PHÂN LOẠI BAYES ĐƠN GIẢN

Phần này sẽ đề cập tới phân loại Bayes đơn giản (Naïve Bayes), một phương pháp phân loại đơn giản nhưng có nhiều ứng dụng trong thực tế như phân loại văn bản, dự đoán sắc thái văn bản, lọc thư rác, chẩn đoán y tế. Phân loại Bayes đơn giản là trường hợp riêng của kỹ thuật học máy Bayes, trong đó các giả thiết về độc lập xác suất được sử dụng để đơn giản hóa việc tính xác suất.

1. Phương pháp phân loại Bayes đơn giản

Tương tự như học cây quyết định ở trên, phân loại Bayes đơn giản sử dụng trong trường hợp mỗi ví dụ được cho bằng tập các thuộc tính $<x_1, x_2, \ldots, x_n>$ và cần xác định nhãn phân loại y, y có thể nhận giá trị từ một tập nhãn hữu hạn C.

Trong giai đoạn huấn luyện, dữ liệu huấn luyện được cung cấp dưới dạng các mẫu $<\mathbf{x}_i, y_i>$. Sau khi huấn luyện xong, bộ phân loại cần dự đoán nhãn cho mẫu mới $\mathbf{x}$.

Theo lý thuyết học Bayes, nhãn phân loại được xác định bằng cách tính xác suất điều kiện của nhãn khi quan sát thấy tổ hợp giá trị thuộc tính $<x_1, x_2, \ldots, x_n>$. Thuộc tính được chọn, ký hiệu c_{MAP} là thuộc tính có xác suất điều kiện cao nhất (MAP là viết tắt của maximum a posterior), tức là:

$$y = c_{MAP} = \arg\max_{c_j \in C} P(c_j \mid x_1, x_2, \ldots, x_n)$$

Sử dụng quy tắc Bayes, biểu thức trên được viết lại như sau

$$c_{MAP} = \arg\max_{c_j \in C} \frac{P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j)}{P(x_1, x_2, \ldots, x_n)}$$

Trong vế phải của biểu thức này, mẫu số không phụ thuộc vào c_j và vì vậy không ảnh hưởng tới giá trị của C_{MAP}. Do đó, ta có thể bỏ mẫu số và viết lại như sau:

$$C_{MAP} = \arg\max_{c_j \in C} P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j)$$

Hai thành phần trong biểu thức trên được tính từ dữ liệu huấn luyện. Giá trị $P(c_j)$ được tính bằng tần suất quan sát thấy nhãn c_j trên tập huấn luyện, tức là bằng số mẫu có nhãn là c_j chia cho tổng số mẫu. Việc tính $P(x_1, x_2, \ldots, x_n \mid c_j)$ khó khăn hơn nhiều. Vấn đề là số tổ hợp giá trị của n thuộc tính cùng với nhãn phân loại là rất lớn khi n lớn. Để tính xác suất này được chính xác, mỗi tổ hợp giá trị thuộc tính phải xuất hiện cùng nhãn phân loại đủ nhiều, trong khi số mẫu huấn luyện thường không đủ lớn.

Để giải quyết vấn đề trên, ta giả sử các thuộc tính là độc lập về xác suất với nhau khi biết nhãn phân loại c_j. Trên thực tế, các thuộc tính thường không độc lập với nhau

như vậy, chẳng hạn đối với ví dụ chơi tennis, khi trời nắng thì xác suất nhiệt độ cao cũng lớn hơn. Chính vì dựa trên giả thiết độc lập xác suất đơn giản như vậy nên phương pháp có tên gọi "Bayes đơn giản". Tuy nhiên, như ta thấy sau đây, giả thiết như vậy cho phép tính xác suất điều kiện đơn giản hơn nhiều và trên thực tế phân loại Bayes có độ chính xác tốt trong rất nhiều ứng dụng.

