Giới thiệu Học máy
Mô hình Naïve Bayes

Tô Hoài Việt
Khoa Công nghệ Thông tin
Đại học Khoa học Tự nhiên TPHCM
thviet@fit.hcmuns.edu.vn

Nội dung

- Giới thiệu Học máy
- Học là gì?
- Các vấn đề và ví dụ của học
- Mô hình Naïve Bayes

Tại sao Học Máy?

- Những tiến bộ gần đây trong thuật toán và lý thuyết
- "Dòng lũ" đang lên của dữ liệu trực tuyến
- Sức mạnh tính toán đã sẵn sàng
- Ngành công nghiệp đang nở rộ

Ba lĩnh vực thích hợp cho học máy

- Khai thác dữ liệu: sử dụng dữ liệu cũ để cải thiện quyết định
- Các ứng dụng phần mềm chúng ta không thể làm bằng tay
- Các chương trình tự tối ưu hoá

Học là gì?

- ghi nhớ điều gì đó
- học các sự kiện qua quan sát và thăm dò
- cải thiện các kỹ năng vận động và/hay nhận thức qua việc luyện tập
- tổ chức tri thức mới thành các biểu diễn tổng quát, hiệu quả

Các loại học

- **Học có giám sát**: cho trước một tập mẫu các cặp input/output, tìm một luật thực hiện việc dự đoán các kết xuất gắn với các input mới
- **Gom cụm:** cho trước một tập mẫu, nhưng chưa gán nhãn, gom nhóm các mẫu thành các cụm "tự nhiên"
- **Học tăng cường:** một agent tương tác với thế giới thực hiện các quan sát, hành động, và được thưởng hay phạt; nó sẽ học để chọn các hành động theo cách để nhận được nhiều phần thưởng

Học một Hàm

Cho trước một tập mẫu các cặp input/output, tìm một hàm làm tốt được công việc biểu diễn mối quan hệ

- Phát âm: hàm ánh xạ từ ký tự sang âm thanh
- Ném một quả bóng: hàm ánh xạ từ vị trí đích thành quỹ đạo cánh tay
- Đọc các chữ viết tay: hàm ánh xạ từ tập các điểm ảnh thành các ký tự
- Chẩn đoán bệnh: hàm ánh xạ từ các kết quả xét nghiệm thành các loại bệnh tật

Các vấn đề để học một hàm

- ghi nhớ
- lấy trung bình
- tổng quát hoá

Bài toán ví dụ

Khi nào thì lái xe (drive or walk) ? Phụ thuộc vào:

- nhiệt độ (temperature)
- mưa tuyết dự kiến (expected precipitation)
- ngày trong tuần (day of the week)
- cô ấy có cần đi mua sắm trên đường về hay không (whether she needs to shop on the way home)
- cô ấy đang mặc gì (what's she wearing)

Ghi nhớ

temp	precip	day	shop	clothes	
80	none	sat	no	casual	**walk**
19	snow	mon	yes	casual	**drive**
65	none	tues	no	casual	**walk**
19	snow	mon	yes	casual	

Ghi nhớ

temp	precip	day	shop	clothes	
80	none	sat	no	casual	**walk**
19	snow	mon	yes	casual	**drive**
65	none	tues	no	casual	**walk**
19	snow	mon	yes	casual	**drive**

Lấy trung bình

Xử lý nhiễu trong dữ liệu

temp	precip	day	shop	clothes	
80	none	sat	no	casual	**walk**
80	none	sat	no	casual	**drive**
80	none	sat	no	casual	**drive**
80	none	sat	no	casual	**walk**
80	none	sat	no	casual	**walk**
80	none	sat	no	casual	**walk**
80	none	sat	no	casual	**walk**
80	none	sat	no	casual	

Lấy trung bình

Xử lý nhiễu trong dữ liệu

temp	precip	day	shop	clothes	
80	none	sat	no	casual	**walk**
80	none	sat	no	casual	**drive**
80	none	sat	no	casual	**drive**
80	none	sat	no	casual	**walk**
80	none	sat	no	casual	**walk**
80	none	sat	no	casual	**walk**
80	none	sat	no	casual	**walk**
80	none	sat	no	casual	**walk**

