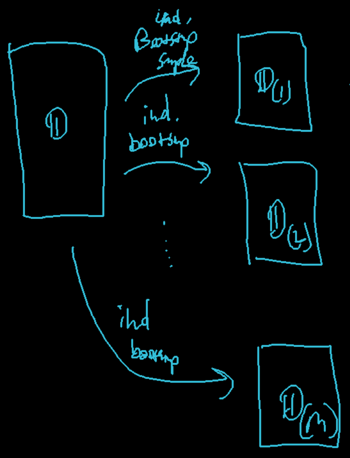

this has $\approx \frac{2}{3}$ obs. of $\mathbb{D}$, $\approx \frac{1}{3}$ left out

" " " " " " " " " " " " but different obs.

" " " " " " " " " " " "

We fit M models using same algorithm:

$$g_{(1)} = \mathcal{A}(\mathbb{D}_{(1)}, \mathcal{H}),\ g_{(2)} = \mathcal{A}(\mathbb{D}_{(2)}, \mathcal{H}), \ldots, g_{(M)} = \mathcal{A}(\mathbb{D}_{(M)}, \mathcal{H})$$

we average all of these models $g_{BAG} := \dfrac{g_{(1)} + g_{(2)} + \ldots + g_{(M)}}{M}$

if $\mathcal{H}$ sufficiently complex relative to f e.g. trees; M large; $\in [0,1]$

$$MSE = \sigma^2 + E_X\left[Bias[g_{BAG}]^2\right] + E_X\left[Var[g_{BAG}]\right] \approx \sigma^2 + E_X[Var[g_{BAG}]] \approx \sigma^2 + \rho E_X[Var[g_{(1)}]]$$

Another benefit of bagging: validation for free using "out-of-bag (oob) validation".

$\mathbb{D}_{oob(1)} := \mathbb{D} \setminus \mathbb{D}_{(1)}$ is a set of about $\frac{1}{3}n$. Thus $g_{(1)}(\mathbb{D}_{oob(1)})$ will give honest predictions.

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

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

⋮

$\mathbb{D}_{oob(M)} := \mathbb{D} \setminus \mathbb{D}_{(M)}$ - - - - - - - - - - - - - -

How do we get validation for g_{BAG}?

$$\hat{y}_{i,oob} := Avg\left(\text{only } g_{(m)} \text{ predictions where } i \text{ is oob}\right).$$

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

OOB validation supposedly $\approx K=2$-fold CV.

Advantages to bagging

1. Obliterates Bias if an $\mathcal{A}$ with complex $\mathcal{H}$ is employed (e.g. trees).
2. Reduces Variance substantially.
3. Free Validation during the fitting step.

F P

Assume we are using trees ....

$MSE = \sigma^2 + \rho E_X[Var[g_{(1)}]]$. How can we make MSE smaller?

ρ = Avg Correlation between two trees each built with a different bootstrap sample.

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

What if during each node's construction 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 would make the trees more different, hence ρ would decrease.

Amazingly this doesn't increase bias too much. Random Forests, RF (Breiman, 2001).