

PHƯƠNG PHÁP TẬP HỢP MÔ HÌNH

ENSEMBLE-BASED METHODS

TS. Trần Nguyễn Minh Thư

1

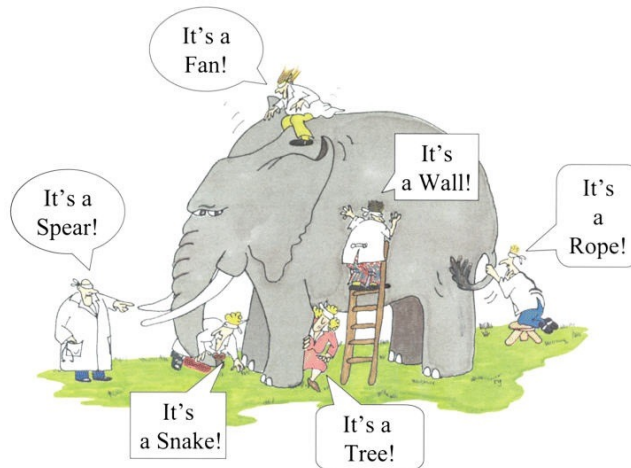
Nội dung

- Giới thiệu về phương pháp tập hợp mô hình
- Tương quan giữa Bias và Variance
- Bagging, Random forests, Boosting
- Kết luận và hướng phát triển

2

2

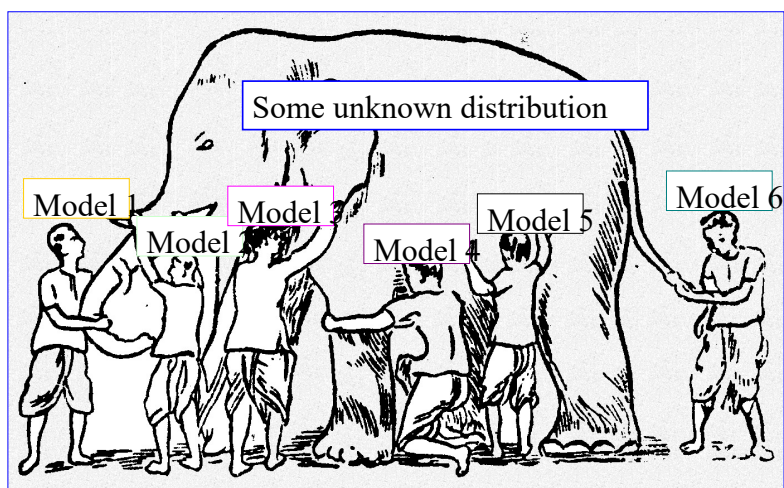
Phương pháp tập hợp mô hình - ensemble methods



https://miro.medium.com/max/1400/1*R-1jRqX1jT2x5ciVYq02og.png

3

Phương pháp tập hợp mô hình - ensemble methods



4

4

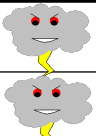






































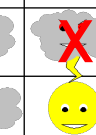



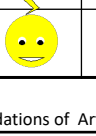



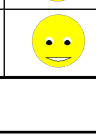

Phương pháp tập hợp mô hình - ensemble methods

- Phương pháp tập hợp mô **hình kết hợp nhiều mô hình cơ sở** dựa trên **tập học** nhằm cải thiện độ chính xác của giải thuật dự đoán.
- **Kết hợp các mô hình phân loại yếu** (weak learner/classifier) thành một mô hình phân loại mạnh (strong classifier)

5

5

Ví dụ: Đưa ra dự đoán dựa vào 5 mô hình

Reality							
1							
2							
3							
4							
5							
Combine							

CS 4700, Foundations of Artificial Intelligence, Carla P. Gomes

6

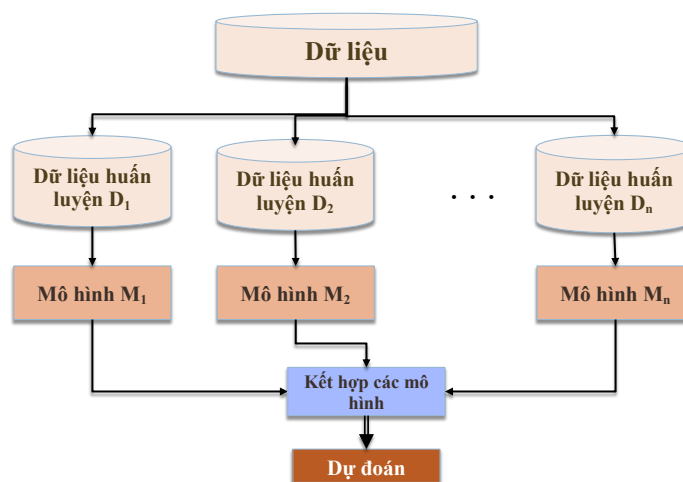
Phương pháp tập hợp mô hình - ensemble methods

