Tìm hiểu mô hình: Naive Bayesian Classification

Người thực hiện: Trần Ngọc Bảo Duy - 51702091

1. Công thức xác suất đầy đủ

Xét n biến $\{A_i\}$, i = 1, 2, … n đầy đủ và B là một biến cố trong phép thử. Ta sẽ có

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$$

Ví dụ: có

- 70 bóng đèn vàng với tỉ lệ hỏng 2%,
- 30 bóng đèn trắng với tỉ lệ hỏng 1%,

a) Tính xác suất chọn được 1 bóng đèn hỏng

b) Chọn được bóng đèn hỏng, tính xác suất để bóng đó thuộc đèn trắng (Naïve Bayesian)

1. Công thức xác suất đầy đủ

a) Gọi A_i lần lượt là biến cố chọn bóng đèn vàng và trắng, i = 1, 2. Theo công thức:

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$$

- Ta có:

$P(A_1) = 70/100 = 0.7$

$P(A_2) = 30/100 = 0.3$

- Gọi B là biến cố chọn được bóng đèn hỏng

$P(B) = P(A_1)*P(B|A_1) + P(A_2)*P(B|A_2) = 0.7*0.02 + 0.3*0.01 = 0.017$

1. Công thức xác suất đầy đủ

b) Theo công thức Naive Bayesian:

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

- Để chọn được bóng đèn hỏng, xác suất để bóng đó thuộc đèn trắng là

$$P(A_2) = \frac{P(A2) * P(B|A2)}{P(B)} = \frac{P(A2) * P(B|A2)}{P(A1) * P(B|A1) + P(A2) * P(B|A2)} = \frac{0.3*0.01}{0.017} = 0.1764705882$$

=> Công thức Naive Bayesian dùng để phân loại các thực thể khi biết xác suất của nó

2. Naive Bayesian

Naive Bayesian là một mô hình xác suất có điều kiện. Xét bài toán classification với C classes 1,2,…, C. Giả sử có một điểm dữ liệu x∈R. Hãy tính xác suất để điểm dữ liệu này rơi vào class C. Nói cách khác, hãy tính:

$$p(y = c|\mathbf{x})$$

Tức tính xác suất để đầu ra là class C biết rằng đầu vào là vector x. Biểu thức này, nếu tính được, sẽ giúp chúng ta xác định được xác suất để điểm dữ liệu rơi vào mỗi class.

=> Từ đó có thể giúp xác định class của điểm dữ liệu đó bằng cách chọn ra class có xác suất cao nhất:

$$c = \arg \max_{c \in \{1,\ldots,C\}} p(c|\mathbf{x})$$

2. Naive Bayesian

$$c = \arg \max_{c \in \{1,\ldots,C\}} p(c|\mathbf{x})$$

Biểu thức trên thường khó được tính trực tiếp. Vì không phải lúc nào cũng có hết p(c|x) ứng với mỗi Class.

Vì vậy quy tắc Bayes thường được sử dụng:

$$\begin{aligned} c &= \arg\max_c p(c|\mathbf{x}) \\ &= \arg\max_c \frac{p(\mathbf{x}|c)p(c)}{p(\mathbf{x})} \\ &= \arg\max_c p(\mathbf{x}|c)p(c) \end{aligned}$$

**Theo công thức Bayes*

**Vì p(x) không phụ thuộc vào c cho nên có p(x) hay không cũng giống nhau, vì mình chọn xác suất lơn nhất*

2. Naive Bayesian

$$
\begin{aligned}
c &= \arg\max_c p(c|\mathbf{x}) \\
&= \arg\max_c \frac{p(\mathbf{x}|c)p(c)}{p(\mathbf{x})} \\
&= \arg\max_c p(\mathbf{x}|c)p(c)
\end{aligned}
$$

- Ta thấy p(c) có thể được hiểu là xác suất để một điểm rơi vào class C. Như ví dụ đầu là xác suất của $p(A_1)$ hay $p(A_2)$
- Thành phần còn lại p(x|c), tức phân phối của các điểm dữ liệu trong class C.

Như ví dụ đầu là $p(B|A_1)$ và $p(B|A_2)$

2. Naive Bayesian

- Thành phần còn lại p(x|c), tức phân phối của các điểm dữ liệu trong class C, thường rất khó tính toán vì x là một biến ngẫu nhiên nhiều chiều, cần rất rất nhiều dữ liệu training để có thể xây dựng được phân phối đó. Để giúp cho việc tính toán được đơn giản, người ta thường giả sử một cách đơn giản nhất rằng các thành phần của biến ngẫu nhiên x là độc lập với nhau, nếu biết c. Tức là:

$$p(\mathbf{x}|c) = p(x_1, x_2, \ldots, x_d|c) = \prod_{i=1}^{d} p(x_i|c)$$

- Như trong bài toán phân loại chủ đề các bài báo (NLP) thì các dữ liệu là tần suất các từ (tuỳ vào ngôn ngữ mà ta quy định chữ bao gồm số lượng từ như thế nào) xuất hiện trong bài báo làm feature cho chủ đề đó, vì nó độc lập với nhau