Nhiễu cảm biến

Xử lý nhiễu trong dữ liệu

temp	precip	day	shop	clothes	
81	none	sat	no	casual	**walk**
82	none	sat	no	casual	**walk**
78	none	sat	no	casual	**drive**
21	none	sat	no	casual	**drive**
18	none	sat	no	casual	**drive**
19	none	sat	no	casual	**drive**
17	none	sat	no	casual	**drive**
20	none	sat	no	casual	

Nhiễu cảm biến

Xử lý nhiễu trong dữ liệu

temp	precip	day	shop	clothes	
81	none	sat	no	casual	**walk**
82	none	sat	no	casual	**walk**
78	none	sat	no	casual	**drive**
21	none	sat	no	casual	**drive**
18	none	sat	no	casual	**drive**
19	none	sat	no	casual	**drive**
17	none	sat	no	casual	**drive**
20	none	sat	no	casual	**drive**

Tổng quát hoá

Xử lý dữ liệu chưa từng gặp trước đây

temp	precip	day	shop	clothes	
71	none	fri	yes	formal	**drive**
38	none	sun	yes	casual	**walk**
62	rain	weds	no	casual	**walk**
93	none	mon	no	casual	**drive**
55	none	sat	no	formal	**drive**
80	none	sat	no	casual	**walk**
19	snow	mon	yes	casual	**drive**
65	none	tues	no	casual	**walk**

Tổng quát hoá

Xử lý dữ liệu chưa từng gặp trước đây

temp	precip	day	shop	clothes	
71	none	fri	yes	formal	**drive**
38	none	sun	yes	casual	**walk**
62	rain	weds	no	casual	**walk**
93	none	mon	no	casual	**drive**
55	none	sat	no	formal	**drive**
80	none	sat	no	casual	**walk**
19	snow	mon	yes	casual	**drive**
65	none	tues	no	casual	**walk**
58	rain	mon	no	casual	

Một ví dụ khác

f_1	f_2	f_3	f_4	y
0	1	0	1	1
0	1	0	1	1
1	1	0	1	1
0	1	0	1	1
0	1	0	1	1
1	1	0	1	0
1	1	0	1	0
1	1	0	1	0
0	1	0	1	0
1	1	0	1	0

<0, 1, 0, 1> = 1

<1, 1, 0, 1> = 0

f_1	f_2	f_3	f_4	y
0	1	0	1	1
0	0	0	1	1
0	1	0	1	1
0	0	0	1	1
0	1	0	1	1
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	1	0	1	0
0	1	0	1	0

<0, 1, 0, 1> = 1

<0, 0, 0, 1> = 0

Một ví dụ khác (tt)

f_1	f_2	f_3	f_4	y
0	1	0	1	1
0	0	0	1	1
1	1	0	1	1
0	0	0	1	1
0	1	0	1	1
1	0	0	1	0
1	0	0	1	0
1	0	0	1	0
0	1	0	1	0
1	1	0	1	0

<0, 0, 0, 1> = ?

<1, 1, 0, 1> = ?

Naïve Bayes

- Dựa trên luật suy diễn xác suất của Bayes
- Cập nhật xác suất của giả thiết (hàm phân lớp) dựa trên chứng cứ
- Chọn giả thiết có xác suất lớn nhất sau khi tích hợp các chứng cứ

- Thuật toán đặc biệt hữu ích cho các lĩnh vực có **nhiều** đặc trưng

Ví dụ

f_1	f_2	f_3	f_4	y
0	1	1	0	1
0	0	1	1	1
1	0	1	0	1
0	0	1	1	1
0	0	0	0	1
1	0	0	1	0
1	1	0	1	0
1	0	0	0	0
1	1	0	1	0
1	0	1	1	0

- $R_1(1,1) = 1/5$: tỷ lệ tất cả các mẫu **dương (y=1)** có đặc trưng 1 **= 1**
- $R_1(0,1) = 4/5$: tỷ lệ tất cả các mẫu **dương** có đặc trưng 1 = 0