- Cho kết quả tốt, tuy nhiên **không thể diễn dịch kết quả sinh ra**
- Ứng dụng thành công trong nhiều lĩnh vực như tìm kiếm thông tin, nhận dạng, phân tích dữ liệu,...

7

7

Phương pháp tập hợp mô hình - ensemble methods



8

8

Phương pháp tập hợp mô hình - ensemble methods

- Phương pháp ensemble-based
 - ❑ Xây dựng **tập hợp các mô hình cơ sở** dựa trên tập học
 - ❑ Tăng hiệu quả của mô hình dựa trên cơ sở **giảm lỗi bias/variance**
 - ❑ Giải thuật cơ sở cho các mô hình con: cây quyết định, SVM, naive Bayes, ...
- Một số giải thuật:
 - ❑ Bagging (Breiman, 1996)
 - ❑ Boosting (Freund & Schapire, 1995)
 - ❑ Random forests (Breiman, 2001)

9

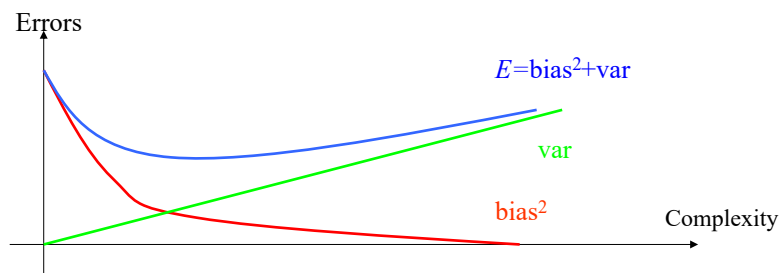
9

Phương pháp tập hợp mô hình - ensemble methods

Khi phân tích thành phần lỗi của giải thuật học, Breiman đã chỉ ra lỗi của giải thuật (E) gồm 2 thành phần là Bias và Variance

$$E = \text{bias}^2 + \text{var}$$

- **bias** : lỗi của mô hình - thành phần lỗi độc lập với mẫu dữ liệu học
- **variance** : thành phần lỗi do biến động liên quan đến sự ngẫu nhiên của tập học

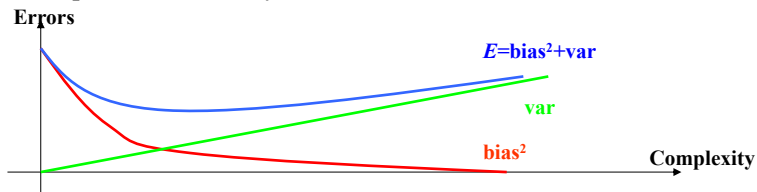


10

10

Bias và variance

- “**Underfitting**”: mô hình là quá "đơn giản" để đại diện cho tất cả đặc điểm của dữ liệu học
 - Bias cao và variance (phương sai) thấp
 - Lỗi quá trình huấn luyện và kiểm tra đều cao



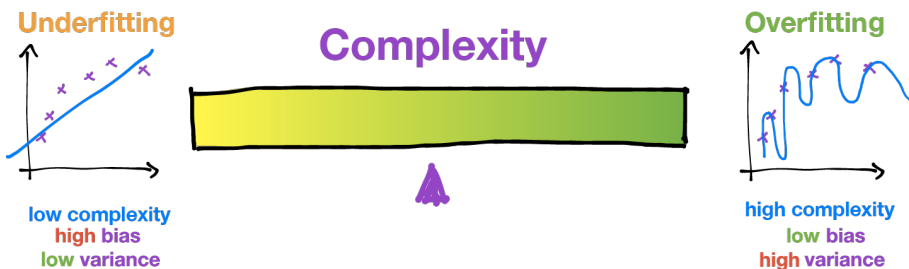
- “**Overfitting**”: mô hình là quá "phức tạp" và phù hợp với duy nhất đặc điểm của dữ liệu đang học, không thích hợp khi gặp dữ liệu nhiều.
 - Bias thấp và variance (phương sai) cao
 - Lỗi quá trình huấn luyện thấp và kiểm tra lỗi cao

