Ensembles: Random Forests
Boosting (XGBoost) } **Scikit learn**

Bias-Variance Decomposition:

Assume we want to estimate some population parameter θ, and we have a sample $x_1, \ldots, x_n$

Bias: $E(\hat{\theta}) - \theta$

Variance: $Var(\hat{\theta})$

concrete:

- $x_1, \ldots, x_n \sim N(\mu, 1)$
- $\theta = \mu$
- $\hat{\theta}_1 = \bar{x}$
- $\hat{\theta}_2$ = sample median

Bias: how far from the truth are we?

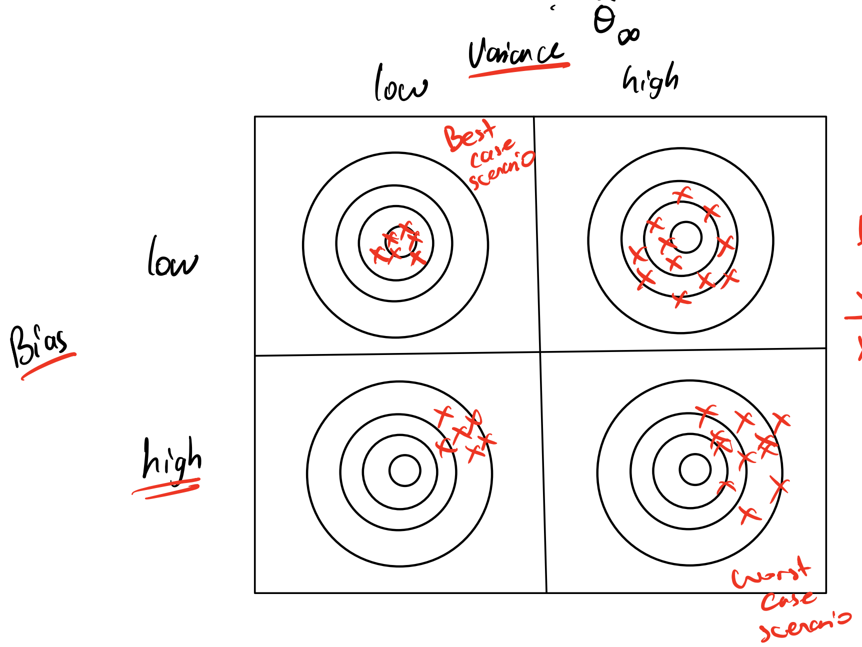
Variance: how noisy our estimator is.

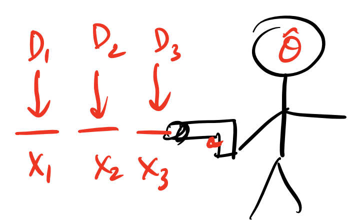
Variance over many datasets:

$D_1 = x_{11}, \ldots, x_n \longrightarrow \bar{x}_1 \quad \hat{\theta}_p$

$D_2 = x_{21}, \ldots, x_{2n} \longrightarrow x_2 \quad \hat{\theta}_2$

$\vdots \qquad \hat{\theta}_\infty$





Bias – Variance Decomposition:

- truth: $f(x)$
- estimate: $\hat{f}(x) = \hat{f}(x, D)$, $D = \{(x_i, y_i) ; i = 1, \cdots, n\}$

  $\frac{1}{n}\sum(x-y)^2$
- Expected : MSE on a test point x_0:

$$\underbrace{\mathbb{E}_{\boxed{D}}(f(x_0) - \hat{f}(x_0; D))^2}_{\text{expected MSE at } x_0} = \text{Bias}_{\boxed{D}}(\hat{f}(x; D))^2 + Var_D(\hat{f}(x_0; D)) + \sigma^2$$

$$= (E_D(\hat{f}(x_0; D) - f(x_0))^2 \text{ — Bias}^2$$

$$+ V_D(\hat{f}(x_0; D)) \text{ — Variance}$$

$$+ \sigma^2$$

σ^2 : irreducible error

reduce test MSE → reduce MSE }
reduce variance

Black Box
Data Generator

$D_1 = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ → $\hat{f}_1$ → NN (5 degrees)

D_2 → $\hat{f}_2$, D_3 → $\hat{f}_3$, . . . D_∞ → $\hat{f}_\infty$

$\mathbb{E}_D(\hat{f}) =$

$V_D(\hat{f}) =$

ML Pipeline:

→ 1. Get a Dataset
2. Build a model
3. Evaluate model
4. Use model in wild

Simulating Dataset

- true function : f

$D_1 = ?$

x_1 $y_1 = f(x_1) + \varepsilon_1$

x_2 $y_2 = f(x_2) + \varepsilon_2$

$\vdots$

x_n $y_n = f(x_n) + \varepsilon_n$

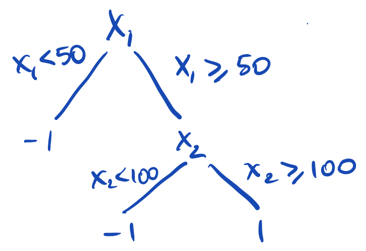
$\varepsilon_i \sim N(0, \sigma^2)$

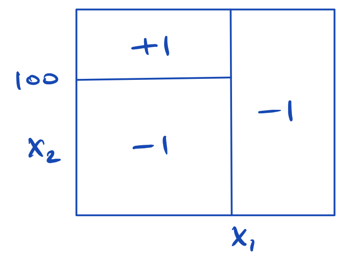
high Bias model: Strong assumptions shape of the fit is predetermined regardless of data.

