**TRƯỜNG ĐẠI HỌC XÂY DỰNG HÀ NỘI**

**KHOA CÔNG NGHỆ THÔNG TIN**

Icon

Description automatically generated with medium confidence

**MÔN HỌC: XỬ LÝ NGÔN NGỮ TỰ NHIÊN**

**TIỂU LUẬN CUỐI KÌ**

**CHƯƠNG 6: PHÂN LOẠI VĂN BẢN**

Nhóm: 3

Giảng Viên: Phạm Hồng Phong

Sinh viên thực hiện:

1. Hoàng Hồng Cẩm (18365)
2. Vũ Thị Thùy Giang (1515865)
3. Phí Hoàng Long (1532064)
4. Nguyễn Ngọc Huy (1654065)

*Hà Nội, 14 tháng 03 năm 2023*

Mục Lục

[Mục Lục 2](#_heading=h.gjdgxs)

[Mục lục hình ảnh 4](#_heading=h.30j0zll)

[Mục lục bảng biểu 4](#_heading=h.1fob9te)

[Tài liệu: 4](#_heading=h.3znysh7)

[**I.**](#_heading=h.2et92p0) **Phân loại được giám sát** 5

[**1.**](#_heading=h.3dy6vkm) **Xác định giới tính** 6

[**2.**](#_heading=h.1t3h5sf) **Chọn tính năng phù hợp** 8

[**3.**](#_heading=h.2s8eyo1) **Phân loại tài liệu** 11

[**4.**](#_heading=h.17dp8vu) **Gắn thẻ một phần của bài phát biểu** 13

[**5.**](#_heading=h.3rdcrjn) **Khái thác ngữ cảnh** 15

[**6.**](#_heading=h.26in1rg) **Phân loại trình tự** 16

[**7.**](#_heading=h.lnxbz9) **Phương pháp để phân loại trình tự** 17

[**II.**](#_heading=h.35nkun2) **Các ví dụ về phân loại có giám sát** 18

[**1.**](#_heading=h.1ksv4uv) **Phân loại câu** 18

[**2.**](#_heading=h.44sinio) **Xác định các hành động đối thoại** 19

[**3.**](#_heading=h.2jxsxqh) **Nhận biết ràng buộc văn bản** 20

[**4.**](#_heading=h.z337ya) **Mở rộng quy mô dữ liệu lớn** 22

[**III.**](#_heading=h.3j2qqm3) **Đánh giá** 23

[**1.**](#_heading=h.4i7ojhp) **Bộ kiểm tra** 23

[**2.**](#_heading=h.2xcytpi) **Độ chính xác** 23

[**3.**](#_heading=h.1ci93xb) **Độ chính xác và thu hồi** 24

[**4.**](#_heading=h.2bn6wsx) **Ma trận nhầm lẫn** 25

[**5.**](#_heading=h.qsh70q) **Xác thực chéo** 27

[**IV.**](#_heading=h.3as4poj) **Cây quyết định** 28

[**1. Entropy và thu được thông tin** 28](#_heading=h.2p2csry)

[**V.**](#_heading=h.3o7alnk) **Bộ phân loại Naive Bayes** 29

[**1.**](#_heading=h.32hioqz) **Mô hình xác suất cơ bản** 30

[**2.**](#_heading=h.41mghml) **Đếm zero và làm mịn** 31

[**3.**](#_heading=h.2grqrue) **Tính năng phi nhị phân** 31

[**4.**](#_heading=h.vx1227) **The Naivete of Independence** 31

[**5.**](#_heading=h.3fwokq0) **Nguyên nhân của việc đếm hai lần** 32

[**VI.**](#_heading=h.19c6y18) **Maximum entropy classifiers** 33

[**1.**](#_heading=h.28h4qwu) **The maximum entropy model** 33

[**2.**](#_heading=h.nmf14n) **Tối đa hóa Entropy** 33

[**VII.**](#_heading=h.2lwamvv) **Mô hình hóa các mẫu ngôn ngữ** 35

[**VIII.**](#_heading=h.111kx3o) **Bài tập** 35

# Mục lục hình ảnh

[Hình 6. 1: Mô hình phân loại có giám sát 5](#_heading=h.tyjcwt)

[Hình 6. 2: Tổ chức kho dữ liệu để huấn luyện bộ phân loại có giá sát 8](#_heading=h.4d34og8)

