TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**NGUYỄN THỊ THÚY VY - 232805404**

**HOÀNG ĐÌNH QUÝ VŨ - 521H0517**

**BÁO CÁO CUỐI KỲ**

**TOÁN CHO KHOA HỌC MÁY TÍNH**

Người hướng dẫn: **PhD. Võ Nguyễn Lê Duy**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**NGUYỄN THỊ THÚY VY - 232805404**

**HOÀNG ĐÌNH QUÝ VŨ - 521H0517**

**BÁO CÁO CUỐI KỲ**

**TOÁN CHO KHOA HỌC MÁY TÍNH**

Người hướng dẫn: **PhD. Võ Nguyễn Lê Duy**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025**

**LỜI CẢM ƠN**

Chúng em xin bày tỏ lòng biết ơn chân thành đến thầy Võ Nguyễn Lê Duy vì  
những kiến thức quý báu mà thầy đã truyền đạt và sự tận tâm trong việc giảng dạy  
môn Toán cho khoa học máy tính. Chúng em cảm nhận được sự chuyên nghiệp và đam mê của thầy trong việc truyền đạt tri thức, và chúng em rất biết ơn vì thầy đã dành thời gian và công sức để hướng dẫn chúng em trong quá trình học tập và tìm hiểu về lĩnh vực này.

Thầy đã truyền đạt những kiến thức sâu sắc và chi tiết về Toán cho khoa học máy tính, giúp chúng em hiểu rõ hơn về khung phát triển này và cách áp dụng vào thực tế. Nhờ những điều thầy đã truyền dạy, chúng em đã nắm vững cách xử lý dữ liệu một cách hiệu quả, đem lại kết quả tốt trong ứng dụng thực tế.

Chúng em cũng biết ơn vì sự quan tâm và hỗ trợ tận tình của thầy trong quá trình học tập. Thầy đã luôn sẵn sàng trả lời các câu hỏi của chúng em và giúp đỡ chúng em vượt qua những khó khăn trong quá trình nắm bắt kiến thức. Nhờ đó, chúng em đã có thêm niềm tin và động lực để tiếp tục khám phá và phát triển trong lĩnh vực Xử lý ảnh số.

Chúng em cảm nhận được sự chuyên nghiệp và đam mê của thầy trong việc  
giảng dạy. Sự cống hiến và tâm huyết của thầy đã giúp chúng em có được nền tảng  
vững chắc, đồng thời truyền cảm hứng để chúng em tiếp tục theo đuổi đam mê và  
ước mơ của bản thân.

Với tấm lòng biết ơn sâu sắc, chúng em xin kính chúc thầy Võ Nguyễn Lê Duy sức khỏe dồi dào, hạnh phúc và ngày càng thành công trong việc truyền tải tri thức và hỗ trợ sinh viên. Mong rằng những đóng góp của thầy sẽ tiếp tục lan tỏa và mang lại những thành tựu to lớn cho thầy và cả khoa Công nghệ thông tin.

*TP. Hồ Chí Minh, ngày 03 tháng 01 năm 2025*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*VY*

*Nguyễn Thị Thúy Vy*

*VŨ*

*Hoàng Đình Quý Vũ*

**CÔNG TRÌNH ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của TS. Võ Nguyễn Lê Duy. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong Dự án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung Dự án của mình**. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 03 tháng 01 năm 2025*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*VY*

*Nguyễn Thị Thúy Vy*

*VŨ*

*Hoàng Đình Quý Vũ*

# TÓM TẮT

Trong thời đại dữ liệu lớn (Big Data), việc khai thác và phân tích dữ liệu trở nên thiết yếu. Một trong những thách thức quan trọng của khoa học máy tính là xây dựng các mô hình phân loại và dự đoán hiệu quả.

Naive Bayes là một thuật toán đơn giản nhưng mạnh mẽ, dựa trên lý thuyết xác suất Bayes, được sử dụng rộng rãi trong các bài toán như phân loại văn bản, phát hiện thư rác, và nhận dạng cảm xúc.

