Previous Class (July 4)

Decision Trees Completed

Today's Class — Random Forests

0) JIO Dataset Problem
1) What are Ensemble Models & Types?
2) Ensemble from DTs → Random Forests (RFs)
3) Hyper-params of RF
4) Sklearn library
5) Hyper-parameter Tuning
6) How to add randomness
7) Quizzes

4 hyper-params

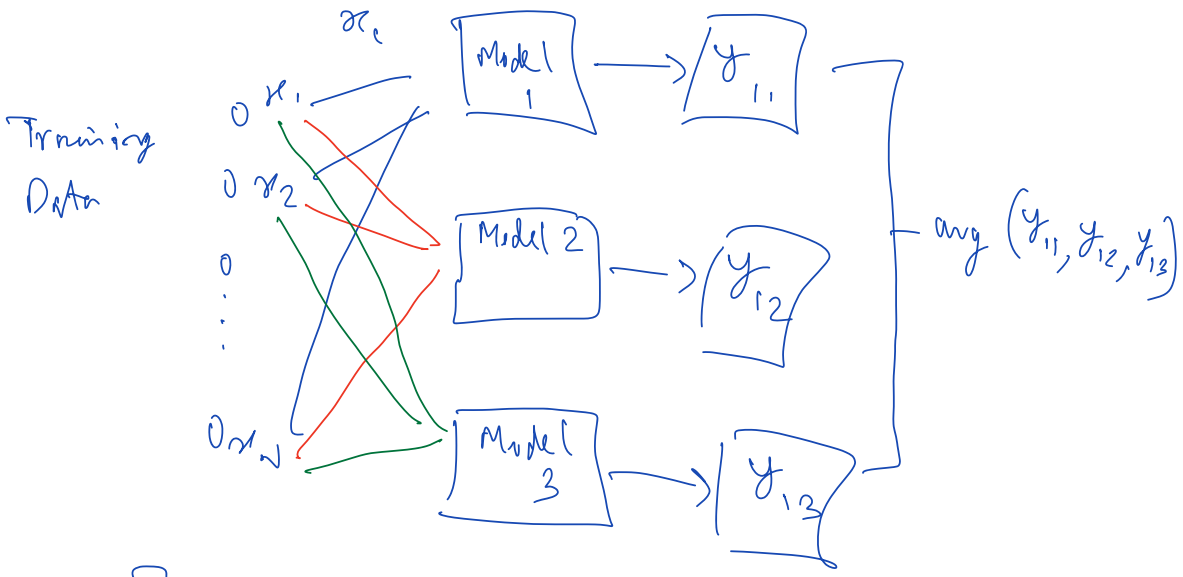
p_1, p_2, p_3, p_4

p_1: 10 values; p_2: 10 values; p_3: 10 values; p_4: 10 values

$10 \times 10 \times 10 \times 10$

p_1 p_2 p_3 p_4

10^4



$X \rightarrow [x_1, x_2, \ldots x_N]$

$\hookrightarrow$ DT $\rightarrow$ DT_1

DT_1 (DT_2) $\quad DT_1 \; != \; DT_2$

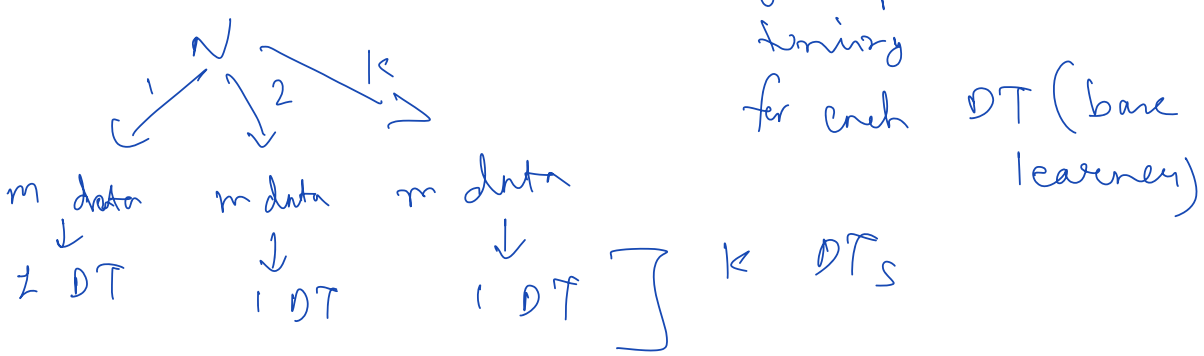
N total points

m samples

$\hookrightarrow$ m_1 $\rightarrow$ unique samples $\rightarrow$ training set

$(N - m_1) \rightarrow$ validation set

$\hookrightarrow$ hyperparameter tuning for each DT (base learners)



