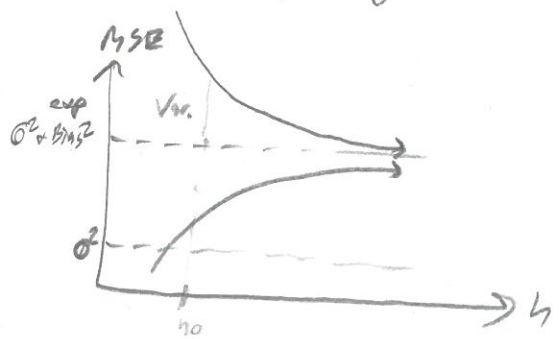
if not, there's a fourth component

$\Rightarrow$ the bias-var decomp. you know about the whole semester!

if $h \to \infty$ $g \to h^*$ You can prove this for many alg's

$\Rightarrow$ exp. est. error $\Rightarrow 0$

but exp. bias is always the same!

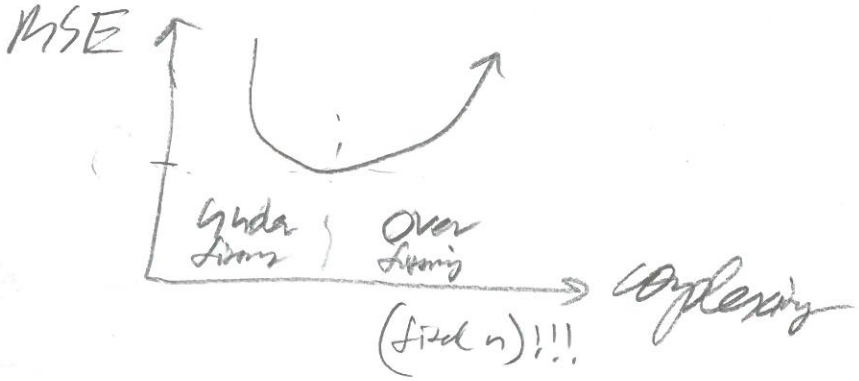


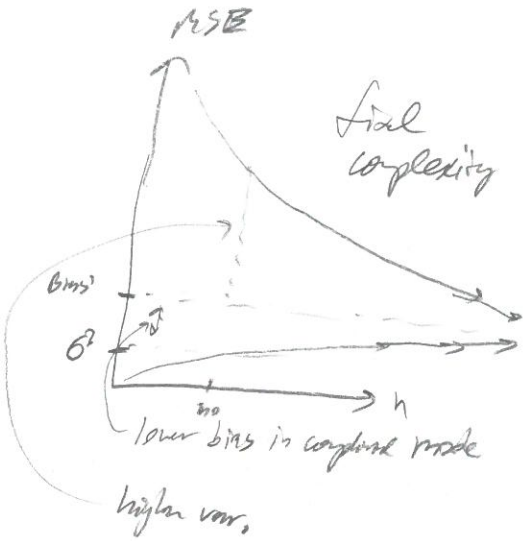
How to minimize bias? Make $h^* \to f$ by increasing model complexity!

Is there a tradeoff? Yes & No.

the tradeoff is between underfitting (high bias / low var) and overfitting (low bias / high var)

bias perfect fitting minimize both





Back to trees ...

regression trees have.. low bias & high var

have the ability to get close to unbiased f

they overfit!

demo

A couple of observations

1. Many models averaged together is very stable!
   We can consider the average model as a model itself!

$$g_{avg.} = \frac{g_1 + g_2 + \ldots + g_m}{m}$$

(2) The avg of many trees is very low in bias

Can we average many trees together?

But with one $\mathcal{D}$ $\Rightarrow$ one tree! It's deterministic! So no...

Unless... we change $\mathcal{D}$ or change A to be random.

What if we change $\mathcal{D}$? Let's sample N rows from $\mathcal{D}$ with replacement. This is called a "Bootstrap Sample". This $\mathcal{D}_{(1)}$ is $\approx \mathcal{D}$ but a little bit different. It has about $\frac{2}{3}$ the original rows and duplicates of those...

So lets fit $g_1 = A(\mathcal{H}, \mathcal{D}_{(1)})$ (regression tree)

Lets then draw another bootstrap sample $\mathcal{D}_{(2)}$ and build another tree:

$$g_2 = A(\mathcal{H}, \mathcal{D}_{(2)}), \text{ etc...}$$

$g_1, \ldots, g_B$ are all regression trees which are similar. Then...

$$g_{avg} = \frac{g_1 + \ldots + g_T}{T}$$ is the "average" or "aggregate" tree

Bootstrap + Aggregation = "Bagging" (Breiman, 1994)

"Bean-Algorithm" which leads to "Improvements for unstable procedures"

HUGE!

high variance algorithms

demo

Here's going on. Imagine a tree g_t. It has low bias & high variance.