# MỤC LỤC

[TÓM TẮT 6](#_Toc186874625)

[MỤC LỤC 7](#_Toc186874626)

[CHƯƠNG 1. PHƯƠNG PHÁP XỬ LÝ BÀI TOÁN 8](#_Toc186874627)

[1.1 Bài toán 8](#_Toc186874628)

[1.2 Cơ sở lý thuyết 8](#_Toc186874629)

[1.2.1 Mô hình phân lớp là gì? 8](#_Toc186874630)

[1.2.2 Mô hình phân lớp Naive Bayes 8](#_Toc186874631)

[1.2.2.1 Ví dụ cụ thể: 9](#_Toc186874632)

[1.2.3 Các biến thể của mô hình Naive Bayes: 10](#_Toc186874633)

[1.2.3.1 Gaussian Naive Bayes: 10](#_Toc186874634)

[1.2.3.2 Multinomial Naive Bayes: 13](#_Toc186874635)

[1.2.3.3 Bernoulli Naive Bayes 15](#_Toc186874636)

[CHƯƠNG 2. KẾT LUẬN 16](#_Toc186874637)

[TÀI LIỆU THAM KHẢO 17](#_Toc186874638)

# PHƯƠNG PHÁP XỬ LÝ BÀI TOÁN

## Bài toán

Trong thời đại dữ liệu lớn (Big Data), việc khai thác và phân tích dữ liệu trở nên thiết yếu. Một trong những thách thức quan trọng của khoa học máy tính là xây dựng các mô hình phân loại và dự đoán hiệu quả.

Naive Bayes là một thuật toán đơn giản nhưng mạnh mẽ, dựa trên lý thuyết xác suất Bayes, được sử dụng rộng rãi trong các bài toán như phân loại văn bản, phát hiện thư rác, và nhận dạng cảm xúc.

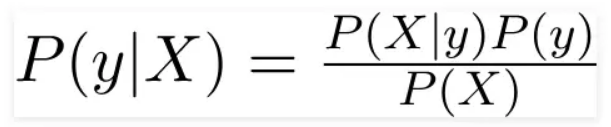
## Cơ sở lý thuyết

### Mô hình phân lớp là gì?

Một mô hình phân lớp là một mô hình Machine Learning dùng để phân loại các vật mẫu dựa trên các đặc tính đã xác định.

### Mô hình phân lớp Naive Bayes

Naive Bayes là một thuật toán phân lớp được mô hình hoá dựa trên định lý Bayes trong xác suất thống kê. Định lý Bayes cho phép tính xác suất xảy ra của một sự kiện ngẫu nhiên y khi biết sự kiện liên quan X đã xảy ra. Xác suất này được ký hiệu là P(y|X)

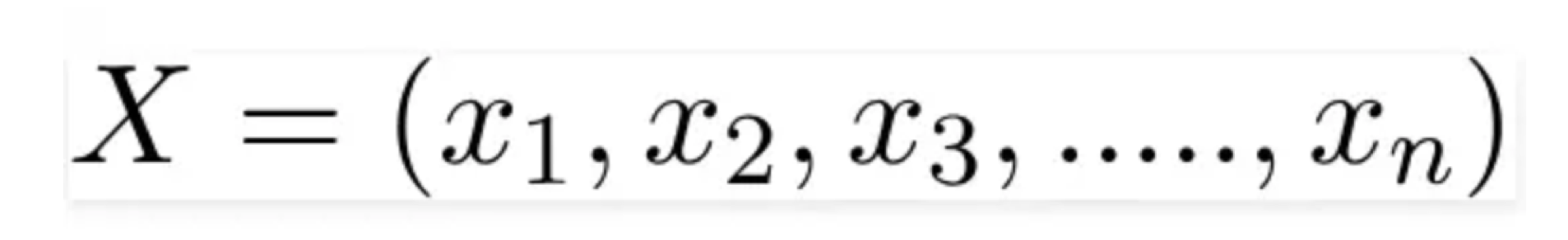


Trong đó:

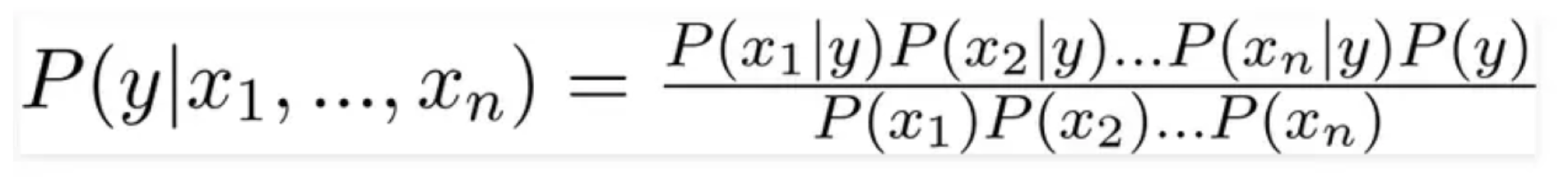
* P(y|X) gọi là xác suất của mục tiêu y với điều kiện có đặc trưng X
* P(X|y) gọi là xác suất của đặc trưng X khi đã biết mục tiêu y (likelihood)
* P(y) gọi là xác suất của mục tiêu y, không quan tâm đến X
* P(X) gọi là xác suất của đặc trưng X, không quan tâm đến y

Trong mô hình Naive Bayes, các đặc trưng đưa vào mô hình là độc lập với nhau. Tức là sự thay đổi giá trị của một đặc trưng không ảnh hưởng đến các đặc trưng còn lại.

X là vector các đặc trưng, có thể viết dưới dạng:

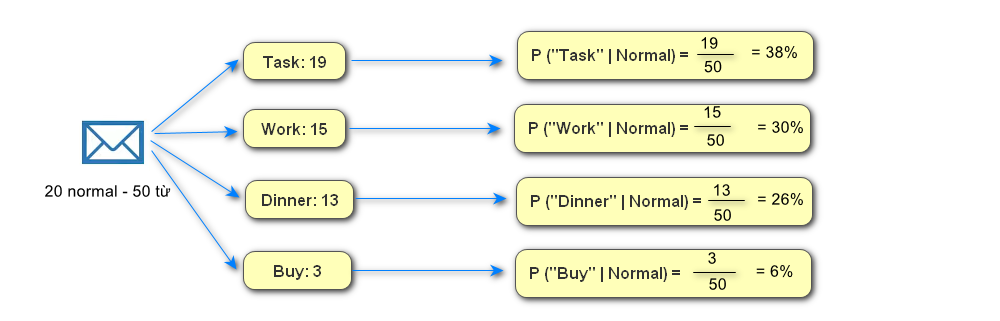


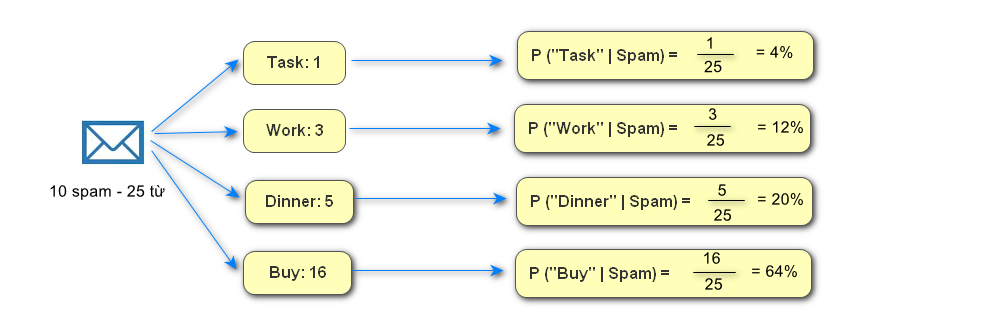
Khi đó, đẳng thức Bayes trở thành:



#### 1.2.2.1 Ví dụ cụ thể:

Ta có tập dữ liệu 20 email normal và 10 email Spam. Cả 2 phân loại có các từ như Task, Work, Dinner, Buy với số lượng như sau:





* P(Normal) = = 67%
* P(Spam) = = 33%

Câu hỏi đặt ra: Nếu nhận được email với nội dung “Task Dinner” thì email này là Normal hay Spam?

P(Normal | “Task Dinner”) =

=

= =

P(Spam | “Task Dinner”) =

=

= =

Kết quả: P(Normal | “Task Dinner”) > P(Spam | “Task Dinner”) nên email có nội dung “Task Dinner” là email **Normal**

### 1.2.3 Các biến thể của mô hình Naive Bayes:

Chúng được sử dụng để xử lý các loại dữ liệu khác nhau, tùy thuộc vào đặc điểm phân phối và cách biểu diễn dữ liệu. Các biến thể này khác nhau chủ yếu ở cách chúng giả định phân phối của các đặc trưng trong từng lớp.

