Plan of Attack

Ensemble Learning techniques

multiple models

Bagging
├ Random forest ①

Boosting
├ Gradient Boost ②
├ xgboost ③

single model

not very difficult

f_1 f_2 f_3

f_1 ← 5

$f_3 > 10$ ⦿
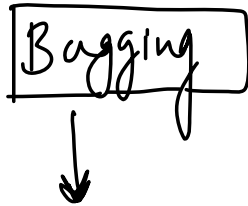
predict

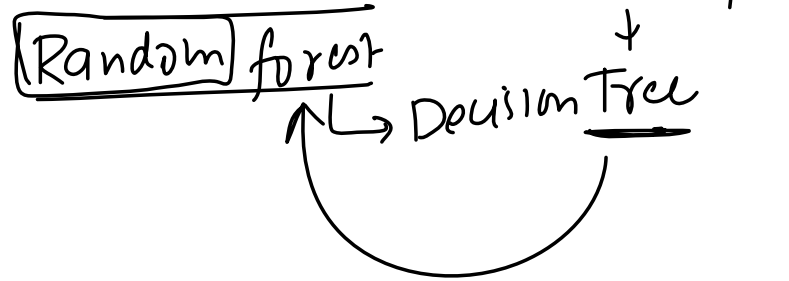| f_1 | f_2 | f_3 |

7 8 9

Introduction to Random Forest

29 July 2023   15:52

Random Forest is a versatile and widely used machine learning algorithm that belongs to the class of ensemble methods. Specifically, it is a type of bagging technique, which involves training many individual models (in this case, decision trees) and combining their outputs to make a final prediction.
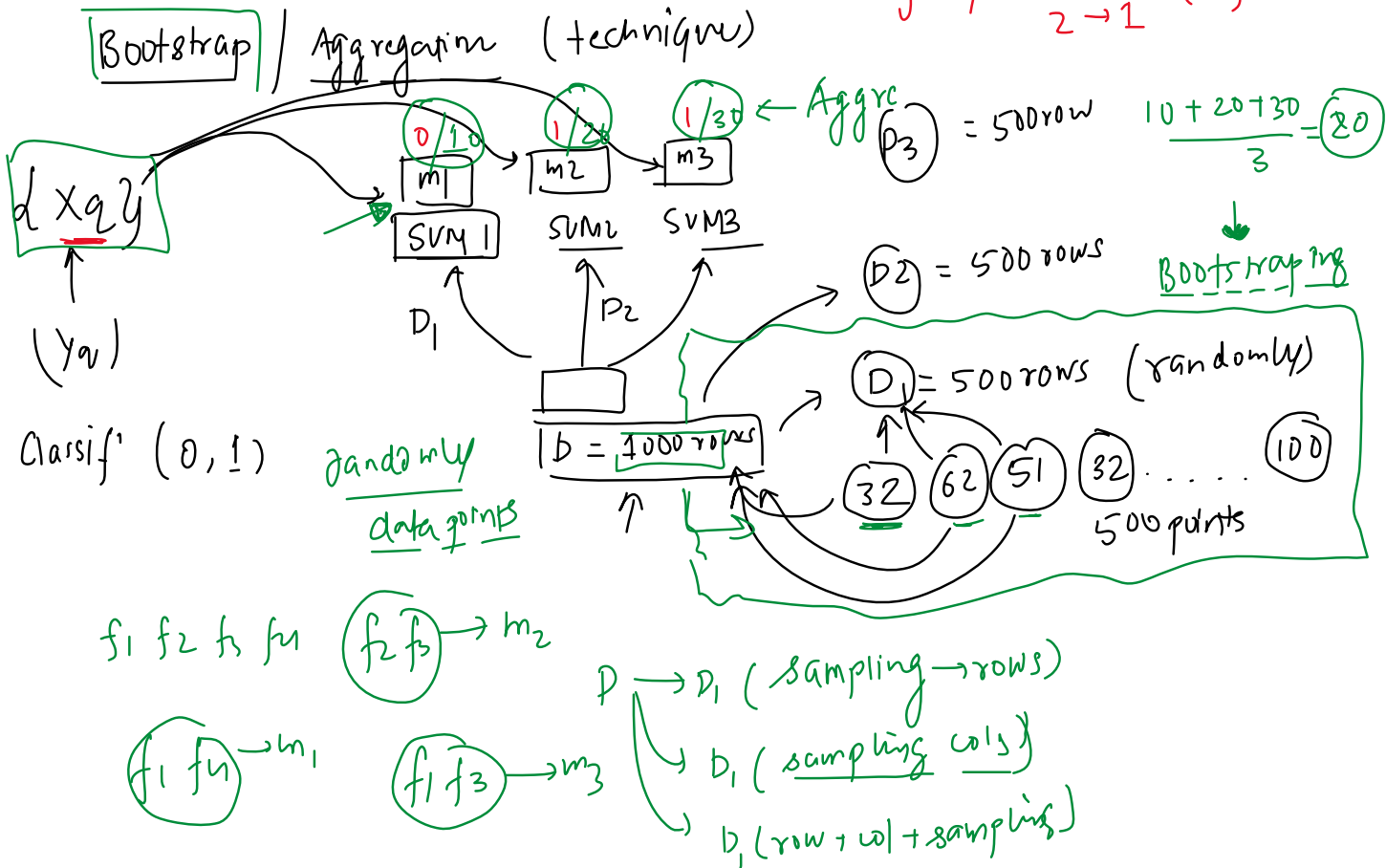
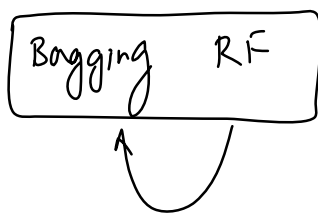Bagging

