zoom.us
Meeting
View
Edit
Window
Help
zoom
56%
Thu 28 Dec 6:17 PM
You are screen sharing
Stop Share
1/3

Collection of Base Models ← Ensemble Models

↳ A Collection of things

I/P → M1 M2 M3 → O/p.

Ensemble

1. Bagging
2. Boosting } → Homogeneous
3. Stacking — Heterogeneous

f1 f2 f3 . . . fd y

1
2
3
.
.
.
n = 1000

f1 f2 . fd

D_n

1. Bagging — Reduces Variance

↳ Base Models

High Variance + Low Bias

Bootstrapping — Sampling with replacement

Bootstrapped Aggregation -

Bagging

Bias & Variance

D_n → Bootstrapping

$n_1 = 200$ D_1 → M_1 → O_1 D — LBHV 35

$n_2 = 200$ D_2 → M_2 → O_2 C — LBHV 43

$n_3 = 300$ D_3 → M_3 → O_3 D — LBHV 32

.
.
.

D_m → M_m → O_m D — LBHV 41

Aggregation

Regression
Avg / Median
→
Majority Vote
classification

LBLV

zoom.us
Meeting
View
Edit
Window
Help
zoom
31%
Thu 28 Dec 7:16 PM
You are screen sharing
Stop Share
2 / 5

Bias → Training Error

↓ Bias → 0

Bias ′↑

Variance – Testing Error.

↑ The model itself changes with small changes in data

Variance ↓

Bias ↓ Low Variance ↓

x_1, x_2 x_3 y $\hat{y}$

1
2
3
4
5
6
7
8
9
10
→ Bias = 0
→ Bias = 0
→ Bias = 0
V

Bagging Reduces Variance → Base Models.

Low Bias & High Variance

Homogeneous Ensembles

10 ppl → Predicting House price.

1 → BTM ✓

2 → Indranagar

3 → HSR

4 → BTM

5 → Indranagar

100

100

70

80

+

-

90% →

90%

Bagging Regressor (M_1 —)

zoom.us
Meeting
View
Edit
Window
Help
46%
Thu 28 Dec 6:39 PM
You are screen sharing
Stop Share
3 / 3

Random Forest → Bagging Technique

1. All Base models are Decision Trees ✓

2. Row sampling + Feature Sampling

↓

↳ only a subset of Features.

$D_n \to D_{n1} = 200$

$f_1 f_2 f_3 \ldots f_d$

D_n

(1200, 3, 5, . .)

DT

$f_1 f_3 f_3$

D_{n1} → O_1 M_1 ✓ →

DT

$f_1 f_2 f_4$

D_{n2} → O_2 M_2 ←

DT

$f_1 f_3 f_5$

D_{n3} → O_3 M_3 →

DT

$f_1 f_2 f_3$

D_{n100} → M_{100} →

Aggregation → O/P

O_1 ✓

Age Sal Loc Cou

→ f_1 f_2 f_3 f_4 f_5 . . f_d y

1

2

3

4

DT

n = 80%.

ML ops →

→ Data Drift

Parallelize the model.

→ Quad core 25%.

25 ↓ C_1

25 ↓ C_2

↓ 25 C_3

↓ 25 C_4

model = RandomForest(n-estimators = 100)

model.fit(Xtr)

model.predict( )

zoom.us
Meeting
View
Edit
Window
Help
zoom
30%
Thu 28 Dec 7:17 PM
You are screen sharing
Stop Share
4 / 5

Only the first model predicts 'y'
All subsequent models predict the errors from the previous

Boosting.

- Reduces Bias ✓
- Sequential Model.
- Additive Model

f_1	f_2	f_3	y	$\hat{y}$	ε_0
x_i	$\leftarrow x_1 \rightarrow$		y_1		
x_2	$\leftarrow x_2 \rightarrow$		y_2		
x_3					
x_n	$\leftarrow x_n \rightarrow$		y_n		

Stage '0': M_0

$D_n \Rightarrow \{x_i, y_i\}_{i=1}^{n}$ $\quad \hat{y} = f_0(x)$ $\quad \varepsilon_0 = y - \hat{y}$

$= y - f_0(x)$

Stage 1: M_1

$\{x_i, \varepsilon_{0i}\}_{i=1}^{n}$ $\quad \hat{\varepsilon}_0 = f_1(x)$ $\quad \varepsilon_1 = \varepsilon_0 - \hat{\varepsilon}_0$

$= y - f_0(x) - f_1(x)$

Stage 2 M_2

$\{x_i, \varepsilon_{1i}\}_{i=1}^{n}$ $\quad \hat{\varepsilon}_1 = f_2(x)$ $\quad \varepsilon_2 = \varepsilon_1 - \hat{\varepsilon}_1$

$= y - f_0(x) - f_1(x) - f_2(x)$

$\Downarrow$

$y = f_0(x) + f_1(x) + f_2(x) \longrightarrow$ Additive.

Base Models

High Bias & Low Variance

LBLV.

ADABOOST ✓
XGBOOST ✓

M_estimators = ?
Each Tree →

Decision stump

Stump

Opera
File
Edit
View
History
Bookmarks
Developer
Window
Help
zoom
25%
Thu 28 Dec 7:28 PM
You are screen sharing
Stop Share
5 / 6

Stage 0 $f_0(x)$

Area	y	$\hat{y}$	ε_0
1200	35	22	13
2300	53	34	19
1700	47	38	9
3000	96	81	15
3500	94	82	12

Stage 1 $f_1(x)$

Area	ε_0	$\hat{\varepsilon}_0$	ε_1
1200	13	8	5
2300	19	10	9
1700	9	2	7
3000	15	6	9
3500	12	5	7

Stage 2 $f_2(x)$

Area	ε_1	$\hat{\varepsilon}_1$	ε_2
1200	5	1	4
2300	9	3	6
1700	7	2	5
3000	9	3	6
3500	7	4	3

$$
\begin{aligned}
y &= f_0(x) + f_1(x) + f_2(x) + \cdots \\
&= 22 + 8 + 1 + -0.5 \\
&= 31
\end{aligned}
$$

-1 -0.8

36

35.2

$\boxed{35}$ 34.95

M_2 31.8

↑ 0.8

M_2 31

↑ 1 ✓

M_1 30

↑ 8 ✓

M_0 22

linear Model

Classification

"Partition the feature space into pure regions belonging to each class"

'Linearly Separable' — linear surface

Opera File Edit View History Bookmarks Developer Window Help
zoom 23% Thu 28 Dec 7:32 PM
You are screen sharing
Stop Share
PL SL species
No
PL<4·5
Yes
Virgi
No
SL>3·5
Yes
Versicolor
Setosa
SL
Plot
3·5
Setosa
Virginica
Versicolor
4·5
PL
f2
Decision Boundary
f1
f3
6 / 6

'Linearly Separable'

Linear Models → Logistic Regression

Non Linear Decision Boundary