+ Giải thuật Naive Bayes hoạt động như thế nào? Hãy giải thích định lý Bayes và giả định "ngây thơ" trong thuật toán này?

Giải thuật Naïve Bayes hoạt động dựa trên định lý Bayes là một phương pháp phân loại xác suất dựa trên công thức:

Trong đó:

* : Xác suất để mẫu x thuộc lớp y
* : Xác suất quan sát được đặc trưng x khi biết lớp y
* : Xác suất tiên nghiệm của lớp y
* : Xác suất quan sát được đặc trưng x (thường xem là hằng số khi so sánh giữa các lớp)

Giả định ngây thơ chính là cho rằng các đặc trưng xi là độc lập có điều kiện với nhau khi biết nhãn lớp y

P(x|y) = P(x1, x2, …,xn|y)

Từ đó tính xác suất của các lớp và chọn ra lớp có xác xuất cao nhất theo đặc trưng để quyết định kết quả

+ Các loại mô hình Naive Bayes (Gaussian, Multinomial, Bernoulli) khác nhau ra sao? Khi nào nên sử dụng từng loại?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mô hình | Loại dữ liệu | Giả định phân phối | Ví dụ | Đặc trưng đáng chú ý |
| |  | | --- | | Gaussian |  |  | | --- | |  | | Liên tục | Chuẩn | Cân nặng, chiều cao, nhiệt độ … | Giả định dữ liệu tuân theo phân phối chuẩn |
| Multinomial | Rời rạc | Đa thức | Phân loại văn bản, email spam | Dựa vào tần suất đặc trưng |
| Bernoulli | Nhị phân | Bernoulli | Phân loại có mua hoặc không mua từ | Không quan tâm tần suất, chỉ cần xuất hiện |

+ Tại sao Naive Bayes được gọi là "ngây thơ"? Giả định về tính độc lập của các đặc trưng ảnh hưởng như thế nào đến hiệu suất của mô hình?

Bởi vì Naive Bayes giả định rằng tất cả những feature đều độc lập có điều kiện với nhau, khi biết nhãn y. Tuy nhiên điều này nó không có đúng với thực tế chính vì sự đơn giản hóa này nên người ta gọi nó là “ngây thơ”

Ảnh hưởng đến hiệu suất của mô hình:

**Ưu điểm:**

Tốc độ cao, dễ huấn luyện:

+ Không cần ước lượng phân phối đồng thời của nhiều đặc trưng

+ Huấn luyện và dự đoán cực kỳ nhanh, ngay cả với tập dữ liệu lớn

Hoạt động tốt ngay cả khi giả định không hoàn toàn đúng:

+ Thực tế, nhiều đặc trưng có tương quan, nhưng mô hình vẫn đạt hiệu quả cao trong nhiều bài toán (Đặc biệt là NLP)

+ Nguyên nhân: dù sai về xác suất tuyệt đối, nhưng xếp hạng xác suất tương đối giữa các lớp vẫn đúng → vẫn phân loại chuẩn

**Nhược điểm:**

Hiệu suất giảm khi các đặc trưng phụ thuộc mạnh:

+ Đây là điều đương nhiên vì Bayes đưa ra các giả định “ngây thơ” rằng các đặc trưng không phụ thuộc nhau, nếu mà trên thực tế các đặc trưng thật sự có mối quan hệ phụ thuộc nhau thì điều này sẽ làm suy giảm đáng kể hiệu năng của mô hình

Không phù hợp cho dữ liệu phi tuyến hoặc phụ thuộc phức tạp:

+ Với dữ liệu có quan hệ giữa các đặc trưng mang tính phi tuyến (nonlinear), Naïve Bayes khó mô hình hóa đúng

+ Ưu điểm và hạn chế của Naive Bayes so với các thuật toán phân loại khác như SVM hoặc Random Forest là gì?

+ Viết đoạn code mẫu bằng Python (sử dụng Scikit-learn) để xây dựng một mô hình Naive Bayes (ví dụ: Gaussian Naive Bayes) không? Hãy mô tả các bước thực hiện

+ Làm thế nào để xử lý dữ liệu phân loại (categorical data) trước khi áp dụng Multinomial Naive Bayes trong Python?

+ Naive Bayes thường được sử dụng trong phân loại văn bản (text classification). Bạn có thể giải thích cách triển khai Naive Bayes cho bài toán này không?