Random forest → Decision Tree

multiple
↓

Bagging

29 July 2023   15:52

Bagging, short for bootstrap aggregating, is a machine learning ensemble method designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also helps to avoid overfitting. The key principle of bagging is to generate multiple subsets of the original data (with replacement), train a separate model for each subset, and then combine the results.
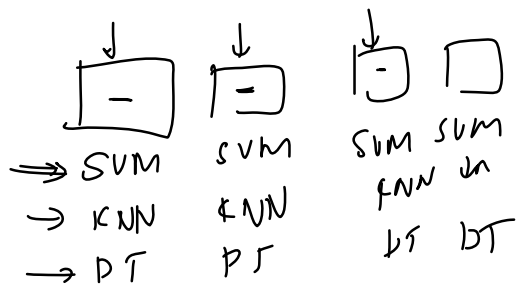
Bagging

majority $1 \to 0$ { 7 }
            $2 \to 1$

Bootstrap / Aggregation (technique)

$d \times q y$ (Yqv)

$m1$ $0/10$ $m2$ $1/20$ $m3$ $1/30$ ← Aggre

SVM 1   SVM2   SVM3 $P3$ = 500 row

D_1 P_2 $\frac{10 + 20 + 30}{3} = 20$

$D2$ = 500 rows   Bootstraping

Classif (0, 1)   randomly data points

D = 500 rows (randomly)

$b = 1000$ rows

32   62   51   32   100

500 points

$f_1 f_2 f_3 f_4$ $f_2 f_3 \to m_2$

$f_1 f_4 \to m_1$ $f_1 f_3 \to m_3$

$P \to D_1$ (Sampling → rows)
  D_1 (sampling cols)
  D_1 (row + col + sampling)

Random Forest Intuition [Code]

Bagging   RF

② major diff

1) Bagging → any ml algorithm

⟹ SVM   SVM   SVM SVM
→ KNN   KNN   KNN DT
→ DT   DT   DT   DT

RF → base model → $\boxed{DT}$

$D \to 1000\ rows/5 cols$ (classification) ⟶ rows = 500 row with replacem

RF (100 DT)