Với giả thiết về tính độc lập xác suất có điều kiện, có thể viết:

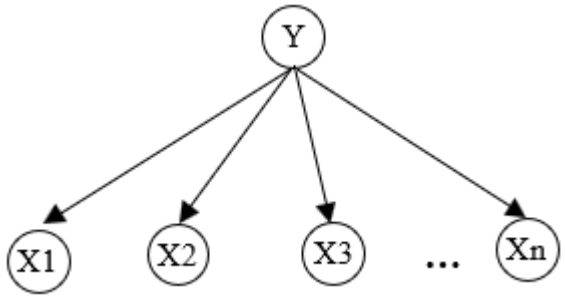
$$P(x_1, x_2,..., x_n \mid c_j) = P(x_1 \mid c_j)\, P(x_2 \mid c_j) \ldots P(x_n \mid c_j)$$

tức là xác suất đồng thời quan sát thấy các thuộc tính bằng tích xác suất điều kiện của từng thuộc tính riêng lẻ. Thay vào biểu thức ở trên, ta được **bộ phân loại Bayes đơn giản** (có đầu ra ký hiệu là c_{NB}) như sau.

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_i P(x_i \mid c_j)$$

trong đó, $P(x_i \mid c_j)$ được tính từ dữ liệu huấn luyện bằng số lần x_i xuất hiện cùng với c_j chia cho số lần c_j xuất hiện. Việc tính xác suất này đòi hỏi ít dữ liệu hơn nhiều so với tính $P(x_1, x_2,..., x_n \mid c_j)$.

Trên hình 1 là biểu diễn mô hình phân loại Bayes đơn giản dưới dạng mạng Bayes. Các thuộc tính không được nối với nhau bởi các cạnh và do vậy các thuộc tính độc lập xác suất với nhau nếu biết giá trị của nhãn phân loại.



Hình 1: Mô hình Bayes đơn giản: các thuộc tính Xi độc lập xác suất với nhau nếu biết giá trị nhãn phân loại Y.

Huấn luyện.

Quá trình huấn luyện hay học Bayes đơn giản là quá trình tính các xác suất $P(c_j)$ và các xác suất điều kiện $P(x_i \mid c_j)$ bằng cách đếm trên tập dữ liệu huấn luyện. Như vậy, khác với học cây quyết định, Học Bayes đơn giản không đòi hỏi tìm kiếm trong không gian các bộ phân loại. Các xác suất $P(c_j)$ và các xác suất điều kiện $P(x_i \mid c_j)$ được tính trên tập dữ liệu huấn luyện theo công thức sau:

$$P(c_j) = \frac{\text{Số mẫu có nhãn là } c_j}{\text{Tổng số mẫu trong tập huấn luyện}}$$

$$P(x_i \mid c_j) = \frac{\text{Số mẫu có giá trị thuộc tính } X_i = x_i \text{ và nhãn là } c_j}{\text{Số mẫu có nhãn là } c_j}$$

Ví dụ.

Để minh họa cho kỹ thuật học Bayes đơn giản, ta sử dụng lại bài toán phân chia ngày thành phù hợp hay không phù hợp cho việc chơi tennis theo điều kiện thời tiết đã được sử dụng trong phần học cây quyết định với dữ liệu huấn luyện cho trong bảng 4.1. Giả sử phải xác định nhãn phân loại cho ví dụ sau:

< Trời = nắng, Nhiệt độ = trung bình, Độ ẩm = cao, Gió = mạnh >

Thay số liệu của bài toán vào công thức Bayes đơn giản, ta có:

$$c_{NB} = \underset{c_j \in C}{\text{argmax}} P(c_j) \prod_i P(x_i \mid c_j)$$

$$= \underset{c_j \in \{co, khong\}}{\text{argmax}}$$

P(Trời=nắng| c_j) P(Nh. độ=t. bình| c_j) P(Độ ẩm=cao| c_j) P(Gió=mạnh |c_j)P (c_j)

Do c_j có thể nhận hai giá trị, ta cần tính 10 xác suất. Các xác suất P(có) và P(không) được tính bằng tất suất "có" và "không" trên dữ liệu huấn luyện.