Ví dụ

f_1	f_2	f_3	f_4	y
0	1	1	0	1
0	0	1	1	1
1	0	1	0	1
0	0	1	1	1
0	0	0	0	1
1	0	0	1	0
1	1	0	1	0
1	0	0	0	0
1	1	0	1	0
1	0	1	1	0

- $R_1(1,1) = 1/5$: tỷ lệ tất cả các mẫu **dương (y=1)** có đặc trưng 1 **= 1**
- $R_1(0,1) = 4/5$: tỷ lệ tất cả các mẫu **dương** có đặc trưng 1 = 0
- $R_1(1,0) = 5/5$: tỷ lệ tất cả các mẫu **âm (y=0)** có đặc trưng 1 **= 1**
- $R_1(0,0) = 0/5$: tỷ lệ tất cả các mẫu **âm** có đặc trưng 1 = 0

Ví dụ

f_1	f_2	f_3	f_4	y
0	1	1	0	1
0	0	1	1	1
1	0	1	0	1
0	0	1	1	1
0	0	0	0	1
1	0	0	1	0
1	1	0	1	0
1	0	0	0	0
1	1	0	1	0
1	0	1	1	0

$R_1(1,1) = 1/5$ $R_1(0,1) = 4/5$

$R_1(1,0) = 5/5$ $R_1(0,0) = 0/5$

$R_2(1,1) = 1/5$ $R_2(0,1) = 4/5$

$R_2(1,0) = 2/5$ $R_2(0,0) = 3/5$

$R_3(1,1) = 4/5$ $R_3(0,1) = 1/5$

$R_3(1,0) = 1/5$ $R_3(0,0) = 4/5$

$R_4(1,1) = 2/5$ $R_4(0,1) = 3/5$

$R_4(1,0) = 4/5$ $R_4(0,0) = 1/5$

Dự đoán

$R_1(1,1) = 1/5$	$R_1(0,1) = 4/5$
$R_1(1,0) = 5/5$	$R_1(0,0) = 0/5$
$R_2(1,1) = 1/5$	$R_2(0,1) = 4/5$
$R_2(1,0) = 2/5$	$R_2(0,0) = 3/5$
$R_3(1,1) = 4/5$	$R_3(0,1) = 1/5$
$R_3(1,0) = 1/5$	$R_3(0,0) = 4/5$
$R_4(1,1) = 2/5$	$R_4(0,1) = 3/5$
$R_4(1,0) = 4/5$	$R_4(0,0) = 1/5$

- Mẫu mới x = <0,0,1,1>
- $S(1) = R_1(0,1) * R_2(0,1) * R_3(1,1) * R_4(1,1) = .205$
- $S(0) = R_1(0,0) * R_2(0,0) * R_3(1,0) * R_4(1,0) = 0$
- Ta có S(1) > S(0), do đó dự đoán lớp 1

Thuật toán Học

- Ước lượng từ dữ liệu, với mọi thuộc tính j, có miền giá trị $D_j = \{v_{1j}, v_{2j}, \ldots v_{nj}\}$, tính

$$P(y=1) = \frac{\#(y^i = 1)}{\#D} \qquad R_j(v_{ij}, 1) = \frac{\#(x_j^i = 1 \wedge y^i = 1)}{\#(y^i = 1)}$$

$$P(y=0) = \frac{\#(y^i = 0)}{\#D} \qquad R_j(v_{ij}, 0) = \frac{\#(x_j^i = 1 \wedge y^i = 0)}{\#(y^i = 0)}$$

Thuật toán Dự đoán

- Cho một mẫu x mới, x = (x_1, x_2,… x_n), tính

$$S(1) = P(y = 1)\prod R_j(x_i,1)$$

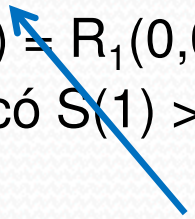
$$S(0) = P(y = 0)\prod R_j(x_i,0)$$

- Xuất ra 1 nếu S(1) > S(0)

