**Khái niệm NB**

Naive Bayes Classification (NBC) là một phương pháp dựa trên định lý Bayes về xác suất, được sử dụng để đưa ra dự đoán và phân loại dữ liệu dựa trên thông tin quan sát và thống kê từ dữ liệu. NBC là một trong những thuật toán phổ biến trong lĩnh vực Machine Learning, được áp dụng rộng rãi để tạo ra các dự đoán chính xác từ tập dữ liệu thu thập được. Thuật toán này được đánh giá cao vì tính dễ hiểu và độ chính xác cao. NBC thuộc nhóm thuật toán Học có giám sát (Supervised Machine Learning Algorithms), nghĩa là thuật toán học từ các ví dụ được cung cấp từ các mẫu dữ liệu đã biết trước.

Định lý Bayes cho phép tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra. Xác suất này được ký hiệu là P(A|B). Đại lượng này được gọi xác suất có điều kiện hay xác suất hậu nghiệm vì được rút ra từ giá trị được cho của B hoặc phụ thuộc vào giá trị đó.

Theo định lí Bayes, P(A|B) sẽ phụ thuộc vào 3 yếu tố:

P(A∣B) là xác suất của sự kiện A xảy ra khi đã biết sự kiện B đã xảy ra.

P(B∣A) là xác suất của sự kiện B xảy ra khi đã biết sự kiện A đã xảy ra.

P(A) và P(B) là xác suất của sự kiện A và B xảy ra mà không cần biết về nhau

* + **Xác định các đặc trưng**

Đầu tiên, thuật toán cần xác định các đặc trưng hoặc thuộc tính của dữ liệu. Ví dụ, trong bài toán phân loại email, các đặc trưng có thể bao gồm từ vựng xuất hiện trong email, số lượng từ trong email, hoặc các đặc điểm của email như độ dài, tỷ lệ các ký tự in hoa, v.v.

* + **Tính toán xác suất của mỗi lớp**

Naive Bayes tính toán xác suất của mỗi lớp dựa trên các đặc trưng của mẫu dữ liệu. Đối với mỗi lớp, tính toán xác suất có điều kiện của mỗi đặc trưng dựa trên dữ liệu huấn luyện.

* + **Áp dụng định lý Bayes**

Sau khi tính toán xác suất có điều kiện cho mỗi lớp, Naive Bayes áp dụng định lý Bayes để tính toán xác suất của mỗi lớp dựa trên các đặc trưng của mẫu dữ liệu. Điều này giúp dự đoán xác suất của mỗi lớp cho mẫu dữ liệu mới.

* + **Chọn lớp có xác suất cao nhất**

Naive Bayes chọn lớp có xác suất cao nhất là lớp dự đoán cho mẫu dữ liệu. Xác định lớp mà mẫu dữ liệu có xác suất cao nhất thuộc vào và gán lớp đó cho dự đoán cuối cùng.

* + **Đưa ra dự đoán**

Cuối cùng, thuật toán đưa ra dự đoán bằng cách gán lớp có xác suất cao nhất cho mẫu dữ liệu. Điều này có nghĩa là xác định lớp mà mẫu dữ liệu được phân loại vào và đưa ra dự đoán tương ứng.

**ƯU ĐIỂM VÀ NHƯỢC ĐIỂM**

Ưu điểm:

1. Dễ triển khai và hiệu quả tính toán:

- Mô hình Naive Bayes là một thuật toán đơn giản và dễ triển khai.

- Việc tính toán xác suất trong mô hình này nhanh chóng và đơn giản, đặc biệt là trên các tập dữ liệu lớn.

2. Hoạt động tốt với dữ liệu lớn:

- Naive Bayes hoạt động tốt trên các tập dữ liệu lớn và có thể được mở rộng để xử lý dữ liệu lớn.

3. Khả năng xử lý các biến đầu vào lớn:

- Mô hình Naive Bayes hoạt động tốt với các biến đầu vào lớn mà không gặp vấn đề về hiệu suất.

4. Hiệu quả trong thời gian thực và phân loại dựa trên dữ liệu thưa:

- Naive Bayes thường hoạt động tốt với dữ liệu thưa (sparse data) và có thể được sử dụng trong thời gian thực.

**Nhược điểm:**

1. Giả định đơn giản:

- Giả định về sự độc lập giữa các biến đầu vào có thể là một giả định quá đơn giản và không phản ánh đúng thực tế trong một số tình huống.

2. Dễ bị ảnh hưởng bởi biến đầu vào không quan trọng:

- Mô hình Naive Bayes dễ bị ảnh hưởng bởi các biến đầu vào không quan trọng hoặc không có ảnh hưởng đến kết quả dự đoán.

3. Xác suất zero và cận zero:

- Trong trường hợp một từ hoặc một biến đầu vào không xuất hiện trong tập dữ liệu huấn luyện, mô hình Naive Bayes có thể dẫn đến xác suất zero hoặc gần zero, gây ra hiện tượng suy giảm hiệu suất dự đoán.

4. Không hiệu quả cho các vấn đề phân loại có mối quan hệ phức tạp giữa các biến đầu vào:

- Trong những trường hợp mối quan hệ giữa các biến đầu vào không phải là độc lập hoặc có mối quan hệ phức tạp, Naive Bayes có thể không hiệu quả.

Kết luận

Việc áp dụng Naive Bayes và TF-IDF đã cho thấy tính ưu việt trong việc xử lý và phân loại văn bản, đảm bảo độ tin cậy. khẳng định tiềm năng và ứng dụng thực tiễn của hệ thống trong lĩnh vực xử lý ngôn ngữ tự nhiên, đồng thời mở ra hướng nghiên cứu và cải tiến tiếp theo nhằm nâng cao hiệu suất và mở rộng phạm vi ứng dụng.

* Nắm được kiến thức cơ bản về hệ thống phân loại.
* Kiến thức về NLP, Machine Learning, TF-IDF, mô hình Naïve Bayes.
* Cách xử lý dữ liệu cũng như huấn luyện mô hình phân loại.