**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**



**KHOA ĐÀO TẠO CHẤT LƯỢNG CAO**

**🕯✡🕮🕮✡🕯**



**BÁO CÁO**

**TÌM HIỂU THUẬT TOÁN NAÏVE BAYES CLASSIFIER VÀ ÁP DỤNG VÀO BÀI TOÁN PHÂN LOẠI BÌNH LUẬN**

**GVHD: TS. Vũ Quang Huy**

**SVTH: MSSV**

**Nguyễn Đặng Nhật Minh 15110081**

**Lê Thị Phương Nha 15110090**

**Tp. Hồ Chí Minh, tháng 10 năm 2018**

# **LỜI NÓI ĐẦU**

Thống kê là bộ môn toán học rất quan trọng và có nhiều ứng dụng to lớn trong thực tế, giúp con người rút ra thông tin từ dữ liệu quan sát, nhằm giải quyết các bài toán thực tế trong cuộc sống.

Trong báo cáo này trình bày về một tiếp cận thống kê trong việc dự đoán sự kiện dựa vào lý thuyết Bayes. Lý thuyết này nói về việc tính xác suất của sự kiện dựa vào các kết quả thống kê các sự kiện trong quá khứ. Sau việc tính toán mỗi sự kiện được gán xác suất hay điểm (tùy vào mỗi phương án đánh giá) ứng với khả năng có thể xảy ra với sự kiện đó. Và cuối cùng dựa vào ngưỡng để phân loại cho các sự kiện.

Sau phần lý thuyết chúng ta sẽ tìm hiểu về bài toán thực tế trong ngành công nghệ thông tin. Bài toán về việc phân loại các bình luận của khách hàng. Giải quyết bài toán này ta dùng bộ lọc Bayes để nó nhận biết phân loại bình luận của khách hàng theo hướng tích cực và tiêu cực.

**MỤC LỤC**

[**CHƯƠNG 1**](#_Toc528104911) [**GIỚI THIỆU CHUNG** 4](#_Toc528104912)

[**1.1. Lý do chọn đề tài** 4](#_Toc528104913)

[**1.2. Mục đích của đề tài** 4](#_Toc528104914)

[**1.3. Ý nghĩa của đề tài** 4](#_Toc528104915)

[**CHƯƠNG 2**](#_Toc528104916) [**NAÏVE BAYES CLASSIFIER** 5](#_Toc528104917)

[**2.1. Tổng quan** 5](#_Toc528104918)

[**2.2. Định luật Bayes** 5](#_Toc528104919)

[**2.3. Ứng dụng** 10](#_Toc528104920)

[**KẾT LUẬN** 11](#_Toc528104921)

# **CHƯƠNG 1**

# **GIỚI THIỆU CHUNG**

## **1.1. Lý do chọn đề tài**

Đối với một người bán hàng hay khách hàng, việc để xem những bình luận tốt hay xấu của tất cả khách là một việc rất khó khăn và tốn rất nhiều thời gian. Vì vậy, việc áp dụng thuật toán Naïve Bayes Classifier vào bài toán này sẽ giúp chúng ta phân loại ra được những bình luận tốt hay chưa tốt.

## **1.2. Mục đích của đề tài**

Nội dung đề tài này định hướng việc phân loại thông tin của một tập lớn dữ liệu dựa trên phương pháp dự đoán hay gợi ý sử dụng thuật toán **Naïve Bayes**. Cụ thể, đề tài sẽ thực hiện phân loại bình luận tốt hay chưa tốt.

## **1.3. Ý nghĩa của đề tài**

Thông qua việc phân tích các thông tin thể hiện trong câu bình luận nhằm hỗ trợ cơ chế trong việc xây dựng hệ thống đánh giá, phân loại thông tin nhằm hạn chế bớt sự phức tạp khi xử lý ngôn ngữ tự nhiên.

# **CHƯƠNG 2**

# **NAÏVE BAYES CLASSIFIER**

## **2.1. Tổng quan**

Naive Bayes Classification (NBC) là một thuật toán dựa trên định lý Bayes về lý thuyết xác suất  để đưa ra các phán đoán cũng như phân loại dữ liệu dựa trên các dữ liệu được quan sát và thống kê. Naive Bayes Classification là một trong những thuật toán được ứng dụng rất nhiều trong các lĩnh vực Machine learning dùng để đưa các dự đoán chính xác nhất dự trên một tập dữ liệu đã được thu thập, vì nó khá dễ hiểu và độ chính xác cao. Nó thuộc vào nhóm Supervised Machine Learning Algorithms (thuật toán học có hướng dẫn), tức là máy học từ các ví dụ từ các mẫu dữ liệu đã có. Naive Bayes Classification (NBC) là một thuật toán dựa trên định lý Bayes về lý thuyết xác suất  để đưa ra các phán đoán cũng như phân loại dữ liệu dựa trên các dữ liệu được quan sát và thống kê. Naive Bayes Classification là một trong những thuật toán được ứng dụng rất nhiều trong các lĩnh vực Machine learning dùng để đưa các dự đoán chính xác nhất dự trên một tập dữ liệu đã được thu thập, vì nó khá dễ hiểu và độ chính xác cao. Nó thuộc vào nhóm Supervised Machine Learning Algorithms (thuật toán học có hướng dẫn), tức là máy học từ các ví dụ từ các mẫu dữ liệu đã có.