P(có) = 9/14 = 0,64

P(không) = 5/14 = 0,36

Các xác suất điều kiện cũng được tính từ dữ liệu huấn luyện, ví dụ ta có:

P(Độ ẩm = cao | có) = 3/9 = 0,33

P (Độ ẩm = cao | không) = 4/5 = 0,8

Thay các xác suất thành phần vào công thức Bayes đơn giản, ta được:

P(có) P(nắng|có) P(trung bình|có) P(cao|có) P(mạnh|có) = 0.0053

P(không) P(nắng|không) P(trung bình|không) P(cao|không) P(mạnh|không) = 0.0206

Như vậy, theo phân loại Bayes đơn giản, ví dụ đang xét sẽ được phân loại là "không". Cần chú ý rằng, 0.0053 và 0.0206 không phải là xác suất thực của nhãn "có"

và "không". Để tính xác suất thực, ta cần chuẩn hóa để tổng hai xác suất bằng 1. Việc chuẩn hoá được thực hiện bằng cách chia mỗi số cho tổng của hai số. Chẳng hạn xác suất có chơi sẽ bằng 0.0053/(0.0053+0.0206) = 0.205.

2. Vấn đề tính xác suất trên thực tế

Phân loại Bayes đơn giản đòi hỏi tính các xác suất điều kiện thành phần $P(x_i \mid c_j)$. Xác suất này được tính bằng n_c / n, trong đó n_c số lần x_i và c_j xuất hiện đồng thời trong tập huấn luyện và n là số lần c_j xuất hiện.

Trong nhiều trường hợp, giá trị n_c có thể rất nhỏ, thậm chí bằng không, và do vậy ảnh hưởng tới độ chính xác khi tính xác suất điều kiện. Nếu $n_c = 0$, xác suất điều kiện cuối cùng sẽ bằng không, bất kể các xác suất thành phần khác có giá trị thế nào.

Để khắc phục vấn đề này, một kỹ thuật được gọi là *làm trơn* thường được sử dụng. Kỹ thuật làm trơn đơn giản nhất sử dụng công thức tính $P(x_i \mid c_j)$ như sau:

$$P(x_i \mid c_j) = (n_c + 1) \ / \ (n + 1)$$

Như vậy, kể cả khi $n_c = 0$, xác suất vẫn nhận giá trị khác 0.

Trong trường hợp chung, có thể sử dụng công thức được làm trơn sau:

$$P(x_i \mid c_j) = \frac{n_c + mp}{n + m}$$

trong đó p là xác suất tiền nghiệm của x_i và m là tham số cho phép xác định ảnh hưởng của p tới công thức. Nếu không có thêm thông tin gì khác thì xác suất tiền nghiệm thường được tính $p = 1 / k$, trong đó k là số thuộc tính của thuộc tính X_i. Ví dụ, nếu không có thêm thông tin gì thêm thì xác suất quan sát thấy Gió = mạnh sẽ là 1/2 do thuộc tính Gió có hai giá trị. Nếu $m = 0$, ta được công thức không làm trơn ban đầu. Ngược lại, khi m → ∞, xác suất hậu nghiệm sẽ bằng p, bất kể n_c thế nào. Trong những trường hợp c n lại, cả n_c / n và p cùng đóng góp vào công thức.

3. Ứng dụng trong phân loại văn bản tự động

Phân loại văn bản tự động là bài toán có nhiều ứng dụng thực tế. Trước tiên, cho một tập huấn luyện bao gồm các văn bản. Mỗi văn bản có thể thuộc vào một trong C loại khác nhau (ở đây ta không xét trường hợp mỗi văn bản có thể thuộc vào nhiều loại khác nhau). Sau khi huấn luyện xong, thuật toán phân loại nhận được văn bản mới và cần xác định phân loại cho văn bản này. Ví dụ, với các văn bản là nội dung thư điện tử,

thuật toán có thể phân loại thư thành "thư rác" và "thư bình thường". Khi huấn luyện, thuật toán học được cung cấp một tập thư rác và một tập thư thường. Sau đó, dựa trên nội dung thư mới nhận, bộ phân loại sẽ tự xác định đó có phải thư rác không. Một ứng dụng khác là tự động phân chia bản tin thành các thể loại khác nhau, ví dụ "chính trị", "xã hội", "thể thao".v.v. như trên báo điện tử.