col sampling   row + col sampli

60 → 0   → 0
40 → 1

D1   D1 D_{100}

→ ☐   ☐ . . . . . ☐ X_q
  m1   m2   m100
   ↓ 1   ↓ 0   ↓ 1
   ↓

Wisdom of the crowd $\longrightarrow$

different background

RF $\longrightarrow$ Boolstc

D1   D2   D3   p4

$\downarrow$ $\downarrow$ $\downarrow$ $\downarrow$

m1   m2   m3   m4

entire

100

Decision tree

DT9

purple $\longrightarrow$ yellow

query

DT   50

yellow

Random forest

purple

avg
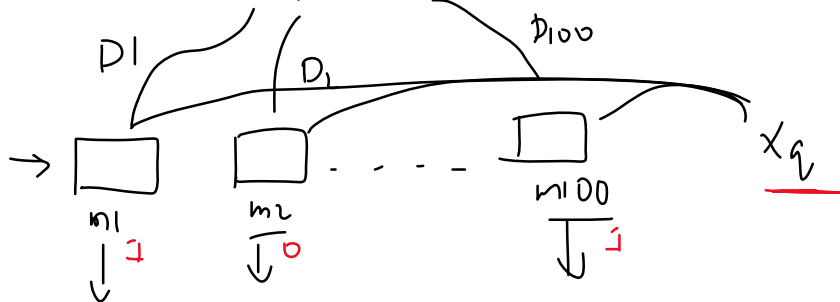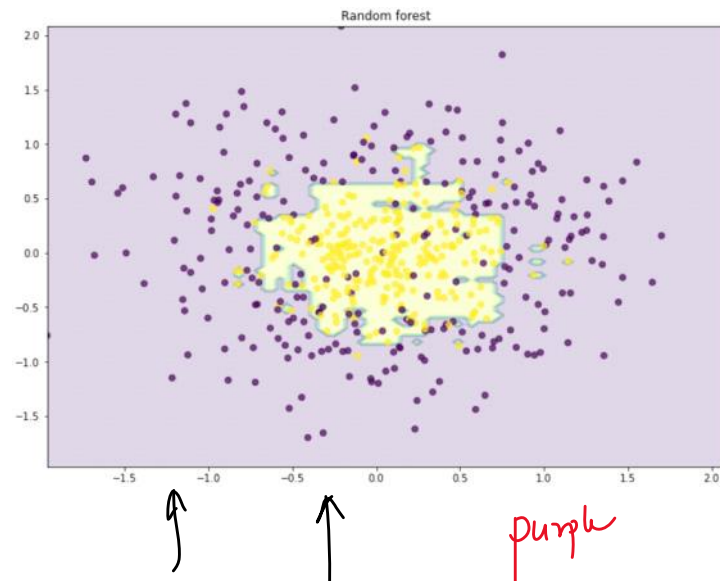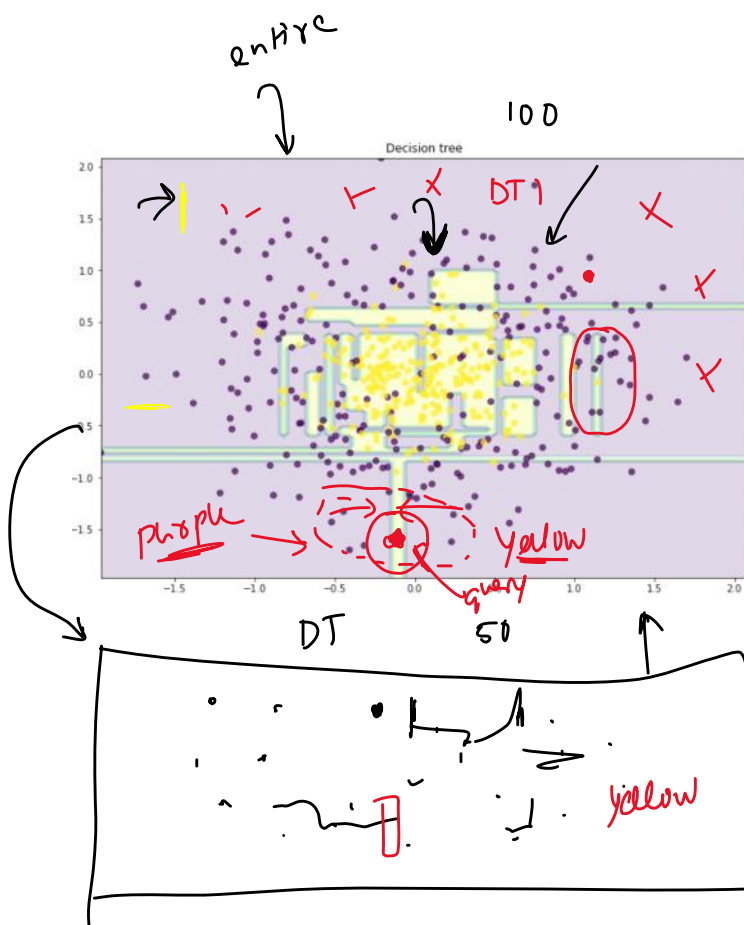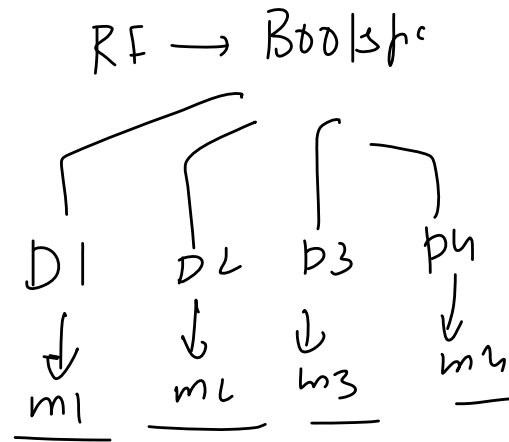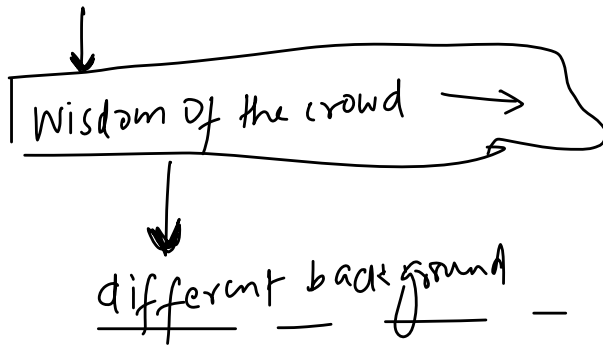
Bias - variance Tradeoff