Slide credit: L. Lazebnik

11

Bias và variance

- “Underfitting – học không thuộc bài”
- “Overfitting – học vẹt”:

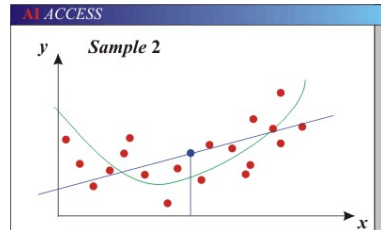


© Machine Learning @ Berkeley

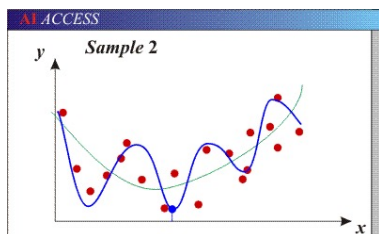
Slide credit: L. Lazebnik

12

Bias và variance



- Các mô hình không chính xác vì có **quá ít tham số** - **lỗi Bias lớn** (mô hình không đủ linh hoạt)



- Các mô hình không chính xác vì có **quá nhiều tham số** - **lỗi variance lớn** (quá nhạy với mẫu học)

Slide credit: D. Hoiem

13

Phương pháp tập hợp mô hình - ensemble methods

- Về bản chất, các phương pháp tập hợp mô hình thường làm **giảm lỗi bias và/hoặc variance** của các giải thuật máy học.
- Lỗi *bias* là lỗi liên quan đến mô hình (bộ phân lớp/ dự đoán) mà không liên quan đến dữ liệu được dùng để huấn luyện
- Lỗi *variance* là lỗi do tính biến thiên của mô hình so với tính ngẫu nhiên của các mẫu dữ liệu học (data samples).

14

14

Phương pháp tập hợp mô hình - ensemble methods

■ Averaging technique

- averaging technique
- xây dựng tập hợp các mô hình cơ sở độc lập nhau
- **kết hợp sự phân loại** của các mô hình
- **giảm variance**
- **bagging** và **random forests**

■ Boosting technique

- xây dựng tập hợp các mô hình cơ sở **tuần tự**
- Tập trung **cải tiến lỗi** sinh ra từ các mô hình trước
- **giảm bias**
- AdaBoost và Arcing

15

15

Averaging technique

16

Averaging technique - Bagging (Breiman, 1996)

■ Bootstrap AGGREGatING - (Breiman, 1996)

■ Bootstrap???

- Từ tập dữ liệu **D** ban đầu có **m** phần tử, người ta thực hiện lấy mẫu có hoàn lại **m** phần tử từ tập **D**, thu được tập **B** (gọi là bootstrap)

$X=(3.12, 0, 1.57, 19.67, 0.22, 2.20)$

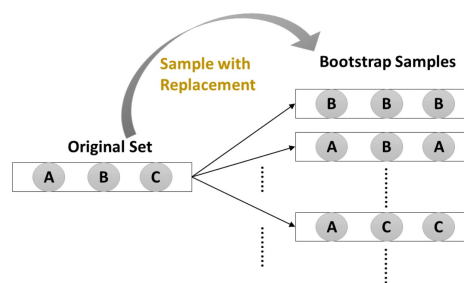
17

17

Averaging technique - Bagging (Breiman, 1996)

■ Bootstrap AGGREGatING - (Breiman, 1996)

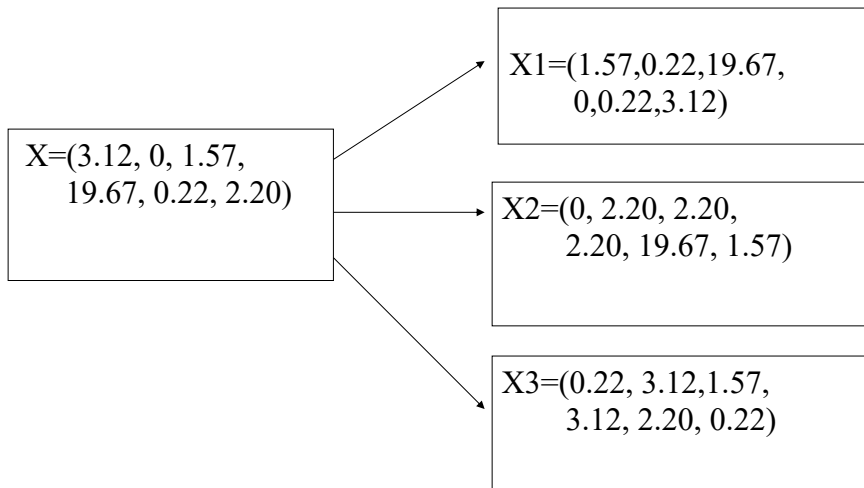
■ Bootstrap???