Phân loại văn bản tự động là dạng ứng dụng trong đó phân loại Bayes đơn giản và các phương pháp xác suất khác được sử dụng rất thành công. Chương trình lọc thư rác mã nguồn mở SpamAssassin (http://spamassassin.apache.org) là một chương trình lọc thư rác được sử dụng rộng rãi với nhiều cơ chế lọc khác nhau, trong đó lọc Bayes đơn giản là cơ chế lọc chính được gán trọng số cao nhất.

Sau đây ta sẽ xem xét cách sử dụng phân loại Bayes đơn giản cho bài toán phân loại văn bản. Để đơn giản, ta sẽ xét trường hợp văn bản có thể nhận một trong hai nhãn: "rác" và "không".

Để sử dụng phân loại Bayes đơn giản, cần giải quyết hai vấn đề chủ yếu: thứ nhất, biểu diễn văn bản thế nào cho phù hợp; thứ hai: lựa chọn công thức cụ thể cho bộ phân loại Bayes.

Cách thông dụng và đơn giản nhất để biểu diễn văn bản là cách biểu diễn bằng "túi từ" (bag-of-word). Theo cách này, mỗi văn bản được biểu diễn bằng một tập hợp, trong đó mỗi phần tử của tập hợp tương ứng với một từ khác nhau của văn bản. Để đơn giản, ở đây ta coi mỗi từ là một đơn vị ngôn ngữ được ngăn với nhau bởi dấu cách. Lưu ý rằng đây là cách đơn giản nhất, ta cũng có thể thêm số lần xuất hiện thực tế của từ trong văn bản. Cách biểu diễn này không quan tâm tới vị trí xuất hiện của từ trong văn bản cũng như quan hệ với các từ xung quanh, do vậy có tên gọi là túi từ. Ví dụ, một văn bản có nội dung "Chia thư thành thư rác và thư thường" sẽ được biểu diễn bởi tập từ {"chia", "thư", "thành", "rác", "và", "thường"} với sáu phần tử.

Giả thiết các từ biểu diễn cho thư xuất hiện độc lập với nhau khi biết nhãn phân loại, công thức Bayes đơn giản cho phép ta viết:

$$c_{NB} = \underset{cj \in \{rac, khong\}}{\text{argmax}} P(c_j) \prod_i P(x_i \mid c_j)$$

$$= \underset{cj \in \{rac, khong\}}{\text{argmax}}$$

$$P(c_j)\ P(\text{"chia"} \mid c_j)\ P(\text{"thư "} \mid c_j)\ P(\text{"thành"} \mid c_j)\ P(\text{"rác"} \mid c_j)\ P(\text{"và"} \mid c_j)\ P(\text{"thường "} \mid c_j)$$

Các xác suất $P(\text{"rác"} \mid c_j)$ được tính từ tập huấn luyện như mô tả ở trên. Những từ chưa xuất hiện trong tập huấn luyện sẽ bị bỏ qua, không tham gia vào công thức.

Cần lưu ý rằng cách biểu diễn và áp dụng phân loại Bayes đơn giản cho phân loại văn bản vừa trình bày là những phương án đơn giản. Trên thực tế có rất nhiều biến thể khác nhau cả trong việc chọn từ, biểu diễn văn bản bằng các từ, cũng như công thức tính xác suất điều kiện của văn bản.

Mặc dù đơn giản, nhiều thử nghiệm cho thấy, phân loại văn bản tự động bằng Bayes đơn giản có độ chính xác khá cao. Trên nhiều tập dữ liệu thư điện tử, tỷ lệ phân loại chính xác thư rác có thể đạt trên 98%. Kết quả này cho thấy, mặc dù giả thiết các từ độc lập với nhau là không thực tế, độ chính xác phân loại của Bayes đơn giản không bị ảnh hưởng đáng kể.