Lec 22 5/3/17 Math 390.4 1

Bias-Var Decomp

One $(\mathbb{X}, \vec{y})$, one $\vec{x}^*$: $MSE(\vec{x}^*) = \sigma^2 + \underbrace{\text{Bias}[g(\vec{x}^*)]^2}_{(g(\vec{x}^*) - f(\vec{x}^*))^2}$

One $\mathbb{X}$, many $\vec{y}$, one $\vec{x}^*$: $MSE(\vec{x}^*) = \sigma^2 + \underbrace{\text{Bias}[g(\vec{x}^*)]^2}_{(E[g(\vec{x}^*)] - f(\vec{x}^*))^2} + \underbrace{\text{Var}[g(\vec{x}^*)]}_{\text{dartboard-dance}}$

Most general situation:

Many $\mathbb{X}$, many $\vec{y}$, many $\vec{x}^*$: $MSE = \sigma^2 + E_X[\text{Bias}[g(\vec{x})]^2] + E_X[\text{Var}[g(\vec{x})]]$

How to make g better. ("Bagging", 1996)

Imagine many models $g_1, \ldots, g_M$ averaged together:

$$g_{avg} := \frac{g_1 + \ldots + g_M}{M}$$ What is MSE?

$$MSE = \sigma^2 + E_X[\text{Bias}[g_{avg}]^2] + E_X[\text{Var}[g_{avg}]]$$

$$= \sigma^2 + E_X\, E\left[\frac{g_1 + \ldots + g_M}{M} - f\right]^2 + E_X \text{Var}\left[\frac{g_1 + \ldots + g_M}{M}\right]$$

$$= \sigma^2 + E_X\, E\left[\frac{1}{M}\left(\underbrace{(g_1 - f) + (g_2 - f) + \ldots + (g_M - f)}_{\text{Bias}[g_1] + \text{Bias}[g_2] + \ldots + \text{Bias}[g_M]}\right)\right]^2 + E_X \frac{1}{M^2} \text{Var}[g_1 + \ldots + g_M]$$

Assume (I) Bias of all $g_1, \ldots, g_M$ same and (II) $g_1, \ldots, g_M$ independent

$$= \sigma^2 + E_X \text{Bias}[g_1]^2 + E_X\left[\frac{\text{Var}[g_1]}{M}\right] = \sigma^2 + E_X \text{Bias}[g_1]^2 = \sigma^2$$

If $M \to \infty$

If Bias is low i.e. overfit model

Can we do this? A little...

(I) It is easy to drive bias → 0. You just overfit. It will find f perfectly on average. But you introduce large variance.

This variance can be driven to zero if ... $g_1 \dots g_m$ indep.

(II) How to get $g_1, \dots, g_m$ independent given one dataset D? You can't!

Enter ~~Breiman~~ Leo Breiman in 1998...

Imagine sampling $\mathbb{D}$ with replacement of size n.

$\mathbb{D}_{(1)} = \text{sample}(\mathbb{D})$ (non-parametric bootstrap sample) → we will explain the meaning of this term shortly.

$\mathbb{D}_{(1)}$ has about $\frac{2}{3}$ of the rows of $\mathbb{D}$ and $\frac{1}{3}$ duplicates

then do it again ... and again

$\mathbb{D}_{(1)}, \mathbb{D}_{(2)}, \ldots, \mathbb{D}_{(M)}$

Each $\mathbb{D}_{(b)}$ is a little bit different from the others since it has slightly different data.

Now build a model on each bootstrap sample

$$g_b = \mathcal{A}_{\mathcal{H}}(\mathbb{D}_{(b)}) \qquad b = 1, \ldots, M$$

and average ..

$$g_{avg} = \frac{1}{M} \sum_{b=1}^{M} g_b$$

done in regression via averaging

this is called bootstrap + aggregation = "bagging"

It is a "meta-algorithm" as in its something done on top of an $\mathcal{A}$ and can yield results for any $\mathcal{A}$.