[Hình 6. 3: Thu thập tài liệu 24](#_heading=h.3whwml4)

[Hình 6. 4: Mô hình cây quyết định cho nhiệm vụ phân loại giới tính của tên 28](#_heading=h.49x2ik5)

[Hình 6. 5: Minh họa quy trình sử dụng bộ phân loại Naive Bayes 29](#_heading=h.23ckvvd)

[Hình 6. 6: Tính khả năng nhãn với Naive Bayes 30](#_heading=h.ihv636)

[Hình 6. 7: Biểu đồ mạng Bayes minh họa quá trình tổng quát được giả định bởi Naive Bayes 30](#_heading=h.1hmsyys)

# Mục lục bảng biểu

[Bảng 6. 1: Tỉ lệ chọn nghĩa chính xác của một từ nhất định 33](#_heading=h.37m2jsg)

[Bảng 6. 2: Tỉ lệ A xuất hiện 34](#_heading=h.1mrcu09)

[Bảng 6. 3: kết quả 34](#_heading=h.46r0co2)

# Tài liệu:

Link tài liệu sách NLTK – chương 6: <https://www.nltk.org/book/ch06.html>

Link code demo: <https://colab.research.google.com/drive/1yjo7pBREI3TDJHisVX2S8FheHcWMLfxL?usp=sharing>

Link bài tập demo:

<https://colab.research.google.com/drive/1Ktt3qYADvUkM1PDdR524ZoRgy-UeDadJ?usp=sharing>

1. **Phân loại được giám sát**

Phân loại là nhiệm vụ chọn nhãn lớp chính xác cho một đầu vào nhất định. Mỗi đầu vào được xem xét tách biệt với tất cả các đầu vào khác và bộ nhẫn được xác định trước.

⬥ Ví dụ về nhiệm vụ phân loại

✔ Quyết định xem một email có phải là thư rác hay không.

✔ Quyết định chủ đề của một bài báo là gì.

⬥ Một số biến thể của nhiệm vụ phân loại:

✔ Trong phân loại nhiều lớp, mỗi trường hợp có thể được gán nhiều nhãn

✔ Trong phân loại mở, bộ nhãn không được xác định trước

✔ Trong phân loại trình tự, một danh sách các đầu vào được phân loại chung

Phân loại có giám sát: là bộ phân loại được xây dựng dựa trên tập dữ liệu huấn luyện chứa nhãn chính xác cho một đầu vào.

Diagram

Description automatically generated

Hình 6. 1 Mô hình phân loại có giám sát

1. ***Xác định giới tính***

Xây dựng bộ phân loại để xác định giới tính qua tên:

⬥ Quyết định các tính năng của đầu vào có liên quan, cách mã hóa tính năng đó

✔ Hàm trích xuất tính năng: xây dựng từ điển chứa thông tin liên quan đến tên đã cho: A picture containing graphical user interface

Description automatically generated

✔ Trả về từ điển- bộ tính năng: ánh xạ tên tính năng sang giá trị của chúng.

✔ Hầu hết các tính năng được mã hóa bằng các loại giá trị đơn giản: boolean, số, chuỗi.

⬥ Chuẩn bị một danh sách các ví dụ, nhãn lớp tương ứng để xác định trình trích xuất tính năng

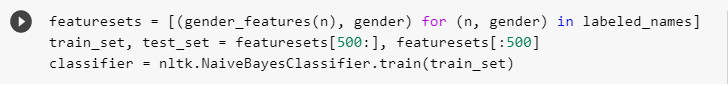
Graphical user interface, text

Description automatically generated with medium confidence

⬥ Dùng trình trích xuất tính năng xử lý dữ liệu, chia danh sách kết quả các bộ tính năng thành:

✔ Tập huấn luyện – huấn luyện bộ phân loại Naïve Bayes

✔ Tập kiểm tra



⬥ Test độ chính xác:

Graphical user interface, text, application, Word

Description automatically generated

⬥ Kiểm tra bộ phân loại để xác định tính năng nào hiệu quả nhất để phân biệt giới tính của tên

Table

Description automatically generated with medium confidence

* Sửa đổi hàm gender\_features() để cung cấp cho bộ phân loại các tính năng mã hóa độ dài của tên, chữ cái đầu tiên của tên và bất kỳ tính năng nào khác có vẻ như chúng có thể cung cấp thông tin. Đào tạo lại trình phân loại với các tính năng mới này và kiểm tra độ chính xác của nó.

