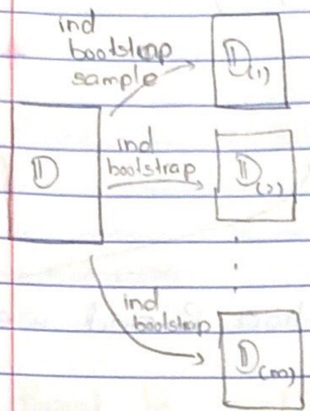
Lecture 24 05/05/2020




this has $\approx 2/3$ obs. of $\mathbb{D}$, $\approx$ left

We fit M models using same algorithm;
$g_{(1)} = \mathcal{A}(\mathbb{D}_{(1)}, \mathcal{H})$, $g_{(2)} = \mathcal{A}(\mathbb{D}_{(2)}, \mathcal{H})$, ..., $g_{(m)} = \mathcal{A}(\mathbb{D}_{(m)}, \mathcal{H})$

We all of these models $g_{BAG} = \dfrac{g_{(1)} + g_{(2)} + \dots + g_{(m)}}{M}$

$$MSE = \sigma^2 + E_x\left[Bias\left[g_{BAG}\right]^2\right] + E_x\left[Var\left[g_{BAG}\right]\right]$$

if H sufficiently complex relative to f e.g. trees.

$$\approx \sigma^2 + E_x\left[Var\left[g_{BAG}\right]\right]$$

M large $\quad \rho \in [0,1]$

$$\approx \sigma^2 + \rho E_x\left[Var\left[g_{BAG}\right]\right]$$

---

Another benefit of bagging: Validation for free using "Out-of-bag (oob) Validation."

$\mathbb{D}_{test(1)}$ $\quad \mathbb{D}_{train(1)}$

$\mathbb{D}_{oob(1)} := \mathbb{D} / \mathbb{D}_{(1)}$ is a set of above $1/3$ n. Thus $g_{(1)}(\mathbb{D}_{oob(1)})$ will give honest prediction.

$\mathbb{D}_{oob(2)} := \mathbb{D} / \mathbb{D}_{(2)}$ is a different set of $\approx 1/3$ n. Thus $g_{(2)}(\mathbb{D}_{oob(2)})$ is honest.

⋮

$\mathbb{D}_{oob(m)} := \mathbb{D} / \mathbb{D}_{(m)}$ " " " "

How do we get validation for $g_{BAG}$?

$$\hat{y}_{i,oob} = \text{Avg}(\text{only } g_{(m)} \text{ prediction where } i \text{ is oob})$$

Each obs is oob $\approx \frac{1}{3} M$. Since $M$ large, each $\hat{y}_{i,oob}$ will be accurate.

OOB validation $\approx k = 2$ - fold CV

Advantage to bagging

1. Obliterates Bias if an $A$ wish complex $H$ is employed (e.g trees)
2. Reduce variance substamtially.
3. Free validation during the fitting step.

Assume we are using tree.

$MSE = \sigma^2 + \rho E_x [Var[g_{(m)}]]$. How can we make MSE smaller?

$\rho$ = Avg correlation between two trees each built with a different bootstrap sample.

How can we further de-correlate the tree during tree construction?

What if during each nodes constraction, you only split on a subset of features of size $P_{try} < p$. i.e. $\{j_1, j_2, \ldots, j_{P_{try}}\} \subset \{1, 2, \ldots, p\}$?

This models make the trees more different, hence $\rho$ decrease. Amazingly this doesn't increase bias too much. Random Forests, RF (Breiman, 2001)