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



**KHOA ĐÀO TẠO CHẤT LƯỢNG CAO**

**🕯✡🕮🕮✡🕯**



**BÁO CÁO**

**PHÂN LỚP NAIVE BAYES**

**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**

Khoa học thống kê đóng một vai trò cực kỳ quan trọng, một vai trò không thể thiếu được trong bất cứ công trình nghiên cứu khoa học, nhất là khoa học thực nghiệm như y khoa, sinh học, nông nghiệp, hóa học, và ngay cả xã hội học. Thí nghiệm dựa vào các phương pháp thống kê học có thể cung cấp cho khoa học những câu trả lời khách quan nhất cho những vấn đề khó khăn nhất.

Khoa học thống kê là khoa học về thu thập, phân tích, diễn giải và trình bày các dữ liệu để từ đó tìm ra bản chất và tính quy luật của các hiện tượng kinh tế, xã hội - tự nhiên. Khoa học thống kê dựa vào lý thuyết thống kê, một loại toán học ứng dụng. Trong lý thuyết thống kê, tính chất ngẫu nhiên và sự không chắc chắn có thể làm mô hình dựa vào lý thuyết xác suất. Vì mục đích của khoa học thống kê là để tạo ra thông tin "đúng nhất" theo dữ liệu có sẵn, có nhiều học giả nhìn khoa thống kê như một loại lý thuyết quyết định.

Thống kê là một trong những công cụ quản lý vĩ mô quan trọng, cung cấp các thông tin thống kê trung thực, khách quan, chính xác, đầy đủ, kịp thời trong việc đánh giá, dự báo tình hình, hoạch định chiến lược, chính sách, xây dựng kế hoạch phát triển kinh tế - xã hội và đáp ứng nhu cầu thông tin thống kê của các tổ chức, cá nhân. Trong số những vai trò quan trọng thì dự báo tình hình là một trong những vai trò mang nhiều ý nghĩa, nó có cả một quá trình huấn luyện bên trong và có tính xử lý tự động khi đã được huấn luyện. Hay nói khác hơn là khi đã có tri thức lấy từ các dữ liệu thống kê hay kinh nghiệm của người dùng kết hợp với một phương pháp học (huấn luyện) dựa trên lý thuyết thống kê ta sẽ có được một cỗ máy có tri thức để tự nó có thể đưa ra được những quyết định với độ chính xác khá cao.

Phân tích thống kê là một khâu quan trọng không thể thiếu được trong các công trình nghiên cứu khoa học, nhất là khoa học thực nghiệm. Một công trình nghiên cứu khoa học, cho dù có tốn kém và quan trọng cỡ nào, nếu không được phân tích đúng phương pháp sẽ không bao giờ có cơ hội được xuất hiện trong các tập san khoa học. Ngày nay, chỉ cần nhìn qua tất cả các tập san nghiên cứu khoa học trên thế giới, hầu như bất cứ bài báo y học nào cũng có phần “Statistical Analysis” (Phân tích thống kê), nơi mà tác giả phải mô tả cẩn thận phương pháp phân tích, tính toán như thế nào, và giải thích ngắn gọn tại sao sử dụng những phương pháp đó để hàm ý “bảo kê” hay tăng trọng lượng khoa học cho những phát biểu trong bài báo. Các tập san y học có uy tín càng cao yêu cầu về phân tích thống kê càng nặng. Không có phần phân tích thống kê, bài báo không thể xem là một “bài báo khoa học”. Không có phân tích thống kê, công trình nghiên cứu chưa được xem là hoàn tất.

Trong khoa học thống kê, có hai trường phái “cạnh tranh” song song với nhau, đó là trường phái tần số (frequentist school) và trường phái Bayes (Bayesian school). Phần lớn các phương pháp thống kê đang sử dụng ngày nay được phát triển từ trường phái tần số, nhưng hiện nay, trường phái Bayes đang trên đà “chinh phục” khoa học bằng một suy nghĩ “mới” về khoa học và suy luận khoa học. Phương pháp thống kê thuộc trường phái tần số thường đơn giản hơn các phương pháp thuộc trường phái Bayes. Có người từng ví von rằng những ai làm thống kê theo trường phái Bayes là những thiên tài!

Để hiểu sự khác biệt cơ bản giữa hai trường phái này, có lẽ cần phải nói đôi qua vài dòng về triết lý khoa học thống kê bằng một ví dụ về nghiên cứu y khoa. Để biết hai thuật điều trị có hiệu quả giống nhau hay không, nhà nghiên cứu phải thu thập dữ liệu trong hai nhóm bệnh nhân (một nhóm được điều trị bằng phương pháp A, và một nhóm được điều trị bằng phương pháp B). Trường phái tần số đặt câu hỏi rằng “nếu hai thuật điều trị có hiệu quả như nhau, xác suất mà dữ liệu quan sát là bao nhiêu”, nhưng trường phái Bayes hỏi khác: “Với dữ liệu quan sát được, xác suất mà thuật điều trị A có hiệu quả cao hơn thuật điều trị B là bao nhiêu”. Tuy hai cách hỏi thoạt đầu mới đọc qua thì chẳng có gì khác nhau, nhưng suy nghĩ kỹ chúng ta sẽ thấy đó là sự khác biệt mang tính triết lý khoa học và ý nghĩa của nó rất quan trọng. Đối với người bác sĩ (hay nhà khoa học nói chung), suy luận theo trường phái Bayes là rất tự nhiên, rất hợp với thực tế. Trong y khoa lâm sàng, người bác sĩ phải sử dụng kết quả xét nghiệm để phán đoán bệnh nhân mắc hay không mắc ung thư (cũng giống như trong nghiên cứu khoa học, chúng ta phải sử dụng số liệu để suy luận về khả năng của một giả thiết).

# **CHƯƠNG 1**

# **NAÏVE BAYES CLASSIFIER**

## **1.1. 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.\

**1.2. 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.

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).