Thuật toán Naïve Bayes

# Giới thiệu thuật toán Naïve Bayes:

Naïve Bayes (NB) là phương pháp phân loại dựa vào xác suất được sử dụng rộng rãi trong lĩnh vực máy học [Mitchell, 1996] [Joachims, 1997] [Jason, 2001], được sử dụng lần đầu tiên trong lĩnh vực phân loại bởi Maron vào năm 1961 [Maron, 1961] sau đó trở nên phổ biến dùng trong nhiều lĩnh vực như trong các công cụ tìm kiếm [Rijsbergen et al, 1970], các bộ lọc email [Sahami et al, 1998]…

# Thuật toán Naïve Bayes:

Thuật toán Naïve Bayes dựa trên định lý Bayes được phát biểu như sau:

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

* D: tập dữ liệu huấn luyện đã được vector hóa dưới dạng
* Ci: phân lớp i, với i = {1,2,…,m}.
* Các thuộc tính độc lập điều kiện đôi một với nhau.

Theo định lý Bayes:

Theo tính chất độc lập điều kiện:

Trong đó:

* là xác suất thuộc phân lớp i khi biết trước mẫu X.
* xác suất là phân lớp i.
* xác suất thuộc tính thứ k mang giá trị xk khi đã biết X thuộc phân lớp i.

Các bước thực hiện thuật toán Naïve Bayes:

Bước 1: Huấn luyện Naïve Bayes (dựa vào tập dữ liệu), tính và

Bước 2: Phân lớp , ta cần tính xác suất thuộc từng phân lớp khi đã biết trước Xnew. Xnew được gán vào lớp có xác suất lớn nhất theo công thức

# Áp dụng cho bài toán phân loại văn bản:

Ý tưởng: Ý tưởng cơ bản của cách tiếp cận Naïve Bayes là sử dụng xác suất có điều kiện giữa từ và chủ đề để dự đoán xác suất chủ đề của một văn bản cần phân loại. Điểm quan trọng của phương pháp này chính là ở chỗ giả định rằng sự xuất hiện của tất cả các từ trong văn bản đều độc lập với nhau. Giả định đó làm cho việc tính toán NB hiệu quả và nhanh chóng hơn các phương pháp khác vì không sử dụng việc kết hợp các từ để đưa ra phán đoán chủ đề. Kết quả dự đoán bị ảnh hưởng bởi kích thước tập dữ liệu, chất lượng của không gian đặc trưng…

Hướng dẫn cài đặt:

Mô tả vector đặc trưng của văn bản: Là vector có số chiều là số đặc trưng trong toàn tập dữ liệu, các đặc trưng này đôi một khác nhau. Nếu văn bản có chứa đặc trưng đó sẽ có giá trị 1, ngược lại là 0.

Thuật toán gồm 2 giai đoạn huấn luyện và phân lớp:

1. Huấn luyện: tính và

Đầu vào:

* Các vector đặc trưng của văn bản trong tập huấn luyện (Ma trận MxN, với M là số vector đặc trưng trong tập huấn luyện, N là số đặc trưng của vector).
* Tập nhãn/lớp cho từng vector đặc trưng của tập huấn luyện.

Đầu ra:

* Các giá trị xác suất và .

Công thức tính đã làm trơn Laplace

Trong đó:

* |docsi|: số văn bản của tập huấn luyện thuộc phân lớp i.
* |total docs|: số văn bản trong tập huấn luyện.
* m số phân lớp

Cài đặt:

* Khởi tạo mảng A, B có kích thước m.
* Duyệt qua các văn bản trong tập dữ liệu, đếm số văn bản trong mỗi phân lớp lưu vào A.
* Tính xác suất cho từng phân lớp theo công thức trên và lưu vào mảng B.

Công thức tính đã làm trơn Laplace:

Trong đó:

* : Số văn bản trong trong phân lớp i có đặc trưng thứ k mang giá trị xk. (hay số văn bản trong lớp i, có xuất hiện/không xuất hiện đặc trưng k)
* : Số văn bản của tập huấn luyện thuộc phân lớp i.
* Số giá trị có thể có của đặc trưng thứ k

Cài đặt:

* Với vector đặc trưng như mô tả bên trên, dk ở đây mang giá trị là 2, tương ứng với xuất hiện và không xuất hiện. Do chỉ có 2 giá trị, ta có thể tính nhanh xác suất không xuất hiện theo công thức
* Khởi tạo mảng 3 chiều C, chiều 1 có kích thước là m (số phân lớp), chiều 2 có kích thước là N (số đặc trưng), chiều 3 có kích là 2 (dk) để lưu các giá trị .
* Duyệt qua các văn bản trong tập dữ liệu, tiến hành thống kê các chỉ số cần thiết để tính xác suất theo công thức trên và lưu vào mảng C.

1. Phân lớp:

Đầu vào:

* Vector đặc trưng của văn bản cần phân lớp.
* Các giá trị xác suất và .

Đầu ra:

* Nhãn/lớp của văn bản cần phân loại.

Công thức tính xác suất thuộc phân lớp i khi biết trước mẫu X

Dựa vào vector đặc trưng của văn bản cần phân lớp, áp dụng công thức trên tính xác suất thuộc từng phân lớp cho văn bản, và chọn ra lớp có xác suất cao nhất.

Ví dụ:

|  |  |  |  |
| --- | --- | --- | --- |
| Docs | Var | Bit | Class |
| Doc1 | 1 | 0 | Math |
| Doc2 | 0 | 1 | Comp |
| Doc3 | 1 | 1 | Comp |
| Doc4 | 1 | 0 | Math |
| Doc5 | 0 | 1 | Math |
| Doc6 | 0 | 1 | Comp |

Docnew = (1,0)

Vậy văn bản mới thuộc lớp Math.

# Tài liệu tham khảo:

[1] Nguyễn Trần Thiên Thanh, Trần Khải Hoàng, *Tìm hiểu các hướng tiếp cận bài toán phân loại văn bản và xây dựng phần mềm phân loại tin tức báo điện tử*, Khóa luận cử nhân tin học 2005.

[2] Christopher D.Manning, Hinrich Schutze, *Foundations Of Statistical Natural Language Processing*.