Naive Bayes

Thuật toán phân loại Naive Bayes (dịch là "Bayes" ngây thơ), là một thuật toán phân loại xác suất dựa trên việc áp dụng định lý của Bayes với các khẳng định độc lập ngây thơ giữa các biến đặc trưng.

Trong định lý Bayes, mình có công thức:

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

Trong đó:

- *P(A|B)*: **Xác suất hậu nghiệm**, dịch từ "Posterior Probability" (Xác suất *P* xảy ra sự kiện *A* sau khi xuất hiện sự kiện *B*).
- *P(B|A)*: **Xác suất có điều kiện**, dịch từ "Conditional Probability" (Xác suất *P* xảy ra sự kiện *B* sau khi xuất hiện sự kiện *A*).
- *P(A)*: **Xác suất tiên nghiệm**, dịch từ "Prior Probability" (Xác suất *P* xảy ra sự kiện *A*.
- *P(B)*: Xác suất *P* xảy ra sự kiện *B*.

Áp dụng định lý Bayes vào học máy, mình sử dụng biến đặc trưng véc-tơ *X*, và các lớp $(y_1, y_2, \ldots, y_c)$ của nhãn/biến mục tiêu *y* với số lượng *C* lớp từ một tập dữ liệu vào công thức trên, mình có:

$$P(y|X) = \frac{P(X|y) * P(y)}{P(X)} \text{ với } y = [y_1, y_2, \ldots, y_c]$$

Trong đó:

- *P(y|X)*: Xác suất *P* xảy ra sự kiện *y* sau khi xuất hiện sự kiện *X*.
- *P(X|y)*: Xác suất *P* xảy ra sự kiện *X* sau khi xuất hiện sự kiện *y*.
- *P(y)*: Xác suất *P* xảy ra sự kiện y.
- *P(X)*: Xác suất *P* xảy ra sự kiện *X*.

Tính toán thành công *P(y|X)* sẽ giúp mình xác định được xác suất dữ liệu từ mỗi lớp của biến mục tiêu *y*. Và từ đó, nó có thể giúp mình dự đoán lớp $\bar{y}$ của điểm dữ liệu mới *X* bằng cách lọc ra lớp *y* có xác suất cao nhất:

$$\bar{y} = argmax\, P(y|X) = argmax\, \frac{P(X|y) * P(y)}{P(X)} = \boldsymbol{argmax P(X|y) * P(y)}$$

$$\text{với } y = [y_1, y_2, \ldots, y_C]$$

Công thức được in đậm không sử dụng mẫu số *P*(*X*) như công thức trước nó, vì nó không phụ thuộc vào lớp *y*.

Đối với *P*(*y*), nó có thể được coi là xác suất để một điểm dữ liệu thuộc về lớp *y*.

Thành phần *P*(*X*|*y*) là một phân phối của các điểm dữ liệu *X* thuộc lớp *y*, việc tính trực tiếp sẽ gặp nhiều khó khăn nếu tập dữ liệu có nhiều biến đặc trưng. Chính vì thế, mình giả sử rằng các biến đặc trưng x_i, với *n* là số lượng biến có trong tập, hoàn toàn độc lập lẫn nhau, nếu biết lớp *y* cụ thể. Mình có:

$$P(X|y) = P(x_1, x_2, \ldots, x_n|y) = P(x_1|y) * P(x_2|y) * \ldots * P(x_n|y) = \prod_{i=1}^{n} P(x_i|y)$$

Việc giả sử các biến đặc trưng độc lập lẫn nhau, biết lớp *y* cụ thể, khá hạn hẹp và hiếm khi tìm được dữ liệu mà các thành phần độc lập nhau. Tuy nhiên, giả thiết ngây thơ ấy mang lại kết quả tốt. Phương pháp xác định lớp *y* của dữ liệu sử dụng giả thiết trên có tên gọi là *Naive Bayes Classifier*.

Nhờ vào cách tính toán ngây thơ này, tốc độ **huấn luyện** và **thử** mô hình thuật toán cũng rất nhanh, khiến nó có khả năng xử lý các bài toán quy mô lớn.

