**LÝ THUYẾT Naive Beyes và TF-IDF**

1. **TF-IDF**

TF-IDF là một kỹ thuật được sử dụng để đo độ quan trọng của một từ trong một tập hợp các văn bản. TF-IDF tính toán giá trị của một từ bằng cách kết hợp hai yếu tố chính: tần suất của từ đó trong một văn bản (TF - Term Frequency) và tần suất nghịch đảo của từ đó trong toàn bộ các văn bản (IDF - Inverse Document Frequency). Cụ thể, giá trị TF-IDF của một từ trong một văn bản được tính như sau:

**TF-IDF(t,d,D)=TF(t,d)×IDF(t,D)**

* t là từ cần tính TF-IDF.
* d là văn bản cần tính TF-IDF
* D là tập hợp tất cả các văn bản.
* TF(t,d) là tần suất của từ t trong văn bản d.
* IDF(t,D) là tần suất nghịch đảo của từ t trong tập hợp các văn bản D.

1. **Mô hình Naive Bayes**

Mô hình Naive Bayes là một mô hình học máy dựa trên xác suất, được sử dụng phổ biến trong phân loại văn bản và phân loại dữ liệu. Mô hình này giả định "ngây thơ" (naive) rằng các đặc trưng (features) là độc lập với nhau khi biết lớp (class). Naive Bayes tính toán xác suất của một lớp cho một mẫu dữ liệu dựa trên xác suất của các đặc trưng khi biết lớp của mẫu dữ liệu đó.

1. **Mối liên hệ giữa TF-IDF và mô hình Naive Bayes**

TF-IDF thường được sử dụng để biểu diễn các văn bản dưới dạng các vector đặc trưng, trong đó mỗi thành phần của vector là giá trị TF-IDF của một từ trong từ điển. Mô hình Naive Bayes sử dụng các vector đặc trưng này để tính toán xác suất của các lớp cho mỗi mẫu dữ liệu, dựa trên giả định "ngây thơ" và xác suất của các từ được biểu diễn bằng giá trị TF-IDF. Do đó, trong bài toán phân loại văn bản, TF-IDF thường được sử dụng để biểu diễn văn bản và làm đầu vào cho mô hình Naive Bayes. TF-IDF giúp mô hình Naive Bayes nhận biết được mức độ quan trọng của từng từ trong việc phân loại văn bản.

1. **Biểu diễn dữ liệu dưới dạng vector đặc trưng**

Mỗi văn bản được biểu diễn dưới dạng một vector đặc trưng, trong đó mỗi phần tử của vector tương ứng với một từ trong từ điển. Giá trị của mỗi phần tử trong vector đặc trưng được tính bằng giá trị TF-IDF của từ đó trong văn bản.

1. **Tính toán xác suất của các lớp cho mỗi mẫu dữ liệu**

Mô hình Naive Bayes tính toán xác suất của mỗi lớp cho mỗi mẫu dữ liệu bằng cách sử dụng giả định "ngây thơ" và xác suất của các từ được biểu diễn bằng giá trị TF-IDF. Xác suất này được tính dựa trên xác suất tiên nghiệm của các lớp và xác suất của từng từ trong mỗi lớp. Cụ thể, trong mô hình Naive Bayes, xác suất của một mẫu dữ liệu

A black text on a white background

Description automatically generated𝑥 thuộc vào lớp 𝑦 được tính bằng công thức Bayes:

P(y) là xác suất tiên nghiệm của lớp 𝑦.

*P*(*xi*​∣*y*) là xác suất của từng tư Xi​ trong mẫu dữ liệu 𝑥 khi biết lớp 𝑦.

P(x) là một hằng số, xác suất tiên nghiệm của mẫu dữ liệu 𝑥.

1. **Phân loại dữ liệu**
2. **TF-IDF**

Tính toán giá trị TF-IDF của mỗi từ trong văn bản. TF-IDF chỉ đánh giá mức độ quan trọng của từ đó trong văn bản, dựa trên tần suất xuất hiện của từ trong văn bản và tần suất nghịch đảo của từ đó trong toàn bộ tập dữ liệu.

1. **Mô hình Naive Bayes**

Sử dụng giá trị TF-IDF được tính toán để biểu diễn văn bản dưới dạng vector đặc trưng. Dựa trên giả định "ngây thơ" và xác suất của các từ (được biểu diễn bằng giá trị TF-IDF), mô hình Naive Bayes tính toán xác suất của các lớp cho mỗi mẫu dữ liệu. Mô hình Naive Bayes quyết định phân loại mỗi mẫu dữ liệu vào lớp có xác suất cao nhất. trong quá trình phân loại, mô hình Naive Bayes sử dụng giá trị TF-IDF đã được tính toán để đưa ra quyết định phân loại dữ liệu vào các lớp tương ứng. TF-IDF giúp mô hình Naive Bayes hiểu được mức độ quan trọng của từng từ trong việc phân loại văn bản.