#### 1.2.3.1 Gaussian Naive Bayes:

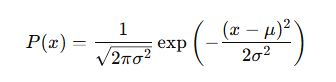
Giả định rằng các đặc trưng (features) tuân theo phân phối chuẩn (Gaussian distribution).

Mô hình phù hợp với dữ liệu liên tục (continuous data**)**

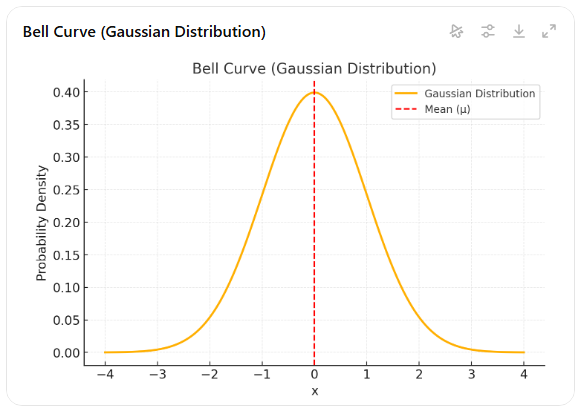
**Định nghĩa phân phối chuẩn:**

Phân phối chuẩn mô tả cách một biến ngẫu nhiên liên tục phân bố xung quanh giá trị trung bình (μ) với một độ lệch chuẩn (

**Công thức tổng quát:**



**Đồ thị:**

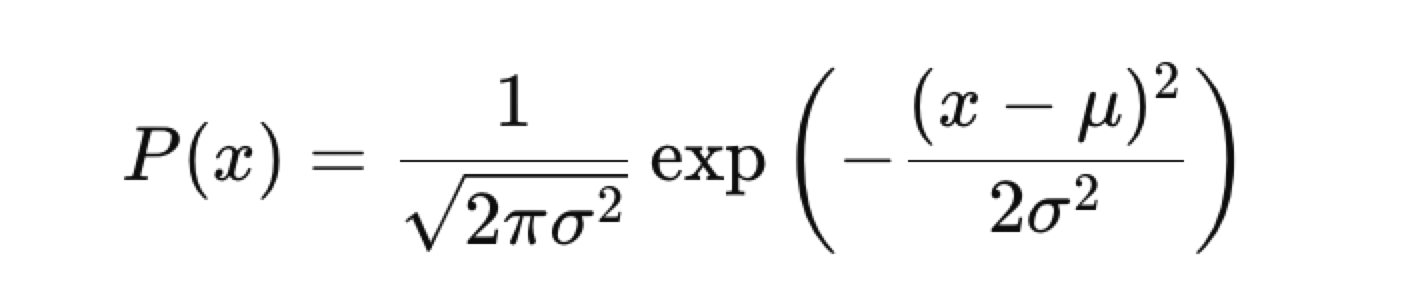
****

**Ví dụ:**

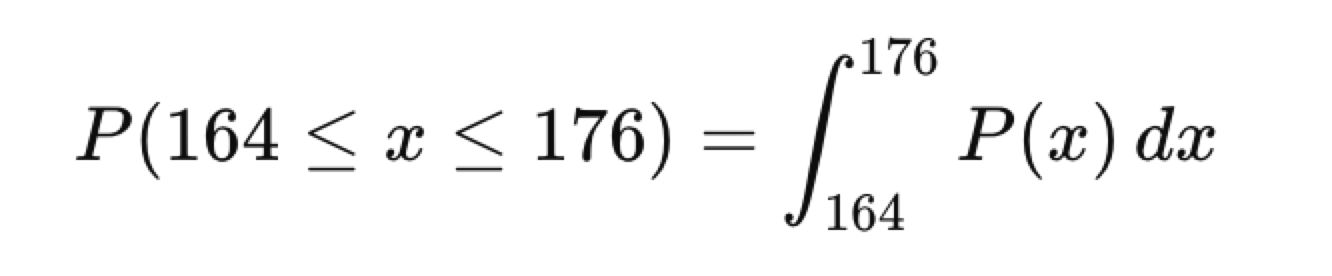
Chiều cao của học sinh trong một trường được giả định tuân theo phân phối chuẩn với: Trung bình (μ): 170 cm, Độ lệch chuẩn (σ): 6 cm.