m data → DT

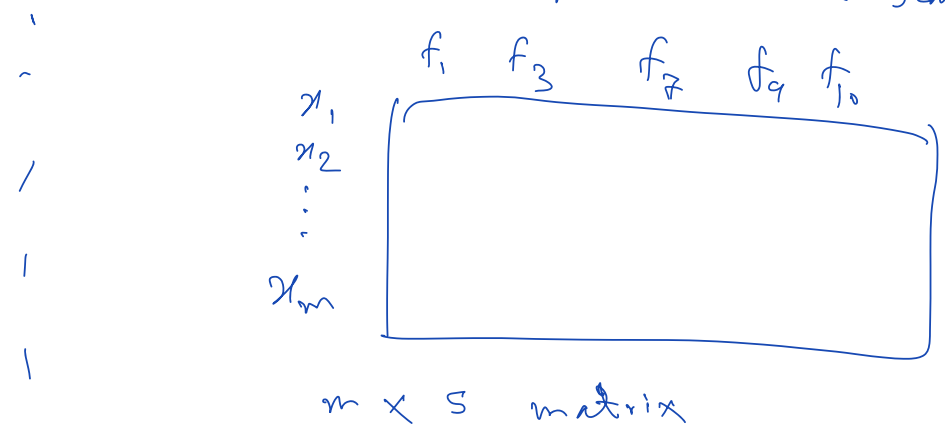
(N − m) samples → validation data

$N \times 10$ → Training data size

N → 100 , d → 10 | $f_1\ f_2\ f_3 \cdots f_{10}$

m → 50 , d′ → 5

Training set 1 → m samples with d′ features



m × 5 matrix

k times

Training set 2 → m × 5 matrix

3

⋮

k → m × 5 matrix

Overfit → DT 1

overfit → DT 2

⋮

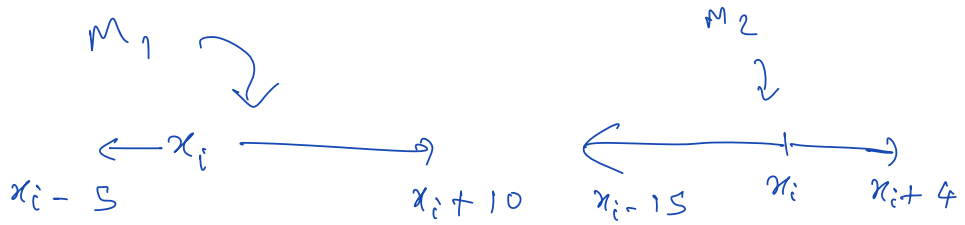
DT k

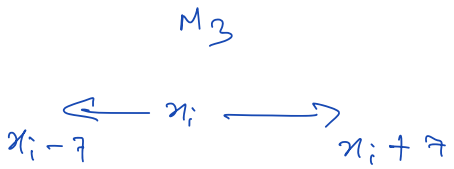
] avg →

Hyper-parameter Tuning

OOB samples → validation data for each DT

RF: Train Val





Error = Bias^2 + Variance + Irreducible Error

loss

LinReg → MSE

Logistic → log loss

min loss fn

min (loss fn + α # Leaf nodes)

$\alpha \uparrow$ LN = k

F = loss fn

$M_1 = F_1 + 0.1 \times 10 = F+1$ $\quad \alpha = 0.1$

$K = 10$

$0.2 \times 10 = F + 2$

$F_2 + 0.2 \times 5 = F + 1$

F_1 / 500 $\quad F_1 + 1.0 \times 500$

$= F_1 + 500$

$F_1 - 0.01$ / 1000 / $\alpha = 1.0$

$F_1 - 0.01 + 1 \times 1000$

$= F_1 + (1000 - 0.01)$

1 2 3 4 5

Horse Cat Dog Elephant Monkey

is_horse is_cat is_Dog is_Elephant

$(w_1 + w_2 + w_3 + w_4)$