Lecture 23 04/30/2020

Data generating process (DGP)

$P(X, Y) = P(Y \mid X = x)\, P(X)$

matrix r.v; vertex r.v.; constant matrix

$\mathcal{D} = \left(\boxed{X}, \boxed{Y} \right)$

Imagine you got to see $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m$ iid from $P(X,Y)$ where m is large

Now you can fit $g_1 = A(\mathcal{D}_1), \ldots, g_m = A(\mathcal{D}_m)$; then you can use the avg:

$$g_{avg} = \frac{g_1 + \ldots + g_m}{m} \approx E[G] \qquad g_t = A(\mathcal{D}_t) \text{ are iid}$$

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

$$= \sigma^2 + E_x\left[E\left[\left(\frac{g_1 + \ldots + g_m}{m} - f\right)^2\right]\right] + E_x\left[Var\left[\frac{g_1 + \ldots + g_m}{m}\right]\right]$$

$$= \sigma^2 + E_x\left[\left(\tfrac{1}{m}\left(\underbrace{(E[g_1] - f)}_{Bias[g_1]} + \ldots + \underbrace{(E[g_m] - f)}_{Bias[g_m]}\right)\right)^2\right] + E_x\left[\tfrac{1}{m^2}\left(Var[g_1] + \ldots + Var[g_m]\right)\right]$$

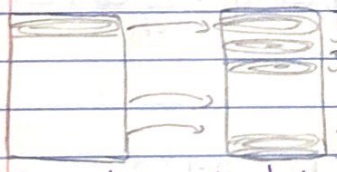
$$\overset{\text{if m is large}}{\approx} \sigma^2 + E_x[Bias[g_t]] + \tfrac{1}{m}\underbrace{E_x[Var[g_t]]}_{\text{constant}}$$

$$\overset{\text{Pick H very expressive}}{\approx} \sigma^2 + E_x[Bias[g_t]] \approx \sigma^2$$ theoretical minimum MSE.

Leo Breiman invented "bagging" (Bootstrap AGGregatING) in 1994, with one $\mathcal{D}$: What now?

Why not consider a sampling of the $\mathcal{D}$: $\mathcal{D}_{(1)}, \mathcal{D}_{(2)}, \ldots, \mathcal{D}_{(m)}$ when m is large

How to sample? Sample n rows/unit with replacement (Non-parameteric bootstrap sampling).



Original dataset — bootstrap sample

$\mathcal{D}_{(1)}$ has both duplicate rows of $\mathcal{D}$ and omissions.

P(learning row one) $= 1 - 1/n$

P(learning row n times) $= (1 - 1/n)^n \overset{n \text{ big}}{\approx} e^{-1} \approx 1/3$

$\Rightarrow$ In each bootstrap sample $\approx 1/3$ omissions, i.e. $2/3$ of the original (D).

Now, if $g_{(1)} = A(D_{(1)}), \ldots, g_{(m)} = A(D_m)$

$g_{(1)}, \ldots, g_{(m)}$ are not the same, but they will be dependent. $g_{avg} = \frac{g_{(1)} + \ldots + g_{(m)}}{m}$

Back to math 241,

Let $x_1, \ldots, x_m$ iid $\overset{\text{ident. distribution}}{\overset{\text{ind}}{=}} \frac{1}{n^2} \cdot \sum \text{Var}[x_i] = 0$. $\quad \text{Var}[\bar{x}] = \text{Var}\left[\frac{1}{n}\sum x_i\right] = \frac{1}{n^2} n\sigma^2 = \sigma^2/n$

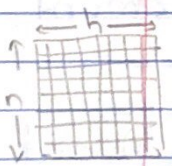
Let $x_1, \ldots, x_n$ be dependent but identically distributed.
Let $\sigma^2 = \text{Var}[x_i]$ and have the same dependance
Let $\rho = \text{Corr}[x_i, x_j]$ $i \neq j$

$$\text{Corr}[x_i, x_j] = \frac{\text{Cov}[x_i, x_j]}{SE[x_i]\,SE[x_j]}$$

$$\Rightarrow \rho = \frac{\text{Cov}[x_i, x_j]}{\sigma \cdot \sigma} \Rightarrow \text{Cov}[x_i, x_j] = \sigma^2\rho$$

$$\text{Var}[\bar{x}] = \frac{1}{n^2} \cdot \text{Var}[\textstyle\sum x_i]$$

$$= \frac{1}{n^2}\left(\text{Var}[x_1] + \ldots + \text{Var}[x_n] + \sum_{i \neq j} \text{Cov}[x_i, x_j]\right)$$



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

if n large, $\approx \sigma^2\rho$.

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

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

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

if $\rho \approx 0$ $\Rightarrow \approx \frac{\sigma^2}{n}$

If $g_{(i)}, g_{(j)}$ have same correlation ρ and we use g_{avg} then ....

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

Let $\mathcal{H}$ be very expressive $\Rightarrow Bias[g_{(i)}] = 0 \Rightarrow Bias[g_{avg}] = 0$

$$= \sigma^2 + E_x\left[Var\left[g_{avg}\right]\right] = \sigma^2 + E_x\left[\rho\, \underbrace{Var[g_{(i)}]}_{\text{constant.}} + \frac{1-\rho}{m} Var[g_{(x)}]\right]$$

if $m \to \infty$

$$\approx \sigma^2 + E_x\left[\rho\, Var[g_{(i)}]\right]$$

↑ gain from bagging.

Since $\mathcal{D}_{(i)}, \mathcal{D}_{(j)}$ share observations $\Rightarrow g_{(i)}, g_{(j)}$ will be positively correlated i.e $\rho \in (0,1)$

We want to minimize ρ as much as possible.