Is $Var[g_{avg}] = \frac{Var(g_i)}{m}$ s.t. $m \to \infty$ $Var[g_{avg}] \to 0$?

No! $g_1, \dots, g_m$ are not independent

share a lot of the same data, therefore $g(1), g(2)$ will be similar

$D_{(1)}$ $D_{(2)}$

How similar?

Math 281 exercise...

$$Var[\bar{X}_n] = Var\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n^2}\left(Var(X_1) + Var(X_2) + \dots + Var(X_n) + \sum_{i \neq j} Cov[X_i, X_j]\right)$$

$\sigma^2 = Var(X_i)$ same for all i

Assume $\sigma_{ij} = Cov(X_i, X_j)$ is the same for all i,j

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

let $\rho = \frac{\sigma_{ij}}{\sigma_i \sigma_j} = \frac{\sigma_{ij}}{\sigma^2}$ (correlation)

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

$$= \rho\sigma^2 + \frac{1-\rho}{n}\sigma^2$$

If $\rho \to 0 \Rightarrow \sigma_{ij} \to 0$ and all X_i's indep $\Rightarrow Var[\bar{X}] \to \frac{\sigma^2}{n}$

otherwise $Var[\bar{X}] > \frac{\sigma^2}{n}$

In expectation, each $g(b)$ has same corr. with another $g(b)$.

$$\Rightarrow MSE = \sigma^2 + E_x\left[Bias[g]^2\right] + E_x\Big[\overbrace{\rho Var[g] + \frac{1-\rho}{M} Var(g)}^{\text{Variance term}}\Big]$$

Now with $M \to \infty$ bootstrap samples...

$$MSE = \sigma^2 + E_x\left[Bias(g)^2\right] + E_x\left[\rho Var(g)\right]$$

If $\rho < 1$, the g_{avg} does better than a single g

And $\rho < 1$ for bootstrapping $\Rightarrow$ MAGIC!

Bootstrap: raise yourself up by pulling on your bootstrap. Get something for nothing!!

Validation for models that implement bagging

Usually $\mathbb{D} = \mathbb{D}_{train} \cup \mathbb{D}_{test}$

Here, you can imagine

$$\mathbb{D} = \mathbb{D}_{(1)} \cup \underbrace{\mathbb{D} \setminus \mathbb{D}_{(1)}}_{\mathbb{D}_{test,(1)}}$$

$\frac{2}{3}$ of train $\mathbb{D}$, ← the $\frac{1}{3}$ left out! "Out of bag" (OOB)

$$\mathbb{D} = \mathbb{D}_{(2)} \cup \mathbb{D}^{test}_{(2)}$$

$\vdots$

etc.

there are n rows and $M >> n$ models built on bootstrap samples where each row has a $\frac{1}{3}$ chance of being left out of each model

$\Rightarrow$ each row has $\approx \frac{M}{3}$ models that are not built with the row,

Thus, we can predict on these $\frac{M}{3}$ models to get an "OOB" estimate. with M large, no problem to get est's. It's similar to leaving out the datum.

Supposedly, theoretically oob is $\approx$ CV with $K=2$.

Let's say we bag trees s.t. nodesize = small. and M = large

$\Rightarrow \text{Bias}[g] \approx 0$

$\Rightarrow MSE \approx \sigma^2 + E_X[\rho \text{Var}[g]]$

Advantages of Bagged trees

① Low MSE Due to low bias since trees are complex enough to find f.

② No need to specify a model since trees are non-parametric

③ Validation for free!

DEMO

How can we do better? Make ρ as small as possible!

How can we "decorrelate" the trees?

Modify the regression tree algorithm. Instead of splitting by trying all p features use $\{j_1, j_2, \ldots, j_{p_{try}}\} \subset \{1, 2, \ldots, p\}$

Sample at every split

This will bring $\rho \downarrow$ but increase $Bias[g]$ but not too much!

Each tree is much more "random" and there are lots of trees

$\Rightarrow$ a "random forest"! (Breiman, 2001)

Concludes the material needed for your project