Câu hỏi: Xác suất để một học sinh có chiều cao trong khoảng 164 cm đến 176 cm?

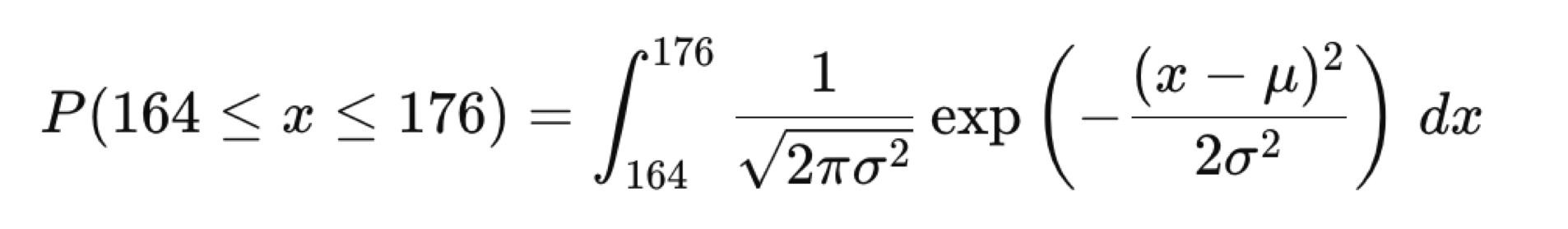
Ta có công thức xác suất phân phối chuẩn:

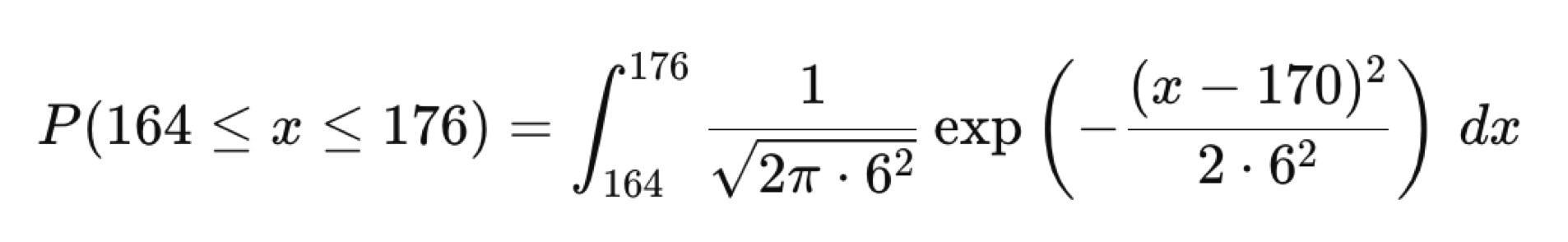


ta cần tính xác suất trong **một khoảng cụ thể** (ở đây là 164≤x≤176) bằng cách tính tích phân của P(x) trong khoảng này.



Thay P(x) vào biểu thức tích phân, ta có:





= 68.2 %

Đồ thị:

A diagram of a distribution curve

Description automatically generated

Kết luận:

* Trung bình (μ):
* **μ** xác định **vị trí trung tâm** của đường cong
* Đường cong luôn đối xứng quanh giá trị trung bình. Nếu thay đổi μ, toàn bộ đồ thị sẽ **tịnh tiến sang trái hoặc phải** mà không thay đổi hình dạng.
* Độ lệch chuẩn (σ): σ quyết định độ rộng và độ dốc của đường cong
* Nếu σ lớn, đường cong sẽ rộng hơn và thấp hơn (phẳng hơn).
* Nếu σ nhỏ, đường cong sẽ hẹp hơn và cao hơn (dốc hơn).

#### 1.2.3.2 Multinomial Naive Bayes:

Mô hình này chủ yếu được sử dụng trong phân loại văn bản mà feature vectors được tính bằng Bags of Words. Đặc trưng đầu vào ở đây chính là tần suất xuất hiện của từ trong văn bản đó.

Mô hình phù hợp với dữ liệu rời rạc (discrete data)

Kết hợp sử dụng phương pháp **Laplace Smoothing** để tránh xác suất bằng 0.