Khi làm việc với kho dữ liệu lớn , phải dùng đến một lượng lớn bộ nhớ, hãy dùng hàm *nltk.classify.apply\_features* để trả về một đối tượng hoạt động giống như một danh sách nhưng không lưu trữ tất cả các bộ tính năng trong bộ nhớ.

1. ***Chọn tính năng phù hợp***

Trình trích xuất đặc trưng:

⬥ Là quá trình chuyển đổi dữ liệu thô thành các đặc trưng số có thể được xử lý trong khi vẫn bảo toàn thông tin trong tập dữ liệu gốc

⬥ Trình trích xuất đặc trưng được xây dựng thông quá trình thử và sai

✔ Kiểm tra xem tính năng nào thực sự hữu ích

Text

Description automatically generated

✔ Trình trích xuất tính năng làm khớp các tính năng giới tính. Các bộ tính năng chứa số lượng lớn các tính năng cụ thể. Tuy nhiên số lượng tính năng có giới hạn nên sử dụng với một thuật toán học tập nhất định.

✔ Lưu ý không nên cung cấp quá nhiều tính năng cho thuật toán dẫn đến vấn đề **trang bị quá mức** các thuật toán sẽ dựa vào đặc điểm riêng của dữ liệu mà không khái quát hóa tốt.

✔ Dùng phương pháp phân tích lỗi để tinh chỉnh bộ tính năng ban đầu.

Làm thế nào để chọn tính năng phù hợp?

⬥ Chọn một bộ phát triển chứa dữ liệu văn bản để tạo mô hình   
⬥ Chia nhỏ bộ phát triển thành tập huấn luyện và tập phát triển- kiểm tra.

Diagram

Description automatically generated with medium confidence

Hình 6. 2 Tổ chức kho dữ liệu để huấn luyện bộ phân loại có giá sát

⬥ Mô hình được đào tạo bằng cách sử dụng tập huấn luyện, sau đó chạy nó trên tập thử nghiệm dành cho nhà phát triển

Text, letter

Description automatically generated

⬥ Một cách để tinh chỉnh tính năng là phân tích lỗi. Chúng ta sẽ tạo một tập phát triển, bao gồm tập huấn luyện và tập dev-test set. Dev-test set phải được tách biệt khỏi bộ thử nghiệm:

Text, application

Description automatically generated

⬥ Tạo danh sách các lỗi mà bộ phân loại mắc phải khi dự đoán giới tính của tên bằng tập phát triển- kiểm tra:

Text

Description automatically generated

Table

Description automatically generated

* Danh sách lỗi cho thấy một số hậu tố nhiều hơn một chữ cái có thể biểu thị giới tính của tên

⬥ Xây dựng lại bộ phân loại bằng trình trích xuất tính năng mới

* Quy trình phân tích lỗi có thể được lặp lại nên chọn một bộ phân chia kiểm tra/ huấn luyện khác nhau để đảm bảo rằng bộ phân loại không bắt đầu phản ánh các đặc điểm riêng trong tập kiểm tra phát triển
* Phải giữ bộ kiểm tra riêng biệt và không được sử dụng cho đến khi quá trình phát triển mô hình hòa tất. Sau đó sử dụng bộ kiểm tra để đánh giá mô hình hoặt động tốt như thế nào trên các giá trị đầu vào mới

1. ***Phân loại tài liệu***

❖ Sử dụng kho văn bản trong đó các tài liệu đã được gắn nhãn với các danh mục để xây dựng các bộ phân loại tự động gắn thẻ các tài liệu mới với các nhãn dán phù hợp.

