Báo cáo thực tập ngày 24/06/2024

Nguyễn Tiến Đạt - Thực tập sinh AI

1 NAIVE BAYES

1.1 Naive Bayes classfier (NBC)

Naive Bayes là kỹ thuật phân loại phổ biến trong học máy có giám sát. Ý tưởng chính của kỹ thuật này dựa vào xác suất có điều kiện giữa từ hay cụm từ và nhãn phân loại để dự đoán văn bản mới cần phân loại thuộc lớp nào. Naive Bayes được ứng dụng nhiều trong giải quyết các bài toán phân loại văn bản, xây dựng bộ lọc thư rác tự động, hay trong bài toán khai phá quan điểm bởi tính dễ hiểu, dễ triển khai cũng như độ chính xác cao.

Thuật toán Naive Bayes dựa trên định lý Bayes:

$$P(Y|X) = \frac{P(XY)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)}$$

Trong đó:

- $P(X|Y)$ là xác suất xảy ra một sự kiện ngẫu nhiên X khi biết sự liên quan Y đã xảy ra.
- $P(Y|X)$ là xác suất xảy ra Y khi biết X đã xảy ra.
- $P(X)$ là xác suất xảy ra của riêng X mà không quan tâm đến Y.
- $P(Y)$ là xác suất xảy ra của riêng Y mà không quan tâm đến X.

Áp dụng trong bài toán phân loại, dữ kiện gồm có:

- D: tập dữ liệu huấn luyện được vectơ hoá dưới dạng $x = [x_1, x_2, ..., x_n]$.
- C_i: phân loại i, với $i = \{1, 2, ..., m\}$.
- Các thuộc tính độc lập điều kiện đôi một với nhau.

Theo định lý Bayes:

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

Theo tính chất độc lập điều kiện:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

Trong đó:

- $P(C_i|X)$ là xác suất thuộc phân loại i khi biết trước mẫu X.
- $P(C_i)$ là xác suất phân loại i.
- $P(x_k|C_i)$ là xác suất thuộc tính thứ k mang giá trị x_k khi đã biết X thuộc phân loại i.

1.2 Các phân phối thường dùng trong NBC

1.2.1 Gaussian naive Bayes

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

Giả thiết, tại chiều dữ liệu thứ $i (i = 1, 2, ..., d)$ và phân lớp c, dữ liệu x_i tuân theo phân bố có kỳ vọng μ_{ci} và phương sai σ_{ci}^2 là:

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

Trong đó bộ tham số $\theta = \left\{\mu_{ci}, \sigma_{ci}^2\right\}$ được xác định bằng Maximum Likelihood.

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

Cách tính này được tham khảo từ thư viện scikit-learn.

1.2.2 Multinomial Naive Bayes

Mô hình này chủ yếu được sử dụng trong phân loại văn bản mà feature vectors được tính bằng *Bag of Words* (BoW). Trong kỹ thuật này mỗi văn bản được biểu diễn bởi một vector có độ dài d - là số từ trong từ điển với số chiều rất lớn. Giá trị thành phần (toạ độ) thứ i của mỗi vector chính là số lần từ thứ i (trong từ điển) xuất hiện trong văn bản.

$p(x_i|c)$ sẽ là tần số xuất hiện từ thứ i trong toàn bộ các văn bản của lớp c. Giá trị này thường được tính theo công thức:

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

Trong đó:

- N_{ci} là tổng số lần từ thứ i xuất hiện trong các văn bản thuộc lớp c. Nó được tính là tổng của tất cả các thành phần (toạ độ) thứ i của các điểm dữ liệu (feature vectors) trong phân lớp c.
- N_c là tổng số từ (kể cả lặp) xuất hiện trong phân lớp c. Tức là N_c bằng tổng độ dài tính theo từ của toàn bộ các văn bản thuộc vào lớp c.
- Có thể suy ra $N_c = \sum_{i=1}^{d} N_{ci}$ và do đó $\sum_{i=1}^{d} \lambda_{ci} = 1$.

Nhược điểm của công thức trên: Nếu một từ không xuất hiện trong phân lớp c thì $\lambda_{ci} = p(x_i|c) = 0$, dẫn dới $p(x|c)$ luôn bằng 0, cho dù các từ có tần suất rất lớn. Đặc điểm này dẫn đến kết quả không chính xác.

Để tránh nhược điểm này, một kỹ thuật được gọi là *Laplace smoothing* được áp dụng:

$$\hat{\lambda}_{ci} = \frac{N_{ci} + \alpha}{N_c + d\alpha}$$

- α là một số dương, thường bằng 1, để tránh trường hợp tỷ số bằng 0.
- Mẫu số được cộng với $d\alpha$ nên có thể đảm bảo bằng tổng xác suất $\sum_{i=1}^{d} \hat{\lambda}_{ci} = 1$.
- Bây giờ mỗi lớp c sẽ được mô tả bằng một d số dương có tổng bằng 1 $\hat{\lambda}_c = \{\hat{\lambda}_{c1}, ..., \hat{\lambda}_d\}$.

