May 0

Bias - Variance Decomposition

(A) One fixed $\langle X, y \rangle$, one $\vec{x}^*$ r.v. $\qquad y = f(x) + \delta$

$$\Rightarrow MSE(\vec{x}^*) = \sigma^2 + \left(Bias[g(\vec{x}^*)]\right)^2$$

where $Bias(g(\vec{x}^*)) = (g(\vec{x}^*) - f(\vec{x}^*))$

(B) One fixed X, but multiple y (r.v. now), one $\vec{x}^*$

$$\Rightarrow MSE(\vec{x}^*) = \sigma^2 + Bias[g(\vec{x}^*)]^2 + Var[g(\vec{x}^*)]$$

where $Bias(g(\vec{x}^*)) = \underbrace{E[g(\vec{x}^*)]}_{\text{mean model}} - f(\vec{x}^*)$

$$Var[g(\vec{x}^*)] = E[g(\vec{x}^*)^2] - E[g(\vec{x}^*)]^2$$

(C) Multiple X, multiple y, multiple $\vec{x}^*$

$$\Rightarrow MSE = \sigma^2 + E_X[Bias[g(\vec{x}^*)]^2 + E_X[Var[g(\vec{x}^*)]]$$

General Error

Trees have no bias, but high variance as a model. (Average of trees converges to actual f as number of trees $\to \infty$)

May 1

Bagging intro

Bagging (Breiman, 1994)

Imagine many models $g_1, \ldots, g_n$ which are averaged

$$g_{avg} = \frac{g_1 + g_2 + \ldots + g_n}{M}$$

What is MSE?

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

↓ won't write E_x (but assume is there)

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

$$= \sigma^2 + E\left[\frac{(g_1 - f) + (g_2 - f) + \ldots + (g_n - f)}{M}\right]^2 + \frac{1}{M^2} Var\left[g_1 + \ldots + g_n\right]$$

$$= \sigma^2 + \frac{1}{M^2}\left(Bias(g_1) + \ldots + Bias(g_n)\right)^2 + \frac{1}{M^2}\left(Var(g_1) + \ldots + Var(g_n)\right)$$

If X, Y are indep

$Var[X+Y] = Var[X] + Var[Y]$

(I) Assume all biases are same,

$$= \sigma^2 + \frac{1}{M^2}\left(M \cdot Bias(g_n)\right)^2 + \frac{1}{M^2}\sum Var[g_m]$$

$$= \sigma^2 + Bias(g_n) + \frac{1}{M^2}\sum Var[g_m]$$

(II) $Bias[g_n] \approx 0$ (for each n)

i.e. $\mathcal{A}$ for g_n is overfit

$$\Rightarrow MSE = \sigma^2 + \frac{1}{M^2}\sum Var[g_m]$$

Assume

(III) Models $g_1, \ldots, g_m$ are independent and have same variance

This is a cheat: only have indep. models if all from different data

$$MSE = \sigma^2 + \frac{Var(g_m)}{M} \quad \text{(m's canceled)}$$

$$MSE = \sigma^2 + \frac{Var[g_m]}{M}$$

if all models use different data, then as $M \to \infty$, requires we have infinite data

(IV) Let $M \to \infty$

$MSE = \sigma^2$, the theoritical limit

fine, but this only works If infinite data

Can you do this with one $\mathbb{D}$ with n observations?

No, but you can do something close

Let $\mathbb{D}_{(1)}$ be a n-size sample with replacement of the observations in $\mathbb{D}$

This is a "non-parametric bootstrap sample"

This sample has $\frac{2}{3}$ of the rows of $\mathbb{D}$ and $\frac{1}{3}$ duplicates.

Let $\mathbb{D}_{(2)}$ be another independently drawn bootstrap sample from $\mathbb{D}$.

Let $\mathbb{D}_{(3)}$ be another.

Let $\mathbb{D}_{(4)}$ be another.

$\vdots$

Let $\mathbb{D}_{(M)}$ be another.

Let $g_1 = \mathcal{A}(\mathbb{D}_{(1)})$

$g_2 = \mathcal{A}(\mathbb{D}_{(2)})$

$\vdots$

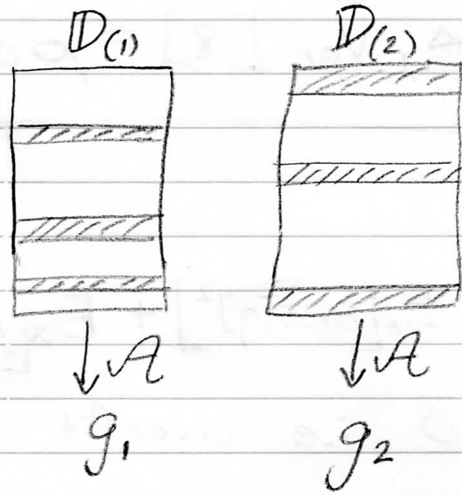
$g_n = \mathcal{A}(\mathbb{D}_{(M)})$

take average of the models (at every value of x)

and $g_{avg} = \dfrac{g_1 + \dots + g_m}{M}$

This is called a "bootstrap aggregator" or "bagging"

Doing this is "Meta-Algorithm"



Are g_1 and g_2 indep.?

