Trường Đại học Trà Vinh

Môn: Khai Khoáng Dữ Liệu

GV: Hà Thị Thúy Vi

Nhóm 5:
110117051 – Đỗ Trọng Hảo
110117035 – Nguyễn Minh Thư
110117079 – Thạch Visal

Tìm Hiểu Về Thuật Toán Naive Bayes

Naive Bayes là gì?

Naive Bayes là một thuật toán trong khai thác dữ liệu, nó dựa trên mô hình xác suất để phân loại dữ liệu.

Các bước thực hiện

Ta có tập dữ liệu huấn luyện như sau:

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	weak	No
sunny	hot	high	strong	No
overcast	hot	high	weak	Yes
rain	mild	High	weak	Yes
rain	cool	Normal	weak	Yes
rain	cool	normal	strong	No
overcast	cool	normal	strong	Yes
sunny	mild	high	weak	No
sunny	cool	normal	weak	Yes
rain	mild	normal	weak	Yes
sunny	mild	normal	strong	Yes
overcast	mild	high	strong	Yes
overcast	hot	normal	weak	Yes
rain	mild	high	strong	No

Yêu cầu đặt ra là hãy cho biết với một thời tiết nhất định thì có nên đi chơi hay không?

Bước 1: Xây dựng mô hình xác suất

Việc đi chơi hay không đi chơi, như vậy ta thấy có hai trường hợp, tạm gọi là lớp gồm lớp **"Yes"** và lớp **"No"**. Như dữ liệu ở trên ta thấy có 14 dòng dữ liệu, ta đếm xem trong 14 dòng đó có mấy dòng là **Yes** và mấy dòng là **No** ! Như vậy ta được xác suất của hai lớp **Yes** và **No** như sau:

P(Yes) = **9/14** và P(No) = **5/14** (P là xác suất)

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	weak	No
sunny	hot	high	strong	No
overcast	hot	high	weak	Yes
rain	mild	High	weak	Yes
rain	cool	Normal	weak	Yes
rain	cool	normal	strong	No
overcast	cool	normal	strong	Yes
sunny	mild	high	weak	No
sunny	cool	normal	weak	Yes
rain	mild	normal	weak	Yes
sunny	mild	normal	strong	Yes
overcast	mild	high	strong	Yes
overcast	hot	normal	weak	Yes
rain	mild	high	strong	No

Tiếp tục với **cột Outlook**, ta thấy có 3 giá trị là **sunny, overcast** và **rain**.

Đầu tiên ở lớp "Yes"!

P(sunny|Y) = 2/9 ; P(overcast|Y) = 4/9 ; P(rain|Y) = 3/9

Ở lớp "No"

P(sunny|N) = 3/5 ; P(overcast|N) = 0 ; P(rain|N) = 2/5

Tương tự với cột **Temperator, Humidty, Windy** được tóm tắt như sau:

Outlook	
P(sunny | y) = 2/9	P(sunny | n) = 3/5
P(overcast | y) = 4/9	P(overcast | n) = 0
P(rain | y) = 3/9	P(rain | n) = 2/5
Temperature	
P(hot | y) = 2/9	P(hot | n) = 2/5
P(mild | y) = 4/9	P(mild | n) = 2/5
P(cool | y) = 3/9	P(cool | n) = 1/5
Humidity	
P(high | y) = 3/9	P(high | n) = 4/5
P(normal | y) = 6/9	P(normal | n) = 1/5
Windy	
P(strong | y) = 3/9	P(strong | n) = 3/5
P(weak | y) = 6/9	P(weak | n) = 2/5

Bước 2: Phân Lớp

Ví dụ: đề cho: **outlook** = **sunny, temperator** = **cool, humidty** = **high, wind** = **strong**. Hãy cho biết có nên đi chơi hay không?

Ta sẽ tính tích của các xác suất này trong cả hai lớp **Yes** và **No**.
Yes: P(Yes) * P(sunny|y) * P(cool|y) * P(high|y) * P(strong|y) = 9/14 * 2/9 * 3/9 * 3/9 * 3/9 = **0.005**
No: P(No) * P(sunny|n) * P(cool|n) * P(high|n) * P(strong|n) = 5/14 * 3/5 * 1/5 * 4/5 * 3/5 = **0.021**
Ta sẽ lấy lớp có đáp án lớn hơn để làm kết quả, như vậy kết quả cho điều kiện thời tiết này là **"No"** tức là không nên đi chơi.

Lưu ý:

Nếu đề bài cho **outlook** = **overcast** thì lớp **No** sẽ bằng 0. Để tránh tình trạng này thì chúng ta phải làm trơn bằng Laplace như sau:

Thêm 1 vào tử số và thêm số lớp vào mẫu số

Ban đầu lớp **Yes** và **No** như sau:

$$P(Yes) = 9/14 \text{ và } P(No) = 5/14$$

Sau khi làm trơn – sẽ thêm 1 vào tử số và thêm 2 vào mẫu số (do có 2 lớp yes và no)

$$P(Yes) = 10/16 \text{ và } P(No) = 6/16$$

Tương tự thử với cột outlook . Do số lớp là 3 (có 3 giá trị) nên cộng 3 vào mẫu.

$$P(sunny|Y) = 3/12 ; P(overcast|Y) = 5/12 ; P(rain|Y) = 4/12$$
$$P(sunny|N) = 4/8 ; P(overcast|N) = 1/8 ; P(rain|N) = 3/8$$

Outlook	
P(sunny | y) = 3/12	P(sunny | n) = 4/8
P(overcast | y) = 5/12	P(overcast | n) = 1/8
P(rain | y) = 4/12	P(rain | n) = 3/8
Temperature	
P(hot | y) = 4/12	P(hot | n) = 3/8
P(mild | y) = 5/12	P(mild | n) = 3/8
P(cool | y) = 4/12	P(cool | n) = 2/8
Humidity	
P(high | y) = 4/11	P(high | n) = 5/7
P(normal | y) = 7/11	P(normal | n) = 2/7
Windy	
P(strong | y) = 4/11	P(strong | n) = 4/7
P(weak | y) = 7/11	P(weak | n) = 3/7

Ta sẽ sử dụng kết quả này để tính toán như ở trên và không bao giờ cho kết quả bằng 0.

Yes: P(Yes) * P(overcast|y) * P(cool|y) * P(high|y) * P(strong|y) = 10/16 * 5/12 * 4/12 * 4/11 * 4/11 = **0.011**

No: P(No) * P(overcast|n) * P(cool|n) * P(high|n) * P(strong|n) = 6/16 * 1/8 * 2/8 *

5/7 * 4/7 = **0.005**

Như kết quả trên chúng ta sẽ chọn lớp **Yes** tức là có thể đi chơi.