1.2.3 Bernoulli Naive Bayes

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

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

Trong trường hợp này, công thức $p(x_i|c)$ như sau:

$$p(x_i|c) = p(i|c)^{x_i}(1 - p(i|c))^{1-x_i}$$

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

1.3 Ví dụ

Xét một bộ dữ liệu đơn giản về việc đi làm muộn của một bạn nhân viên. Bộ dữ liệu được biểu diễn dạng bảng dưới đây:

	Giờ dậy (x1)	Sức khoẻ (x2)	Thời tiết (x3)	Đi muộn (y)
1	Sớm	Tốt	Nắng	Không
2	Sớm	Tốt	Mưa	Không
3	Bình thường	Tốt	Nắng	Có
4	Muộn	Xấu	Nắng	Có
5	Sớm	Xấu	Nhiều mây	Không
6	Bình thường	Xấu	Nhiều mây	Không
7	Muộn	Tốt	Nắng	Có
8	Bình thường	Tốt	Nắng	Không
9	Sớm	Xấu	Nhiều mây	Có
10	Muộn	Tốt	Mưa	Có

Nhận thấy rằng ở đây có hai lớp "Muộn" và "Không muộn", ta cần tìm *p(Có)* và *p(Không)* dựa trên tần suất xuất hiện của mỗi class. Ta có:

p(Có)=$\frac{5}{10}$, *p(Không)*=$\frac{5}{10}$

Tập hợp toàn bộ các từ trong từ điển là:

V = {Sớm, Bình thường, Muộn, Tốt, Xấu, Nắng, Mưa, Nhiều mây}

Tổng cộng số phần tử trong từ điển là $|V| = 8$.

Giả sử để dự đoán một ngày x_{11}=[Muộn, Xấu, Mưa]. Ta có quá trình huấn luyện và kiểm thử cho bài toán khi sử dụng Multinomial Naive Bayes, trong đó *Laplace smoothing* được sử dụng với $\alpha = 1$.

class = "Không muộn"

	Sớm	Bình thường	Muộn	Tốt	Xấu	Nắng	Mưa	Nhiều mây
d1: x_1	1	0	0	1	0	1	0	0
d2: x_2	1	0	0	0	1	0	1	0
d5: x_5	1	0	0	0	1	0	0	1
d6: x_6	0	1	0	0	1	0	0	1
d8: x_8	0	1	0	1	0	1	0	0
Total	3	2	0	2	3	2	1	2
$\Rightarrow \widehat{\lambda}_K$	$\frac{4}{21}$	$\frac{3}{21}$	$\frac{1}{21}$	$\frac{3}{21}$	$\frac{4}{21}$	$\frac{3}{21}$	$\frac{2}{21}$	$\frac{3}{21}$

$d = |V| = 8$
$N_K = 15$
$(23 = N_K + |V|)$

Hình 1: Minh hoạ NBC với Multinomial Naive Bayes với class "Không muộn"

$d11: \ x_{11} = [0, 0, 1, 0, 1, 0, 1, 0]$

$p(K|d_{11}) \propto p(K) \prod_{i=1}^{d} p(x_i|K) = \frac{5}{10} * \frac{1}{23} * \frac{4}{23} * \frac{2}{23} = 3.29 * 10^{-4}$

class = "Muộn"

	Sớm	Bình thường	Muộn	Tốt	Xấu	Nắng	Mưa	Nhiều mây
d3: x_3	0	1	0	1	0	1	0	0
d4: x_4	0	0	1	0	1	1	0	0
d7: x_7	0	0	1	1	0	1	0	0
d9: x_9	1	0	0	0	1	0	0	1
d10: x_{10}	0	0	1	1	0	0	1	0
Total	1	1	3	3	2	3	1	1
$\Rightarrow \widehat{\lambda}_C$	$\frac{2}{23}$	$\frac{2}{23}$	$\frac{4}{23}$	$\frac{4}{23}$	$\frac{3}{23}$	$\frac{4}{23}$	$\frac{2}{23}$	$\frac{2}{23}$

$N_C = 15$
$23 = N_C + |V|$
$d11: \ x_{11} = [0, 0, 1, 0, 1, 0, 1, 0]$

Hình 2: Minh hoạ NBC với Multinomial Naive Bayes với class "Muộn"

$p(C|d_{11}) \propto p(C) \prod_{i=1}^{d} p(x_i|C) = \frac{5}{10} * \frac{4}{23} * \frac{3}{23} * \frac{2}{23} = 9.86 * 10^{-4}$

$\Rightarrow p(C|d_{11}) > p(K|d_{11}) \Rightarrow d(11) \in class$(Muộn)

Hai giá trị tìm được $3.29 * 10^{-4}$ & $9.86 * 10^{-4}$ không phải là hai xác suất cần tìm mà chỉ là hai đại lượng tỉ lệ thuận với hai xác suất đó. Để tính cụ thể, ta có thể làm như sau:

$$p(C|d_{11}) = \frac{9.86*10^{-4}}{3.29*10^{-4} \ \& \ 9.86*10^{-4}} \approx 0.75$$

$$\Rightarrow p(K|d_{11}) = 1 - p(C|d_{11}) \approx 1 - 0.75 = 0.25$$

Như vậy xác suất để d_{11} rơi vào class "Muộn" là 75%, vào class "Không muộn" là 25%.