Ví dụ như ta có thể ứng dụng vào việc thiết kế một ứng dụng nghe nhạc có thể phán đoán được sở thích của nghe nhạc của người dùng dựa trên các hành vi như nhấn nút “thích” bài hát, “nghe đi nghe” lại nhiều lần các bài hát,  “bỏ qua” các bài hát không thích…. Dựa trên tập dữ liệu đó ta có thể áp dụng NBC để tính toán ra các phong cách nhạc mà người dùng thích nhất, từ đó chúng ta có thể đưa ra các “gợi ý” nghe nhạc gần đúng nhất cho người dùng từ việc học hỏi từ những thói quen đó.\

## **2.2. Định luật Bayes**

Định luật Bayes được phát biểu như sau:

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)**, và đọc là “xác suất của A nếu có B”. Đại lượng này được gọi xác suất có điều kiện hay xác suất hậu nghiệm vì nó được rút ra từ giá trị được cho của B hoặc phụ thuộc vào giá trị đó.

Theo định lí Bayes, xác suất xảy ra A khi biết B sẽ phụ thuộc vào 3 yếu tố:

\* Xác suất xảy ra A của riêng nó, không quan tâm đến B. Kí hiệu là **P(A)**và đọc là xác suất của A. Đây được gọi là xác suất biên duyên hay xác suất tiên nghiệm, nó là “tiên nghiệm” theo nghĩa rằng nó không quan tâm đến bất kỳ thông tin nào về B.

\* Xác suất xảy ra B của riêng nó, không quan tâm đến A. Kí hiệu là **P(B)**và đọc là “xác suất của B”. Đại lượng này còn gọi là hằng số chuẩn hóa (normalising constant), vì nó luôn giống nhau, không phụ thuộc vào sự kiện A đang muốn biết.

\* Xác suất xảy ra B khi biết A xảy ra. Kí hiệu là **P(B|A)** và đọc là “xác suất của B nếu có A”. Đại lượng này gọi là khả năng (likelihood) xảy ra B khi biết A đã xảy ra. Chú ý không nhầm lẫn giữa khả năng xảy ra B khi biết A và xác suất xảy ra A khi biết B.

Tóm lại định lý Bayes sẽ giúp ta tính ra xác suất xảy ra của một giả thuyết bằng cách thu thập các bằng chứng nhất quán hoặc không nhất quán với một giả thuyết nào đó. Khi các bằng chứng tích lũy, mức độ tin tưởng vào một giả thuyết thay đổi. Khi có đủ bằng chứng, mức độ tin tưởng này thường trở nên rất cao hoặc rất thấp, tức là xác xuất sảy ra giả thuyết sẽ thay đổi thì các bằng chứng liên quan đến nó thay đổi.

Công thức của định luật Bayes được phát biểu như sau:

Trong đó

– P(A|B) là  xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra.

– P(B|A) là xác suất xảy ra B khi biết A xảy ra

– P(A) là xác suất sảy ra của riêng A mà không quan tâm đến B.

– P(B) là xác suất xảy ra của riêng B mà không quan tâm đến A.

Ở trên ta có thể thấy xác suất sảy ra của giả thuyết A phụ thuộc và xác suất của giả thuyết B, nhưng trong thực tế xác suất A có thể phụ thuộc vào xác suất của nhiều các giác thuyết khác có thể là B1, B2, B3 … Bn. Vậy định luật Bayes có thể được mở rộng bằng công thức sau:

**Ví dụ**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Chills** | **Runny nose** | **Headache** | **Fever** | **Flu?** |
| Y | N | Mild | Y | N |
| Y | Y | No | N | Y |
| Y | N | Strong | Y | Y |
| N | Y | Mild | Y | Y |
| N | N | No | N | N |
| N | Y | Strong | Y | Y |
| N | Y | Strong | N | N |
| Y | Y | Mild | Y | Y |

Lớp: C1: Flu = “Y” (số người bị cúm)

C2: Flu = “N” (số người không bị cúm)

Chọn mẫu thử X= (Chills = “Y”, Runny nose = “N”, Headache = “Mild”, Fever = “N”)

**Tính P():**

P(C1) = = 0.625 (có 5 người bị cúm trong tổng 8 người khám)

P(C2) = = 0.375 (có 3 người không bị cúm trong tổng 8 người khám)

**Tính P(X|) cho mỗi lớp:**

P(Chills = “Y” | Flu = “Y”) = = 0.6 (có 3 người có triệu chứng ớn lạnh trong 5 người bị cúm)

P(Chills = “Y” | Flu = “N”) = = 0.33 (có 1 người có triệu chứng ớn lạnh trong 3 người không bị cúm)