⬥ Xây dựng một danh sách với các tài liệu được gán nhãn thích hợp

Text

Description automatically generated

⬥ Xác định trình trích xuất tính năng cho các tài liệu

✔ Xác định chủ đề: xác định tính năng cho mỗi từ, cho biết tài liệu có chứa từ đó không  
✔ Giới hạn số lượng tính năng:

• Xây dựng danh sách từ thường gặp trong văn bản tổng thể

• Định nghĩa trình trích xuất tính năng

Graphical user interface, text, application

Description automatically generated

⬥ Sử dụng trình trích xuất tính năng đào tạo một trình phân loại để gán nhãn các bài đánh giá mới

Graphical user interface, text, application

Description automatically generated

⬥ Tính toán độ chính xác của bộ phân loại kết quả trên tập kiểm tra để kiểm tra độ tin cậy

A picture containing text

Description automatically generated

## ***Gắn thẻ một phần của bài phát biểu***

Xem cấu tạo bên trong của từ để xây dựng trình gắn thẻ biểu thức chính quy chọn thẻ một phần lời nói cho một từ

⬥ Huấn luyện bộ phân loại tìm ra hậu tố có nhiều thông tin nhất

Graphical user interface, text, application

Description automatically generated

⬥ Định nghĩa hàm trích xuất đặc trưng để kiểm tra một từ đã cho cho các hậu tố

Text

Description automatically generated

⬥ Xác định trình trích xuất đặc trưng – để huấn luyện một trình phân loại cây quyết định mới

Graphical user interface, text

Description automatically generated

* Bộ phân loại bắt đầu bằng cách kiểm tra xem từ có kết thúc bằng dấu phẩy/the/s không sau đó bắt đầu phân loại

In cây quyết định:

Text

Description automatically generated

## ***Khái thác ngữ cảnh***

Để phù hợp với tính năng phụ thuộc vào ngữ cảnh của một từ, sửa lại mẫu đã sử dụng để xác định trình trích xuất tính năng. Thay vì chuyển vào từ cần được gắn thẻ , hãy chuyển vào một câu hoàn chỉnh (không được gắn thẻ), cùng với chỉ mục của từ mục tiêu

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

* Việc khai thác tính năng theo ngữ cảnh sẽ cải thiện hiệu suất của trình gắn thẻ phần lời nói

## ***Phân loại trình tự***

Các mô hình phân loại chung chọn một nhãn thích hợp cho một tập hợp các đầu vào có liên quan. Với việc gắn thẻ một phần lời nói, một số mô hình phân loại trình tự khác nhau có thể được sử dụng để cùng chọn các thẻ một phần lời nói cho các từ trong một câu nhất định.

Phân loại liên tiếp (còn gọi là phân loại trình tự tham lam) tìm nhãn lớp có khả năng nhất cho đầu vào đầu tiên, sau đó sử dụng nhãn này để tìm nhãn tốt nhất cho đầu vào tiếp theo. Quá trình này được lặp lại cho đến khi tất cả các đầu vào đã được dán nhãn.

Chúng tôi có thể cập nhật pos\_features để lưu trữ các thẻ của từ trước đó và từ đó xây dựng trình phân loại trình tự:

***Text

Description automatically generated with medium confidence***

***Text

Description automatically generated***

***Text

Description automatically generated***

***Graphical user interface, text

Description automatically generated***

## ***Phương pháp để phân loại trình tự***

Gán điểm cho tất cả các trình tự có thể của phần gắn thẻ lời nói và chọn trình tự có tổng điểm cao nhất. tuy nhiên phương pháp này có nhiều hạn chế

# **Các ví dụ về phân loại có giám sát**

## ***Phân loại câu***

Phân đoạn câu có thể được xem như một nhiệm vụ phân loại dấu câu: bất cứ khi nào chúng ta gặp một biểu tượng có thể kết thúc câu, chẳng hạn như dấu chấm hoặc dấu chấm hỏi, chúng ta phải quyết định xem nó có kết thúc câu trước đó hay không.

⬥ Lấy một số dữ liệu đã được phân đoạn thành các câu và chuyển đổi nó thành dạng phù hợp để trích xuất các tính năng

Graphical user interface, text, application

Description automatically generated

⬥ Chỉ định các tính năng của dữ liệu sẽ được sử dụng để quyết định xem dấu chấm câu có biểu thị ranh giới câu hay không

Graphical user interface, text, application

Description automatically generated

⬥ Tạo danh sách các bộ tính năng được gắn nhãn bằng cách chọn tất cả các mã thông báo dấu chấm câu và gắn thẻ xem chúng có phải là mã thông báo ranh giới hay không

Graphical user interface, text, application

Description automatically generated

⬥ Huấn luyện và đánh giá bộ phân loại dấu chấm câu

Text

Description automatically generated with medium confidence

⬥ Để sử dụng bộ phân loại này để thực hiện phân đoạn câu, cần kiểm tra từng dấu chấm câu để xem liệu nó có được gắn nhãn là ranh giới hay không; và chia danh sách các từ tại các dấu ranh giới.