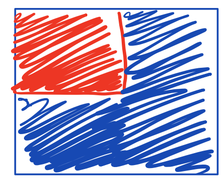
high Variance model: Error due to fitting randomness

Back to Ensembles $\longrightarrow$ how do ensembles approach B-V decomposition.

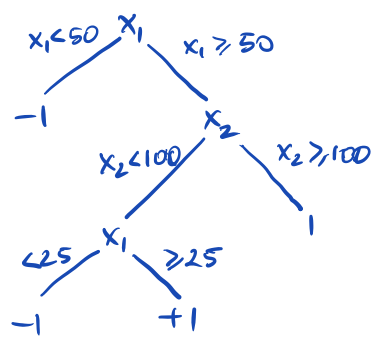
Decision Trees:

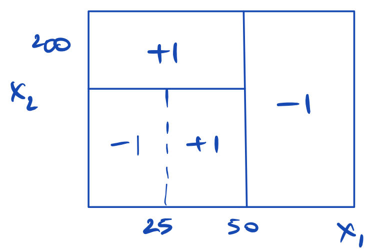


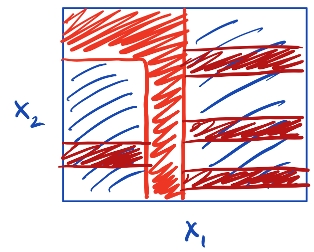


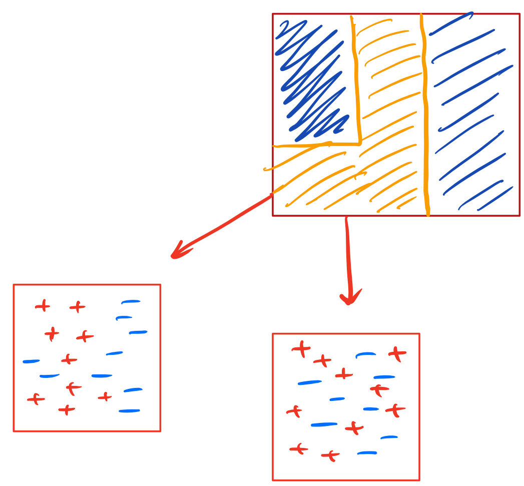


DT's: low bias
high variance

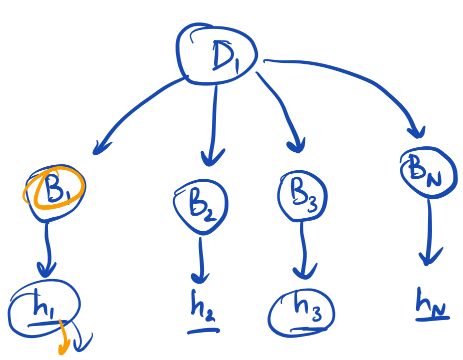








Bagging : Bootstrap Aggregation



$x_1 = (x_{11}, x_{12}, \dots, x_{1d})$

$\downarrow$

$\tilde{X}_{B_1} = \underbrace{(x_{11}, x_{13}, x_{16}, x_{1t})}$

B_i : Sample from D_1 with replacement

m points

where $m < n$

$D_1 = \{(x_1, y_1), \dots, (x_n, y_n)\}$

h_i = tree built on B_i

Ensemble $\rightarrow$ Bagging Classifier

new point $\quad C(x) = \begin{cases} +1 & \text{if } \frac{1}{n}\sum_{i=1}^{m/n} h_i(x) > 0 \\ -1 & \text{otherwise} \end{cases}$

majority vote

Average over many trees.

Why?

Randomisation induces independence

1. Randomise across samples
2. Randomise across features

(Random Forest)

Intuition: Ensembles : wisdom of crowds versus traditional building expert.

Why does model averaging work?

$Y_1, \dots, Y_n \overset{\text{independent}}{\sim} (\mu, \sigma^2)$

$E(Y_i) = \mu$

$V(Y_i) = \sigma^2$

$\text{bias}(\hat{f}) = E(\hat{f}) - f$

$= b$

$$E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

$$V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} Y_i\right) + 0$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2$$

$$= \frac{n\sigma^2}{n^2}$$

$$= \frac{\sigma^2}{n} < \sigma^2$$

so: average has the same mean and lower variance than individual values.

Average of N models with bias b is b, but the variance is lower.

Appeal of Ensembles:

- Base classifier (Decision Trees) is really simple/fast to train
- RF $\longrightarrow$ N DT's,
- Deep learning

Disadvantage:

- DT's are really simple to interpret
- RF's lose any easy interpretability