No, but they are between fully indep. and fully dependent.

for arbitrary $X_1, \dots, X_n$ r.v.'s

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

Assume/Use

(I) $\sigma^2 = Var[X_i] \;\forall i$

(II) $\sigma_{ij} = Cov[X_i, X_j] \;\forall i \neq j$

let $\rho = \frac{\sigma_{ij}}{\sigma^2}$ (ρ is unitless and scaled $\in [-1, 1]$)

correlation

so

$$Var\left[\frac{X_1 + \dots + X_n}{n}\right] = \frac{1}{n^2}\left(n\sigma^2 + n(n-1)\sigma_{ij}\right)$$

cancel n

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

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

$$Var[\bar{X}] = \frac{1}{n}\left(\sigma^2 + n\sigma^2\rho\right)$$

$$Var[\bar{X}] = \frac{\sigma^2}{n} + \sigma^2\rho - \frac{\sigma^2\rho}{n}$$

$$Var[\bar{X}] = \rho\sigma^2 + \frac{1-\rho}{n}\sigma^2$$

If $n \to \infty$, $Var[\bar{X}] = \rho\sigma^2$

Use g_{avg}

$$MSE = \sigma^2 + E_x[Bias[g]^2] + E_x\left[\rho Var[g] \cdot \frac{1-\rho}{M} Var[g]\right]$$

(I) $Bias[g] \approx 0$ i.e. overfit

(II) $Var[g]$ is same for all models

(III) $M \to \infty$

get

$$MSE = \sigma^2 + E_x[\rho Var[g]]$$

$\rho < 1$, which is definately the case during bagging

compare

without bagging, would be

$$MSE = \sigma^2 + E_x[Var[g]]$$ (as $M \to \infty$)

Random Forests

by Breiman

At each iteration during regression tree algorithm, it tests every single $X_j < X_k$ rule at which there are $(n-1)p$

Breiman imagined a tree algorithm such that you only test $\{j_1, j_2, \ldots, j_{m+n}\} \subset \{1, 2, \ldots, p\}$ covariates randomly chosen at each iteration

only use some of the features to make each tree

(trees less accurate)

so... bias increases, but still negligible i.e. bias still ≈ 0

Thankfully, the trees are <u>less</u> correlated (better savings)

$$MSE = \sigma^2 + E(\rho \, Var[g])$$

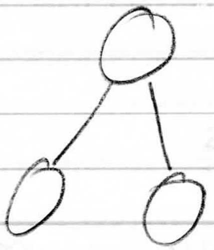
decreased this over just doing Bagging

less error overall!

Averaging these kinds of trees is called <u>Random Forests</u>

Regular Trees vs. Random Forest

→ check all possible splits for all features



$x_1 < x_{11}$

$\vdots$

$x_1 < x_{1(n-1)}$

$x_2 < x_{21}$

$\vdots$

$x_2 < x_{2(n-1)}$

$\vdots$

$x_n < x_{n1}$

$\vdots$

$x < x_{p(n-1)}$

all $p(n-1)$ possible rules

→ check all possible split for random selection of features

$p = 10$

use p_m only

$p_m = 5$

example of possible set of 5 features selected

is $\{1, 3, 4, 8, 10\}$

$\downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow$

$x_1 \quad x_3 \quad x_4 \quad x_8 \quad x_{10}$

used to make the trees

only $p_m(n-1)$ possible rules

more bagging ... validation

How do we validate g_1?

$$\mathbb{D} = \underbrace{\mathbb{D}_{(1)}}_{\frac{2}{3} \text{ of } \mathbb{D}} \cup \underbrace{(\mathbb{D} - \mathbb{D}_{(1)})}_{= \mathbb{D}_{test}}$$

about $\frac{1}{3}$ of $\mathbb{D}$

to validate prediction $\mathbb{D} - \mathbb{D}_{(1)}$

$\mathbb{D} - \mathbb{D}_{(1)}$ also written $\mathbb{D} \setminus \mathbb{D}_{(1)}$

How do we validate g_2?

$$\mathbb{D} = \underbrace{\mathbb{D}_{(2)}}_{\frac{2}{3} \text{ of } \mathbb{D}} \cup \underbrace{(\mathbb{D} - \mathbb{D}_{(2)})}_{= \mathbb{D}_{test}}$$

$= \mathbb{D}_{test}$ (about $\frac{1}{3}$ of $\mathbb{D}$)

use $\mathbb{D} - \mathbb{D}_{(2)}$ to validate predition

we use $\frac{1}{3}$ is actually more like $\frac{1}{e}$ "hat problem"

How to validate g_{avg}?

g_{bagged}

for observation 1, around $M/3$ models left out

for obs. 2, around $M/3$ models left out

If I predict for $\vec{x}_1$, on the models left out and average, I get oos $\hat{y}_1$

similarly for $\vec{x}_2, \vec{x}_3, \ldots, \vec{x}_n$

for each obs.

Models left out are called "out of bag" models

If I predict for $\vec{x}_1$ on these models and average, I get oos $\hat{y}_1$ ("oob" out of bag estimate)

If do bagging, get validation for free

⋮

If I predict for $\vec{x}_2$ on these models and average, I get oos $\hat{y}_2$

⋮ $\hat{y}_3$

" " " for $\vec{x}_n$ " ⋮ $\hat{y}_n$

doing this is like $K=2$ split

Average the residuals for all i
(each residual is $e_i = y_i - \hat{y}_i$)

Random forest does same thing, but with modified trees.