18

18

Bootstrap - Từ tập dữ liệu D ban đầu có m phần tử, người ta thực hiện lấy mẫu có hoàn lại m phần tử từ tập D, thu được tập B (gọi là bootstrap)



19

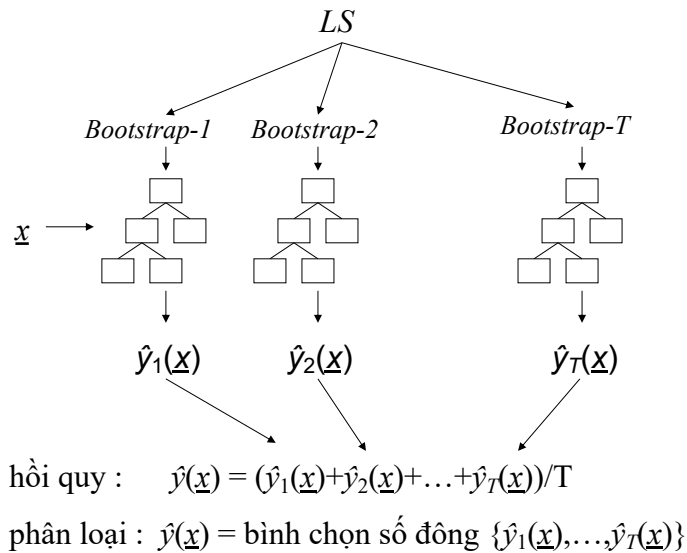
Bootstrap – lấy mẫu có hoàn lại

Original Data	1	2	3	4	5	6	7	8	9	10
Bagging (Round 1)	7	8	10	8	2	5	10	10	5	9
Bagging (Round 2)	1	4	9	1	2	3	2	7	3	2
Bagging (Round 3)	1	8	5	10	5	5	9	6	3	7

- Mỗi mẫu được lấy có xác suất $1/N$
- Với mỗi mẫu không được chọn sau N lần, thì xác suất là $(1 - 1/N)^N$ (Khi N lớn thì xác suất này gần bằng $1/e$)
- Với mỗi mẫu được chọn, sau N lần thì xác suất là $1 - 1/e = 0.632$
=> Lấy mẫu Bootstrap chứa 63% dữ liệu gốc

20

Bagging (Breiman, 1996)



21

Bagging (Breiman, 1996)

■ Bootstrap AGGREGatING

- từ tập học LS (learning set) có N phần tử
- xây dựng tập hợp T **mô hình cơ sở độc lập nhau**
- mô hình thứ i được xây dựng trên tập mẫu bootstrap
- 1 bootstrap : lấy mẫu N phần tử có hoàn lại từ tập LS
- khi phân loại : sử dụng luật bình chọn số đông (majority vote)
- hồi quy : tính giá trị trung bình của dự đoán của các mô hình

22

22

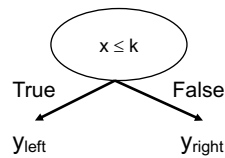
Bagging Example

- Xét tập dữ liệu sau với thuộc tính “x” và nhãn dự đoán “y”

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
y	1	1	1	-1	-1	-1	-1	1	1	1

- Nhãn dự đoán dựa vào luật:

- ▣ $x \leq k$ hay $x > k$
- ▣ Điểm phân hoạch dựa trên giá trị entropy



23

23

Bagging Example

Bagging Round 1:

x	0.1	0.2	0.2	0.3	0.4	0.4	0.5	0.6	0.9	0.9
y	1	1	1	1	-1	-1	-1	-1	1	1

$x \leq 0.35 \rightarrow y = 1$
 $x > 0.35 \rightarrow y = -1$

3/14/22

24

Bagging Example

Bagging Round 1:

x	0.1	0.2	0.2	0.3	0.4	0.4	0.5	0.6	0.9	0.9
y	1	1	1	1	-1	-1	-1	-1	1	1

$x \leq 0.35 \rightarrow y = 1$
 $x > 0.35 \rightarrow y = -1$

Bagging Round 2:

