Back to Math 241/368... X_1, ..., X_n dependent rv's

$$\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$$

Assume $\sigma^2 = Var[X_i]$ same

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

$$Cov[X_i, X_j] = E[X_i X_j] - E[X_i]E[X_j] \overset{\text{assume}}{=} \sigma_{ij} \text{ (same)}$$

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

$$\rho := Corr[X_i, X_j] := \frac{Cov[X_i, X_j]}{SD[X_i]\,SD[X_j]} = \frac{\sigma_{ij}}{\sigma\sigma} \Rightarrow \sigma_{ij} = \sigma^2\rho \quad \in [-1, 1]$$

but in our case, $\in (0,1)$

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

Check: $\rho = 0$

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

if $\rho > 0$

Also $> \frac{\sigma^2}{n}$

Now let's apply this to the MSE decomposition formula for a model average g_avg where the constituent models are zero bias i.e. overfit:

$$MSE = \sigma^2 + E_X\left[Bias[g_1]^2\right] + E_X\left[Var[g_{avg}]\right]$$

$$= \sigma^2 + E_X\left[Var[g_{avg}]\right]$$

$$= \sigma^2 + E_X\left[\rho\, Var[g_1] + \frac{1-\rho}{M} Var[g_1]\right]$$

$M \to \infty$

$$= \sigma^2 + \rho\, E_X\left[Var[g_1]\right], \quad \rho \in (0,1)$$

$$< \sigma^2 + E_X\left[Var[g_1]\right]$$

which means we get a discount almost for free

"Raise oneself by their Bootstrap" means getting something for nothing. The slight cost is that you're model-averaging meaning you lose visibility into how your model is coming up with its predictions.

There is another bonus to this bagging algorithm: free validation.

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

In a bootstrap sample:

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

rows that are missing in the boostrap sample provides a natural Dtest for g1

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

$$\vdots$$

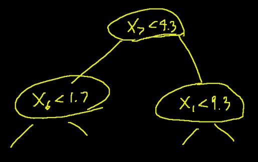
$$\mathbb{D} = \mathbb{D}_{(M)} \cup (\mathbb{D} \setminus \mathbb{D}_{(M)})$$

Since M is large, M >> n. Since each bootstrap sample has ~1/3 left out, each observation in D has ~ M/3 models that are built without seeing it. We can predict that observation on these M/3 models. Then we do this for all 1...n observations. These "out of sample" predictions y-hats are called "out of bag" (OOB). This procedure is called "boostrap validation". Theoretically, they say boostrap validation is approximately equivalent to K=2 cross fold validation.

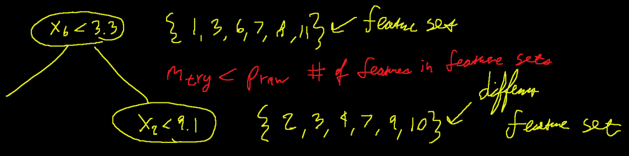
Why use trees in the bagging procedure? They have no bias. There is no need to specify transformations of the p_raw features. Then the bagging gives you a discount on the variance. And the bagging gives you validation without further work.

$$MSE = \sigma^2 + \underbrace{\rho}_{<1} E_X\left[Var[g_1]\right]$$

Leo Breiman hits the scene again 7 yr later. In 2001, he said "we can do better" i.e. we can make ρ even smaller. In the context of bagged trees, we can do the following during the individual tree construction:



In the original 1984 CART algorithm, at every split point, every feature 1...p is searched to find the best split point. What if at every split point, we take a random *subset* of the p features?



If you build trees like this on the boostrapped samples, you decorrelate the tree models even more! You get a further discount on the MSE via the variance term. And... bias doesn't suffer that much!

Default mtry for regression is floor(praw / 3). Default mtry for clsasificaiton is floor(sqrt(praw)). The hyperparameter mtry is definitely cross-validated over because it really matters.

How to name this procedure? There's lots of trees. They're all random... "Random Forests" (RF).