KHAI PHÁ DỮ LIỆU

VO DUC QUANG – VINH UNIVERSITY

QUANGVD@VINHUNI.EDU.VN

Naïve Bayes Classifier Algorithm

- Thuật toán Naïve Bayes là một thuật toán học có giám sát, dựa trên định lý Bayes và được sử dụng để giải các bài toán phân loại
- Là một bộ phân loại theo xác suất, có nghĩa là nó dự đoán trên cơ sở xác suất của một đối tượng
- Mô hình Naive Bayes ghi lại tần suất xuất hiện của giá trị thuộc tính nhãn lớp cùng với giá trị của các thuộc tính đầu vào
- Là một trong những thuật toán Phân loại đơn giản và hiệu quả nhất giúp xây dựng các mô hình học máy nhanh có thể đưa ra dự đoán nhanh chóng

Naïve Bayes Classifier Algorithm

- Thuật toán Naïve Bayes bao gồm hai từ Naïve và Bayes, có thể được mô tả như sau:
  - **Naïve:** Giả định rằng sự xuất hiện của một thuộc tính nhất định là độc lập với sự xuất hiện của các thuộc tính khác.

    Ví dụ:

    Chẳng hạn như nếu trái cây được xác định dựa trên các cơ sở về màu sắc, hình dạng và mùi vị, thì trái cây màu đỏ, hình cầu và ngọt được nhận biết là một quả táo.

    Do đó, mỗi đặc điểm riêng lẻ góp phần xác định đó là một quả táo mà không phụ thuộc vào nhau.
  - **Bayes:** được gọi là Bayes vì phụ thuộc vào nguyên lý của Định lý Bayes

Naïve Bayes Classifier Algorithm

❑ Định lý Bayes:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Công thức chỉ ra xác suất của A xảy ra nếu B cũng xảy ra, ta viết là $P(A|B)$. Và nếu ta biết xác suất của B xảy ra khi biết A, ta viết là $P(B|A)$ cũng như xác suất độc lập của A và B

- $P(A|B)$ là "xác suất của A khi biết B"
- $P(A)$ là xác suất xảy ra của A
- $P(B|A)$ là "xác suất của B khi biết A"
- $P(B)$ là xác suất xảy ra của B
- A và B là các biến cố độc lập với nhau

Naïve Bayes Classifier Algorithm

- Liên quan đến tập dữ liệu dùng trong học máy, áp dụng định lý Naïve Bayes như sau:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

  - y là nhãn lớp của tập dữ liệu
  - X là vector các thuộc tính phụ thuộc lẫn nhau (có n thuộc tính)

    $X = (x_1, x_2, x_3, \ldots, x_n)$
  - Ví dụ:

ID	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY GOLF
0	Rainy	Hot	High	False	No

- X = (Rainy, Hot, Hight, False); y = No
- $P(y|X)$ ở đây có nghĩa là, xác suất “Not playing golf” khi điều kiện thời tiết là “Rainy outlook”, “Temperature is hot”, “high humidity” and “no wind”.

Naïve Bayes Classifier Algorithm

- Giả sử các thuộc tính là độc lập dữ liệu với nhau, có thể biểu diễn lại công thức Bayes cho tập dữ liệu như sau:

$$P(y|x_1,\ldots,x_n) = \frac{P(x_1|y)P(x_2|y)\ldots P(x_n|y)P(y)}{P(x_1)P(x_2)\ldots P(x_n)}$$

- $P(y|x_1,\ldots,x_n)$ được gọi là xác suất hậu nghiệm. Bộ phân lớp sẽ gán nhãn cho mẫu dữ liệu được xét là lớp có xác suất hậu nghiệm lớn nhất.

- Do $P(x_1), P(x_2), \ldots, P(x_n)$ là như nhau với mọi lớp nên có thể bỏ qua. Khi đó, để tính $P(y|x_1,\ldots,x_n)$ cho từng lớp ta chỉ cần tính các xác suất thành phần $P(y|x_1)$, $P(y|x_2)$, ..., $P(y|x_n)$.

- Chú ý rằng y ở đây có thể biểu thị cho nhiều lớp, P(y_i) được ước lượng bằng |Di|/|D|, trong đó Di là tập các phần tử dữ liệu thuộc lớp y_i, D là tổng phần tử dữ liệu của dataset.

Naïve Bayes Classifier Algorithm

- Ví dụ: Dự báo thuê bao rời mạng với thuật toán Naïve Bayes

ID	Partner	PhoneService	StreamingTV	Churn
1	Yes	No	No	Yes
2	No	Yes	Yes	No
3	No	Yes	No	Yes
4	Yes	No	No internet service	No
5	No	Yes	Yes	Yes

Giả sử ta có một khách hàng mới X có giá trị tương ứng với các thuộc tính như sau:

S= (Partner = No, PhoneService = No, Streaming = Yes)

Bây giờ cần xác định xem khách hàng X có thuộc lớp Cyes (Khách hàng rời mạng) hay không?

Naïve Bayes Classifier Algorithm

- Ví dụ: Dự báo thuê bao rời mạng với thuật toán Naïve Bayes

Ta tính như sau:

P(Cyes) = 3/5 = 0.6

P(Cno)=2/5=0.4

Các xác suất thành phần:

P(Partner = No|Cyes) = 2/3 = 0.667

P(Partner = No|Cno) = 1/2 = 0.5

P(PhoneService = No|Cyes) = 1/3 = 0.333

P(PhoneService = No|Cno) = 1/2 = 0.5

P(StreamingTV = Yes|Cyes) = 1/3 = 0.333

P(StreamingTV = Yes|Cno) = 1/2 =0.5

Cuối cùng ta có:

P(X|Cyes) = 0.667 * 0.333 * 0.333 = 0.074

P(X|Cno) = 0.5 * 0.5 * 0.5 = 0.125

P(X|Cyes) * P(Cyes) = 0.074 * 0.6 = 0.044

P(X|Cno) * P(Cno) = 0.125 * 0.4 = 0.05

➔ Từ kết quả trên ta thấy P(X|Cno) * P(Cno) có giá trị lớn nhất, do đó thuật toán Naïve Bayes sẽ kết luận rằng khách hàng X sẽ không rời mạng

Naïve Bayes Classifier Algorithm

- Một số bộ phân loại Naive Bayes khác:
  - Multinomial Naive Bayes: sử dụng cho bài toán phân loại tài liệu, tức là liệu một tài liệu có thuộc thể loại thể thao, chính trị, công nghệ hay không, v.v.
  - Bernoulli Naive Bayes: Tương tự như Multinomial Naive Bayes nhưng biến được dự đoán là các giá trị Boolean
  - Gaussian Naive Bayes: Dùng trong trường hợp yếu tố dự đoán nhận các giá trị liên tục và không rời rạc

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$