Ví dụ: Ta có tập dữ liệu các đoạn text đã được phân loại theo 2 nhóm là Sports và Not Sports theo bảng bên dưới. **Giả sử ta có đoạn text “A very close game” thì nó thuộc loại nào?**

|  |  |
| --- | --- |
| **Text** | **Category** |
| A Great Game | Sports |
| The Election was over | Not Sports |
| Very clean match | Sports |
| A clean but forgettable game | Sports |
| It was a close election | Not Sports |

Theo câu hỏi trên thì kết quả cần tìm là:

P(Not Sports | “a very close game”) và P(Sports | “a very close game”)

Lưu ý: P(close|Sports) =0 nên sử dụng **Laplace Smoothing** để xác suất khác 0

Xác suất kết hợp với Laplace Smoothing:

P(word | Category) =

trong đó V là tổng số từ.

Ta có bảng tính xác suất của từng từ “a very close game” như sau:

|  |  |  |
| --- | --- | --- |
| **Word** | **P(word | Sports)** | **P(word | Not Sports)** |
| a |  |  |
| very |  |  |
| close |  |  |
| game |  |  |

**P(Sports | “A very close game”)** =

=

=

**P(Not Sports | “A very close game”)** =

=

=

Ta thấy P(Sports | “A very close game”) sẽ cho kết quả cao hơn, nên “A very close game” sẽ thuộc category **Sports**.

#### 1.2.3.3 Bernoulli Naive Bayes

Bernoulli Naive Bayes là một biến thể của thuật toán Naive Bayes, được thiết kế đặc biệt cho dữ liệu nhị phân (0/1). Thuật toán này sử dụng Định lý Bayes để tính xác suất một mẫu thuộc về một lớp nhất định, với giả định rằng các đặc trưng trong dữ liệu là độc lập. Mỗi đặc trưng được biểu diễn dưới dạng nhị phân, đại diện cho sự có mặt hoặc vắng mặt của một yếu tố.

BNB thường được ứng dụng trong phân loại văn bản, ví dụ như phát hiện thư rác (spam) hoặc phân tích cảm xúc, nơi dữ liệu được biểu diễn bằng cách mã hóa các từ có xuất hiện (1) hoặc không (0) trong tài liệu. Điểm mạnh của nó là nhanh, hiệu quả với dữ liệu lớn, nhưng giả định độc lập giữa các đặc trưng có thể hạn chế độ chính xác trong một số trường hợp thực tế.



# CHƯƠNG 2. KẾT LUẬN

Naive Bayes là một thuật toán phân loại mạnh mẽ, đơn giản và hiệu quả, đặc biệt phù hợp với các bài toán có yêu cầu về tốc độ và khả năng xử lý dữ liệu lớn. Các biến thể của Naive Bayes mở rộng khả năng của thuật toán này để xử lý các loại dữ liệu và bài toán khác nhau.

Ưu điểm:

* **Đơn giản và nhanh chóng:**Với giả định độc lập có điều kiện, Naive Bayes có thời gian huấn luyện và suy luận nhanh.
* **Hiệu quả trên dữ liệu nhỏ:**Với một tập dữ liệu nhỏ, Naive Bayes thường hoạt động tốt và có thể khái quát hóa tốt.
* **Ứng dụng đa dạng:**Được sử dụng rộng rãi trong các bài toán phân loại văn bản (như lọc spam, phân loại cảm xúc) và nhận dạng mẫu.

Hạn chế:

* **Giả định độc lập có điều kiện:**Đây là một giả định mạnh mẽ, và trong thực tế, các đặc trưng thường không hoàn toàn độc lập. Điều này có thể làm giảm độ chính xác của mô hình trong một số bài toán.
* **Khả năng xử lý dữ liệu phức tạp:**Naive Bayes không hoạt động tốt khi dữ liệu không thể biểu diễn tốt bằng các phân phối mà các biến thể của nó hỗ trợ (như Gaussian, Multinomial, Bernoulli).

# TÀI LIỆU THAM KHẢO

<https://www.youtube.com/watch?v=rdJ_CMYLiqY>

<https://viblo.asia/p/mo-hinh-phan-lop-naive-bayes-vyDZO0A7lwj>

<https://machinelearningcoban.com/2017/08/08/nbc/>

<https://towardsdatascience.com/gaussian-naive-bayes-explained-a-visual-guide-with-code-examples-for-beginners-04949cef383c>