Dự đoán

$R_1(1,1) = 1/5$	$R_1(0,1) = 4/5$
$R_1(1,0) = 5/5$	$R_1(0,0) = 0/5$
$R_2(1,1) = 1/5$	$R_2(0,1) = 4/5$
$R_2(1,0) = 2/5$	$R_2(0,0) = 3/5$
$R_3(1,1) = 4/5$	$R_3(0,1) = 1/5$
$R_3(1,0) = 1/5$	$R_3(0,0) = 4/5$
$R_4(1,1) = 2/5$	$R_4(0,1) = 3/5$
$R_4(1,0) = 4/5$	$R_4(0,0) = 1/5$

- Mẫu mới x = <0,0,1,1>
- $S(1) = R_1(0,1) * R_2(0,1) * R_3(1,1) * R_4(1,1) = .205$
- $S(0) = R_1(0,0) * R_2(0,0) * R_3(1,0) * R_4(1,0) = 0$
- Ta có S(1) > S(0), do đó dự đoán lớp 1



lưu ý cả hai lớp đều có tỷ lệ bằng nhau = 0.5

Thuật toán Dự đoán

- Cho một mẫu x mới, x = (x_1, x_2,… x_n), tính

$$\log S(1) = \log P(y=1) + \sum_{j=1} \log R_j(x_i, 1)$$

$$\log S(0) = \log P(y=0) + \sum_{j=1} \log R_j(x_i, 0)$$

- Xuất ra 1 nếu logS(1) > logS(0)

Cộng log sẽ dễ dàng hơn nhiều so với nhân các số nhỏ

Phép sửa lỗi Laplace

- Tránh sự xuất hiện của 1 hoặc 0 trong xác suất

$$P(y=1)=\frac{\#(y^i=1)+1}{\#D+2} \qquad R_j(v_{ij},1)=\frac{\#(x_j^i=1\wedge y^i=1)+1}{\#(y^i=1)+2}$$

$$P(y=0)=\frac{\#(y^i=0)+1}{\#D+2} \qquad R_j(v_{ij},0)=\frac{\#(x_j^i=1\wedge y^i=0)+1}{\#(y^i=0)+2}$$

có 2 phân lớp

x có 2 giá trị

Ví dụ với Sửa lỗi

f_1	f_2	f_3	f_4	y
0	1	1	0	1
0	0	1	1	1
1	0	1	0	1
0	0	1	1	1
0	0	0	0	1
1	0	0	1	0
1	1	0	1	0
1	0	0	0	0
1	1	0	1	0
1	0	1	1	0

$R_1(1,1) = 2/7$ $R_1(0,1) = 5/7$

$R_1(1,0) = 6/7$ **$R_1(0,0) = 1/7$**

$R_2(1,1) = 2/7$ $R_2(0,1) = 5/7$

$R_2(1,0) = 3/7$ $R_2(0,0) = 4/7$

$R_3(1,1) = 5/7$ $R_3(0,1) = 2/7$

$R_3(1,0) = 2/7$ $R_3(0,0) = 5/7$

$R_4(1,1) = 3/7$ $R_4(0,1) = 4/7$

$R_4(1,0) = 5/7$ $R_4(0,0) = 2/7$

Dự đoán

$R_1(1,1) = 2/7$	$R_1(0,1) = 5/7$
$R_1(1,0) = 6/7$	$R_1(0,0) = 1/7$
$R_2(1,1) = 2/7$	$R_2(0,1) = 5/7$
$R_2(1,0) = 3/7$	$R_2(0,0) = 4/7$
$R_3(1,1) = 5/7$	$R_3(0,1) = 2/7$
$R_3(1,0) = 2/7$	$R_3(0,0) = 5/7$
$R_4(1,1) = 3/7$	$R_4(0,1) = 4/7$
$R_4(1,0) = 5/7$	$R_4(0,0) = 2/7$

- Mẫu mới x = <0,0,1,1>
- $S(1) = R_1(0,1) * R_2(0,1) * R_3(1,1) * R_4(1,1) = .156$
- $S(0) = R_1(0,0) * R_2(0,0) * R_3(1,0) * R_4(1,0) = .017$
- Ta có S(1) > S(0), do đó dự đoán lớp 1

Điều cần nắm

- Các vấn đề của học máy
- Hiểu và sử dụng được mô hình Naïve Bayes
- Nắm được các vấn đề của Naïve Bayes