$\downarrow$

Bias $\propto \dfrac{1}{variance}$
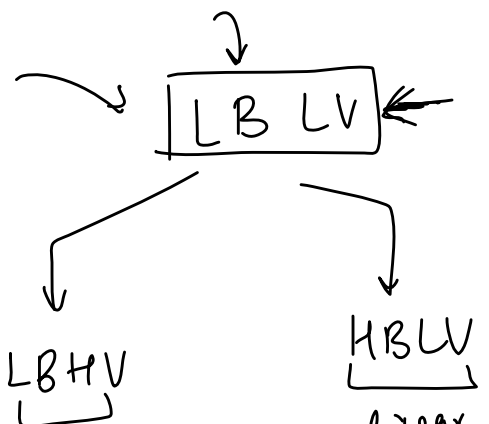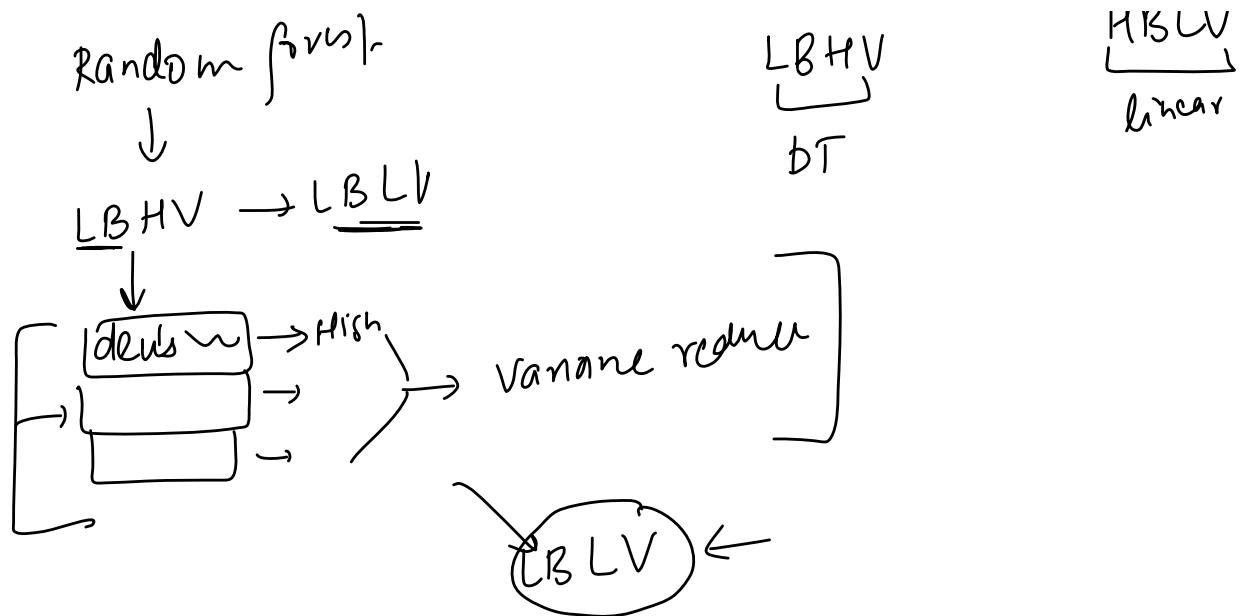
LB LV

LBHV   HBLV

Random forest

Random Forest

$\downarrow$

LB HV $\rightarrow$ $\underline{L\underline{B}\underline{LV}}$

$\downarrow$

High $\rightarrow$ Variance reduce

LB LV

LB HV
DT

HBLV
linear

Baggin (dt) $\rightarrow$ RF ?