x	0.1	0.2	0.3	0.4	0.5	0.5	0.9	1	1	1
y	1	1	1	-1	-1	-1	1	1	1	1

$x \leq 0.7 \rightarrow y = 1$
 $x > 0.7 \rightarrow y = -1$

Bagging Round 3:

x	0.1	0.2	0.3	0.4	0.4	0.5	0.7	0.7	0.8	0.9
y	1	1	1	-1	-1	-1	-1	-1	1	1

$x \leq 0.35 \rightarrow y = 1$
 $x > 0.35 \rightarrow y = -1$

Bagging Round 4:

x	0.1	0.1	0.2	0.4	0.4	0.5	0.5	0.7	0.8	0.9
y	1	1	1	-1	-1	-1	-1	-1	1	1

$x \leq 0.3 \rightarrow y = 1$
 $x > 0.3 \rightarrow y = -1$

Bagging Round 5:

x	0.1	0.1	0.2	0.5	0.6	0.6	0.6	1	1	1
y	1	1	1	-1	-1	-1	-1	1	1	1

$x \leq 0.35 \rightarrow y = 1$
 $x > 0.35 \rightarrow y = -1$

25

25

Bagging Example

Bagging Round 6:

x	0.2	0.4	0.5	0.6	0.7	0.7	0.7	0.8	0.9	1
y	1	-1	-1	-1	-1	-1	-1	1	1	1

$x \leq 0.75 \rightarrow y = -1$
 $x > 0.75 \rightarrow y = 1$

Bagging Round 7:

x	0.1	0.4	0.4	0.6	0.7	0.8	0.9	0.9	0.9	1
y	1	-1	-1	-1	-1	1	1	1	1	1

$x \leq 0.75 \rightarrow y = -1$
 $x > 0.75 \rightarrow y = 1$

Bagging Round 8:

x	0.1	0.2	0.5	0.5	0.5	0.7	0.7	0.8	0.9	1
y	1	1	-1	-1	-1	-1	-1	1	1	1

$x \leq 0.75 \rightarrow y = -1$
 $x > 0.75 \rightarrow y = 1$

Bagging Round 9:

x	0.1	0.3	0.4	0.4	0.6	0.7	0.7	0.8	1	1
y	1	1	-1	-1	-1	-1	-1	1	1	1

$x \leq 0.75 \rightarrow y = -1$
 $x > 0.75 \rightarrow y = 1$

Bagging Round 10:

x	0.1	0.1	0.1	0.1	0.3	0.3	0.8	0.8	0.9	0.9
y	1	1	1	1	1	1	1	1	1	1

$x \leq 0.05 \rightarrow y = 1$
 $x > 0.05 \rightarrow y = 1$

26

26

Bagging Example

- Kết quả tổng của 10 lần lặp:

Round	Split Point	Left Class	Right Class
1	0.35	1	-1
2	0.7	1	1
3	0.35	1	-1
4	0.3	1	-1
5	0.35	1	-1
6	0.75	-1	1
7	0.75	-1	1
8	0.75	-1	1
9	0.75	-1	1
10	0.05	1	1

3/14/22

Introduction to Data Mining, 2nd Edition

27

27

Bagging Examp

- Sử dụng “majority vote” để xác định lớp của bộ phân loại tổng hợp

Round	Split Point	Left Class	Right Class
1	0.35	1	-1
2	0.7	1	1
3	0.35	1	-1
4	0.3	1	-1
5	0.35	1	-1
6	0.75	-1	1
7	0.75	-1	1
8	0.75	-1	1
9	0.75	-1	1
10	0.05	1	1

Round	x=0.1	x=0.2	x=0.3	x=0.4	x=0.5	x=0.6	x=0.7	x=0.8	x=0.9	x=1.0
1	1	1	1	-1	-1	-1	-1	-1	-1	-1
2	1	1	1	1	1	1	1	1	1	1
3	1	1	1	-1	-1	-1	-1	-1	-1	-1
4	1	1	1	-1	-1	-1	-1	-1	-1	-1
5	1	1	1	-1	-1	-1	-1	-1	-1	-1
6	-1	-1	-1	-1	-1	-1	-1	1	1	1
7	-1	-1	-1	-1	-1	-1	-1	1	1	1
8	-1	-1	-1	-1	-1	-1	-1	1	1	1
9	-1	-1	-1	-1	-1	-1	-1	1	1	1
10	1	1	1	1	1	1	1	1	1	1
Sum	2	2	2	-6	-6	-6	-6	2	2	2
Sign	1	1	1	-1	-1	-1	-1	1	1	1

Predicted Class

28

28