***Graphical user interface, text, application, email

Description automatically generated***

## ***Xác định các hành động đối thoại***

Sử dụng dữ liệu để xây dựng bộ phân loại có thể xác định các loại hành động đối thoại cho các bài đăng nhắn tin tức thời mới

⬥ Trích xuất dữ liệu tin nhắn cơ bản



⬥ Xác định trình trích xuất tính năng đơn giản để kiểm tra xem bài đăng chứa những từ nào

Text

Description automatically generated

⬥Xây dựng dữ liệu đào tạo và thử ngiệm bằng cách áp dụng trình trích xuất tính năng cho từng bài đăng và tạo một bộ phân loại mới

Graphical user interface, text

Description automatically generated

## ***Nhận biết ràng buộc văn bản***

❖ Sự kéo theo văn bản (RTE) là nhiệm vụ xác định xem một đoạn văn bản *T* nhất định có kéo theo một văn bản khác được gọi là "giả thuyết"

❖ Coi RTE như một nhiệm vụ phân loại, trong đó chúng ta cố gắng dự đoán nhãn *Đúng* / *Sai* cho mỗi cặp. Nếu có thông tin được tìm thấy trong giả thuyết mà không có trong văn bản, thì sẽ không có sự kéo theo, ngược lại

❖ Các từ (nghĩa là các loại từ) đóng vai trò là đại diện cho thông tin và các tính năng của chúng tôi đếm mức độ chồng chéo từ và mức độ có các từ trong giả thuyết nhưng không có trong văn bản

⬥ Trình trích xuất tính năng nhận dạng ràng buộc văn bản

Graphical user interface, text, application

Description automatically generated

* Các từ quan trọng trong giả thuyết có trong văn bản

Test:

Graphical user interface, application

Description automatically generated with medium confidence

## ***Mở rộng quy mô dữ liệu lớn***

Nên khám phá các cơ sở của NLTK để giao tiếp với các gói máy học bên ngoài để đào tạo bộ phân loại với lượng lớn dữ liệu đào tạo hoặc nhiều tính năng

1. ***Đánh giá***

Để quyết định xem một mô hình phân loại có nắm bắt chính xác một mẫu hay không, phải đánh giá mô hình đó

1. ***Bộ kiểm tra***

❖ Các kỹ thuật đánh giá tính điểm cho một mô hình bằng cách so sánh các nhãn mà nó tạo ra cho các đầu vào trong tập kiểm tra(hoặc tập đánh giá) với các nhãn chính xác cho các đầu vào đó

❖ Tập kiểm tra thường cùng định dạng với tập huấn luyện nhưng không được là tập là tập huấn luyện đó

❖ Xây dựng tập kiểm tra, có sự đánh đổi giữa dữ liệu có sẵn để kiểm tra và dữ liệu có sẵn để huấn luyện

⬥ Lưu ý : nếu tác vụ phân loại có số lượng nhãn lớn hoặc bao gồm các nhãn hiếm gặp , nên chọn kích thước của tập kiểm tra đảm bảo nhãn thường xuyên nhất

❖ Khi các thể hiện trong bộ thử nghiệm và các thể hiện trong bộ phát triển càng giống nhau thì kết quả đánh giá tổng quát cảng giảm cho các bộ dữ liệu khác

1. ***Độ chính xác***

❖ Số liệu đơn giản nhất có thể được sử dụng để đánh giá bộ phân loại, độ chính xác, đo lường tỉ lệ phần trăm đầu vào trong tập kiểm tra mà bộ phân loại đã dán nhãn chính xác

❖ Khi diễn giải điểm chính xác của một bộ phân loại, phải xem xét tần số của các nhãn lớp riêng lẻ trong bộ kiểm tra

1. ***Độ chính xác và thu hồi***

❖ Vì số lượng tài liệu không liên quan lớn hơn nhiều so với số lượng tài liệu liên quan nên điểm chính xác có thể gây nhầm lẫn

Diagram

Description automatically generated

Hình 6. 3 Thu thập tài liệu

❖ Sử dụng tập hợp các biện pháp khác nhau cho các nhiệm vụ tìm kiếm, dựa trên số lượng mục đích trong mỗi loại được hiển thị ở hình trên