Baggin $\downarrow$

SVM/FNN/DT

DT

NO $\rightarrow$ 2nd reason

col sampling   (0.5) $\rightarrow$ 50%

2.5 $\rightarrow$ (2)

col sampling (50%)

| f_1 | f_2 | f_3 | f_4 | f_5 |

Bagging

2 cols

tree level
col samp

$f_1 f_2$

baggins works

2 features

$(f_1 f_3)$ D_1 → m1

$f_1 f_5$ D_2 → m2

D

$f_2 f_3$ D_3 → m3

$f_4 f_5$ D_4 → m4

$f_1 f_3$ D_5 → m5
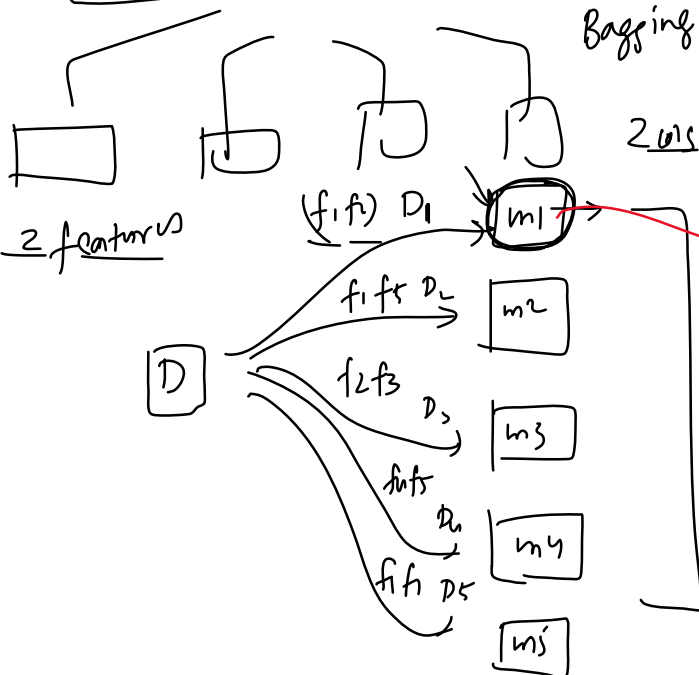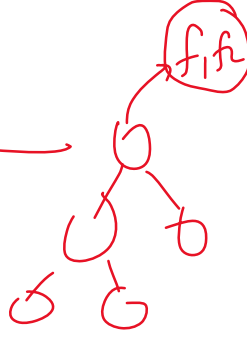
| f_1 | f_2 | f_3 | f_4 | f_5 | $\rightarrow$ col samp (0.5) $\rightarrow$ 2 cols |

node level $f_1 f_3$

$f_4 f_5$

$f_1 f_1$

D1

D2

D

D3

D4

D5

f_1
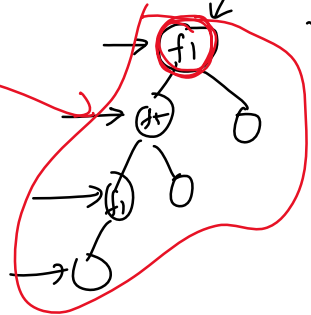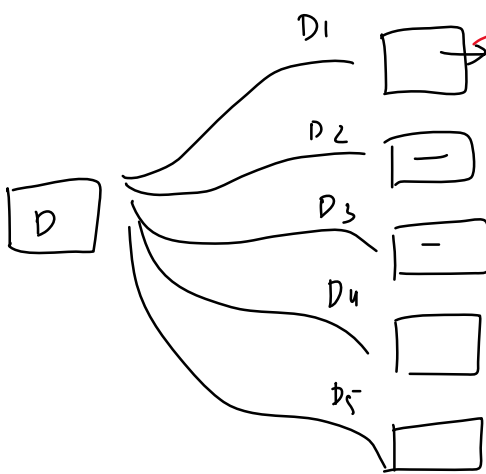
Bagging $\rightarrow$ RF $\rightarrow$ extra randomness

tree level   node level   D D D D

Feature Importance [Code]

$D \longrightarrow$ | f1 f2 f3 f4 f5 |

DT1   DT2   DT3

DT1 - f1 → 0.1
DT2 → f1 → 0.3
DT3 → f1 → $\dfrac{0.4}{\dfrac{0.8}{3}}$

R.F.

f1 → 0.2
f2 → 0.3
f3 → 0.25
f4 → 0.15
f5 → 0.10

→ feature selection

→ decision tree → f1 → imp'