T Many trees averaged $g_{bagged} = \frac{g_1 + \dots + g_T}{T}$

also have low bias since

$$\text{Bias} := E[g_{bagg}] - f = E\left[\frac{g_1 + \dots + g_T}{T}\right] - f$$

$$= \frac{1}{T} E[g_1 + \dots + g_T] - \frac{Tf}{T} = \frac{1}{T}\left(E[g_1 - f + g_2 - f + \dots + g_T - f]\right)$$

$$= \frac{1}{T}\left(\underbrace{(E[g_1] - f)}_{\substack{\text{Bias of} \\ 1^{st} \text{ tree} \\ (\text{small})}} + \underbrace{(E[g_2] - f)}_{\substack{\text{Bias of} \\ 2^{nd} \text{ tree} \\ (\text{small})}} + \dots + \underbrace{(E[g_T] - f)}_{\substack{\text{Bias of} \\ \text{last tree} \\ (\text{small})}}\right) = \frac{T}{T}(\text{small})$$

But the variance term looks like:

$$Var(g_{bagged}) = Var\left[\frac{g_1 + \dots + g_T}{T}\right] = \frac{1}{T^2} Var\left[\sum_{t=1}^{T} g_t\right]$$

if $g_1, \dots, g_T$ independent

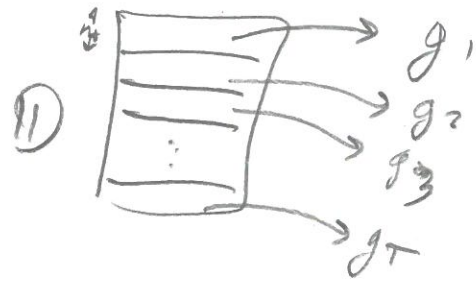
$$= \frac{1}{T^2} \sum_{\text{all trees}} Var[g_i] = \frac{Var[g_i]}{T}$$

which means as $T \to \infty$, the variance term vanishes!

$$\Rightarrow MSE = \sigma^2 + \underbrace{\frac{Var(g_i)}{T}}_{\text{goes to zero}} + \underbrace{\text{Bias}}_{\text{small}} \approx \sigma^2 \quad \text{PERFECT!!!}$$

What's the problem?

$g_1, \ldots, g_T$ not independent... Why? Usually from same data (bootstrap samples). But if they are from different data, this works!



As long as n_t large enough to create an unbiased model then your MSE will be very small.

Usually, you use all ① to build g bagged... in which case the trees are dependent because bootstrap samples contain a lot of the same data!

Back to March 24!

If $X_1, \ldots, X_n$ identically distr but not independent, what is $\mathrm{Var}[\bar{X}_n]$?

$$\mathrm{Var}[\bar{X}_n] = \mathrm{Var}\left[\frac{1}{n}\sum X_i\right] = \frac{1}{n^2}\left(\mathrm{Var}[X_1] + \ldots + \mathrm{Var}[X_n] + \sum_{i\neq j} \mathrm{Cov}[X_i, X_j]\right)$$

Why? ↑ assume all terms the same and positive!!

$$= \frac{1}{n^2}\left(n\sigma^2 + (n^2-n)\,\sigma_{ij}\right)$$

$$= \frac{1}{n}\left(\sigma^2 + (n-1)\,\sigma_{ij}\right)$$

this is the correct but let's get it into correlation form

Note $\rho_{ij} := \frac{\sigma_{ij}}{\sigma_i \sigma_j} = \frac{\sigma_{ij}}{\sigma^2} \Rightarrow \sigma_{ij} = \rho\sigma^2$

↑↑ same

$$= \frac{1}{n}\left(\sigma^2 + (n-1)\rho\sigma^2\right)$$

$$= \frac{1}{n}\left(n\rho\sigma^2 + \sigma^2 - \rho\sigma^2\right) = \rho\sigma^2 + \frac{1-\rho}{n}\sigma^2$$

Note $\rho \to 0$ $\mathrm{Var}[\bar{X}] \to \frac{\sigma^2}{n}$ as expected

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

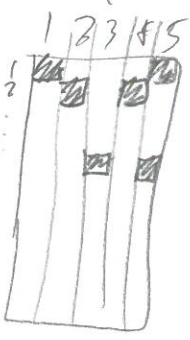
$\ll Var[g_0]$ itself since $\rho \neq 1$

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.



Validate 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 compare e_i's and do an error metric e.g. R^2_{oos}, $s_{e,oos}$, etc.

⇒ Out of bag error est (oob error). It works! Empyc does it. Theoretically it's ≈ K-fold CV with $K=2$. But I don't know much about it.