Ở bước **huấn luyện**, các phân phối *P*(*y*) và *P*(x_i|*y*), với *n* số lượng đặc trưng, sẽ được xác định dựa trên **dữ liệu huấn luyện**. Sử công thức **ước tính khả năng xảy ra tối đa (Maximum Likelihood Estimation)**, hoặc **hậu nghiệm A tối đa (Maximum A Posteriori)**. (1)

Đối với bước **thử**, sử dụng một điểm dữ liệu *X* mới, lớp mà thuật toán dự đoán được xác định bởi công thức:

$$\bar{y} = argmax\ P(y) * \prod_{i=1}^{n} P(x_i|y)$$

$$\text{với } y = [y_1, y_2, \ldots, y_C]$$

Trong trường hợp thuật toán được sử dụng trên một tập dữ liệu có nhiều *n* số lượng đặc trưng và các xác suất nhỏ, vế phải của công thức trên sẽ là 1 con số khá nhỏ, tính toán có khả năng lớn gặp sai số. Vì thế, công thức thường được viết lại dưới dạng tương tự bằng cách lấy hàm *log*:

$$\bar{y} = argmax\ log(P(y)) + \sum_{i=1}^{n} log(P(x_i|y))$$

$$\text{với } y = [y_1, y_2, \ldots, y_c]$$

Việc tính toán $P(x_i|y)$ dựa vào loại dữ liệu bên trong tập dữ liệu. Có ba loại thường xuyên được sử dụng:

- Gaussian Naive Bayes (cho dữ liệu liên tục/dữ liệu số)
- Multinomial Naive Bayes (cho dữ liệu văn bản)
- Bernoulli Naive Bayes (cho dữ liệu nhị phân)

1. Gaussian Naive Bayes:

Với mỗi biến đặc trưng x_i, khẳng định nó đi theo một phân phối chuẩn, cùng xác suất của x_i nếu biết trước lớp *y*. Mình có:

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

$$\text{với } y = [y_1, y_2, \ldots, y_c]$$

Trong đó:

- μ_y: Trung bình (Mean) các đặc trưng x_i cho lớp *y*.
- σ_y^2: Phương sai (Variance) các đặc trưng x_i cho lớp *y*.

2. Multinomial Naive Bayes:

Đặc trưng của thuật toán cho loại này là khả năng phân loại văn bản dựa trên số lần xuất hiện của các từ ngữ cụ thể trong nó. Lúc này, $P(x_i|y)$ là tỉ lệ tần suất của từ ngữ thứ i (hoặc là đặc trưng thứ i trong trường hợp tổng quát) xuất hiện trong các văn bản thuộc lớp y. Mình có:

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

Trong đó:

- $N(x_i|y)$: Số lượng lần từ ngữ x_i xuất hiện trong lớp y.
- $N(X|y)$: Tổng số lượng từ ngữ trong lớp y.

Công thức trên có một hạn chế, là nếu có một từ ngữ mới chưa bao giờ xuất hiện trong lớp y, kết quả $P(x_i|y)$ cho từ ngữ ấy luôn luôn sẽ bằng 0, điều này sẽ khiến cho kết quả dự đoán ở bước **thử** cũng bằng 0.

Để tránh trường hợp này, mình sử dụng phương pháp **làm mượt Laplace (Laplace Smoothing)**:

$$P(x_i|y) = \frac{N(x_i|y) + \alpha}{N(X|y) + |V|\alpha}$$

Trong đó:

- α: Tham số **làm mượt** (Thường thì = 1, tránh trường hợp tử số bằng 0).
- $|V|$: Kích thước từ vựng của tập dữ liệu (Số lượng từ ngữ duy nhất xuất hiện trong tập).

3. Bernoulli Naive Bayes:

Dùng cho tập dữ liệu với các đặc trưng nhị phân - giữa **0** và **1**. Cũng có thể được sử lên dữ liệu văn bản, nhưng thay vì tính tần suất xuất hiện của 1 từ ngữ trong văn bản, mình chỉ cần quan tâm rằng từ ngữ đó có xuất hiện trong văn bản hay không. Mình có công thức

$$P(x_i|y) = P(i|y)^{x_i} * (1 - P(i|y))^{1-x_i}$$

Trong đó, $P(i|y)$ có thể được hiểu là xác suất từ ngữ thứ i xuất hiện trong các văn bản thuộc lớp y.

Đọc thêm:

(1): Maximum Likelihood Estimation or Maximum A Posteriori