3. Train và Test dùng NBC

- Khi training data

Các phân phối p (c) và p (x_i|c), i = 1 , … , d sẽ được xác định dựa vào training data. Việc xác định các giá trị này có thể dựa vào Maximum Likelihood Estimation – là một kĩ thuật đi tìm bộ tham số θ sao cho xác suất sau đây đạt giá trị lớn nhất với xi các điểm dữ liệu với i $\in N$

$$\theta = \max_{\theta} p(\mathbf{x}_1, \ldots, \mathbf{x}_N | \theta)$$

Ví dụ khi xử lý ngôn ngữ tự nhiên, để phân loại các bài báo vào các chủ đề, ta có thể đi word to vec bằng cách dùng những từ thường xuyên xuất hiện của 1 từ trong 1 chủ đề (độ frequence – hay gọi là xác suất, tần suất xuất hiện), nhưng không được xuất hiện quá nhiều (bộ tham số θ với xác suất cao nhất để dễ dàng phân loại chủ đề) trong các chủ đề khác để làm feature cho training data

3. Train và Test dùng NBC

- Khi testing data

Ở bước **test**, với một điểm dữ liệu mới x, class của nó sẽ được xác đinh bởi:

$$c = \arg \max_{c \in \{1,\ldots,C\}} p(c) \prod_{i=1}^{d} p(x_i|c)$$

Có nghĩa là đầu vào là feature của dữ liệu cần predict, đầu ra sẽ là các xác suất của từng class và lấy giá trị nào lớn nhất, đó chính là class mà dữ liệu đầu vào thuộc về.

4. Kết luận

- Cả việc training và test của NBC là cực kỳ nhanh khi so với các phương pháp classification phức tạp khác.

- Việc giả sử các thành phần trong dữ liệu là độc lập với nhau, nếu biết class, khiến cho việc tính toán mỗi phân phối p(xi|c) trở nên cực kỳ nhanh.

- Mỗi giá trị p(c), c =1, 2, …, C có thể được xác định như là tần suất xuất hiện của class c trong training data. Việc tính toán p(xi|c) phụ thuộc vào loại dữ liệu.

- Có ba loại được sử dụng phổ biến là: Gaussian Naive Bayes, Multinomial Naive Bayes, và Bernoulli Naive .

5. Gaussian Naive Bayes

Mô hình này được sử dụng chủ yếu trong loại dữ liệu mà các thành phần là các biến liên tục.

Với mỗi chiều dữ liệu ii và một class c, xi tuân theo một phân phối chuẩn có kỳ vọng μci và phương sai σ2ci:

$$p(x_i|c) = p(x_i|\mu_{ci}, \sigma_{ci}^2) = \frac{1}{\sqrt{2\pi\sigma_{ci}^2}} \exp\left(-\frac{(x_i - \mu_{ci})^2}{2\sigma_{ci}^2}\right)$$

Trong đó, bộ tham số θ={μci,σ2ci} được xác định bằng Maximum Likelihood:

$$(\mu_{ci}, \sigma_{ci}^2) = \arg\max_{\mu_{ci}, \sigma_{ci}^2} \prod_{n=1}^{N} p(x_i^{(n)}|\mu_{ci}, \sigma_{ci}^2)$$

6. Multinomial Naive Bayes

Mô hình này chủ yếu được sử dụng trong phân loại văn bản mà feature vectors được tính bằng Bags of Words. Lúc này, mỗi văn bản được biểu diễn bởi một vector có độ dài d chính là số từ trong từ điển. Giá trị của thành phần thứ i trong mỗi vector chính là số lần từ thứ i xuất hiện trong văn bản đó. Khi đó, p(xi|c) tỉ lệ với tần suất từ thứ i (hay feature thứ i cho trường hợp tổng quát) xuất hiện trong các văn bản của class c. Giá trị này có thể được tính bằng cách:

$$\lambda_{ci} = p(x_i|c) = \frac{N_{ci}}{N_c}$$

Trong đó:

- Nci là tổng số lần từ thứ i xuất hiện trong các văn bản của class c, nó được tính là tổng của tất cả các thành phần thứ i của các feature vectors ứng với class c.
- Nc là tổng số từ (kể cả lặp) xuất hiện trong class c. Nói cách khác, nó bằng tổng độ dài của toàn bộ các văn bản thuộc vào class c.

7. Bernoulli Naive Bayes

Mô hình này được áp dụng cho các loại dữ liệu mà mỗi thành phần là một giá trị binary - bẳng 0 hoặc 1.

Ví dụ: cũng với loại văn bản nhưng thay vì đếm tổng số lần xuất hiện của 1 từ trong văn bản, ta chỉ cần quan tâm từ đó có xuất hiện hay không.

Khi đó, p(xi|c) được tính bằng: $p(x_i|c) = p(i|c)^{x_i}(1 - p(i|c)^{1-x_i}$

với p(i|c) có thể được hiểu là xác suất từ thứ i xuất hiện trong các văn bản của class c.