P(Runny nose = “N” | Flu = “Y”) = = 0.2 (có 1 người không có triệu chứng sổ mũi trong 5 người bị cúm)

P(Runny nose = “N” | Flu = “N”) = = 0.66 (có 2 người không có triệu chứng sổ mũi trong 3 người không bị cúm)

P(Headache = “Mild” | Flu = “Y”) = = 0.4 (có 2 người có triệu chứng đau đầu nhẹ trong 5 người bị cúm)

P(Headache = “Mild” | Flu = “N”) = = 0.33 (có 1 người có triệu chứng đau đầu nhẹ trong 3 người không bị cúm)

P(Fever = “N” | Flu = “Y”) = = 0.2 (có 1 người không có triệu chứng sốt trong 5 người bị cúm)

P(Fever = “N” | Flu = “N”) = = 0.66 (có 2 người không có triệu chứng sốt trong 3 người không bị cúm)

**Tính P(X):**

P(Chills = “Y”) = = 0.5 (có 4 người bị ớn lạnh trên 8 người khám triệu chứng ớn lạnh)

P(Runny nose = “N”) = = 0.375 (có 3 người không bị sổ mũi trên 8 người khám triệu chứng sổ mũi)

P(Headache = “Mild”) = = 0.375 (có 3 người bị đau đầu nhẹ trên 8 người khám triệu chứng đau đầu nhẹ)

P(Fever = “N”) = = 0.375 (có 3 người không bị sốt trên 8 người khám triệu chứng sốt)

P(X) = (Xác suất của mẫu X)

= = 0.026

Xác suất của mẫu thử X có bị cúm:

P(X | Flu = “Y”) =

P(Headache = “Mild” | Flu = “Y”)

P(Fever = “N” | Flu = “Y”)

P(Flu = “Y”)

= 0.6 0.2 0.4 0.2 0.625 = 0.006

Xác suất của mẫu thử X không bị cúm:

P(X | Flu = “N”) =

P(Headache = “Mild” | Flu = “N”)

P(Fever = “N” | Flu = “N”)

P(Flu = “N”)

= 0.33 0.66 0.33 0.66 0.375 = 0.018

Vì P(X | Flu = “Y”) < P(X | Flu = “N”) nên mẫu X là không bị cúm.

Chúng ta có thể ứng dụng Naive Bayes Classification để tính tỷ lệ xác suất với rất nhiều các dạng bài toán khác nhau, với dữ liệu càng nhiều thì độ chính xác của thuật toán sẽ càng cao, và khi dữ liệu thay đổi thì kết quả cũng thay đổi theo.

**2.3. Ứng dụng**

Thuật toán Naive Bayes Classification được áp dụng vào các loại ứng dụng sau:

**Real time Prediction:** NBC chạy khá nhanh nên nó thích hợp áp dụng ứng dụng nhiều vào các  ứng dụng chạy thời gian thực, như hệ thống cảnh báo, các hệ thống trading …

**Multi class Prediction:**Nhờ vào định lý Bayes mở rộng ta có thể ứng dụng vào các loại ứng dụng đa dự đoán, tức là ứng dụng có thể dự đoán nhiều giả thuyết mục tiêu.

**Text classification/ Spam Filtering/ Sentiment Analysis:**NBC cũng rất thích hợp cho các hệ thống phân loại văn bản hay ngôn ngữ tự nhiên vì tính chính xác của nó lớn hơn các thuật toán khác. Ngoài ra các hệ thống chống thư rác cũng rất ưu chuộng thuật toán này. Và các hệ thống phân tích tâm lý thị trường cũng áp dụng NBC để tiến hành phân tích tâm lý người dùng ưu chuộng hay không ưu chuộng các loại sản phẩm nào từ việc phân tích các thói quen và hành động của khách hàng.

**Recommendation System:**Naive Bayes Classifier và Collaborative Filtering được sử dụng rất nhiều để xây dựng cả hệ thống gợi ý, ví dụ như xuất hiện các quảng cáo mà người dùng đang quan tâm nhiều nhất từ việc học hỏi thói quen sử dụng internet của người dùng, hoặc như ví dụ đầu bài viết đưa ra gợi ý các bài hát tiếp theo mà có vẻ người dùng sẽ thích trong một ứng dụng nghe nhạc …

# **KẾT LUẬN**

Naive Bayes cho kết quả tốt trong thực tế mặc dù chịu những giả thiết về tính độc lập có điều kiện (khi được cho nhãn/lớp) của các thuôc tính. Phân lớp không yêu cầu phải ước lượng một cách chính xác xác suất. Dễ cài đặt, học nhanh, kết quả dễ hiểu sử dụng trong phân loại text, spam, etc.Tuy nhiên khi dữ liệu có nhiều thuộc tính dư thừa thì naïve Bayes không còn hiệu quả dữ liệu liên tục có thể không tuân theo phân phối chuẩn (=> kernel density estimators).