⬥ True positives : tài liệu có liên quan được xác định có liên quan

⬥ True negatives : tài liệu không liên quan được xác định là không liên quan

⬥ False positives : tài liệu không liên quan được xác định không chính xác là có liên quan

⬥ False negatives : tài liệu có liên quan được xác định không chính xác là không liên quan

❖ Từ đó ta xác định được các số liệu

⬥ Độ chính xác

A picture containing text

Description automatically generated

⬥ Recall

A picture containing text

Description automatically generated

⬥ F1-score

Text, letter

Description automatically generated

1. ***Ma trận nhầm lẫn***

❖ Ma trận nhầm lẫn là một bảng trong số đó mỗi ô [i, j] cho biết tần suất nhãn j đã được dự đoán chính xác là i.

* Các mục tiêu nằm trên đường chéo biểu thị nhãn đã được dự đoán chính xác, các mục tiêu nằm ngoài đường chéo biểu thị lỗi

Text

Description automatically generated

Table

Description automatically generated

1. ***Xác thực chéo***

❖ Một phần dữ liệu chú thích được dành ra cho bộ thử nghiệm nhằm đánh giá các mô hình

❖ Xác nhận chéo: Để tập kiểm tra không nhỏ hơn hay lớn hơn tập huấn luyện nên thực hiện đánh giá trên các bộ kiểm tra khác nhau, sau đó kết hợp từ những đánh giá đó

⬥ Sử dụng xác thực chéo cho phép kiểm tra mức độ thay đổi của hiệu suất giữa các tập huấn luyện khác nhau

1. **Cây quyết định**

Cây quyết định là một lưu đồ đơn giản chọn nhãn cho các giá trị đầu vào. Lưu đồ này bao gồm các nút quyết định kiểm tra các giá trị tính năng và các nút lá gán nhãn

Diagram

Description automatically generated

Hình 6. 4 Mô hình cây quyết định cho nhiệm vụ phân loại giới tính của tên

Xây dựng một gốc quyết định

▪ Xây dựng gốc quyết định cho từng tính năng, xem tính năng nào đạt độ chính xác cao nhất trên dữ liệu huấn luyện

▪ Gán nhãn dựa trên những nhãn thường xuyên nhất

Kiểm tra độ chính xác cho từng lá trên tập huấn luyện sau khi đã tìm được gốc quyết định tổng thể

***1. Entropy và thu được thông tin***

❖ Dùng mức tăng thông tin để xác định đặc điểm cung cấp nhiều thông tin nhất cho gốc quyết định, đo lường mức độ tổ chức của các giá trị đầu vào, đo mức độ lộn xộn bằng entropy

⬥ Công thức entropy:

⬥ Khi tính toán entropy của tập hợp ban đầu các giá trị đầu vào có thể xác định mức độ tổ chức của các nhãn sau khi áp dụng gốc quyết định

❖ Thuật toán đơn giản để chọn gốc quyết định phải xây dựng gốc quyết định ứng cử viên cho một tính năng có thể và quy trình phải được lặp lại cho mọi nút trong cây quyết định đã xây dựng

❖ Cây quyết định đặc biệt phù hợp với các trường hợp có thể tạo nhiều sự phân biệt phân loại theo thứ bậc

❖ Để tránh trường hợp các nút quyết định thấp hơn overfit tập huấn luyện ta có hai giải pháp:

⬥ Ngừng phân chia các nút khi dữ liệu huấn luyện quá nhỏ

⬥ Phát triển cây quyết định đầy đủ, sau đó loại bỏ các nút quyết định không cải thiện hiệu suất

❖ Với cây quyết định chúng buộc các tính năng phải được kiểm tra theo thứ tự cụ thể. Do không gian gần đỉnh của cây bị hạn chế nên hầu hết các tính năng này bị lặp lại trên nhiều nhánh khác nhau của cây và số lần lặp lại có thể là rất lớn vì số nhánh cây tăng theo cấp số nhân

❖ Cây quyết định không sử dụng tốt các tính năng là yếu tố dự báo yếu của các nhãn chính xác, thay vào đó có thể xem xét tác động của các tính năng này trên toàn bộ tập huấn luyện để đưa ra một số kết luận và cách các tính năng ảnh hưởng đến việc lựa chọn nhãn

1. **Bộ phân loại Naive Bayes**

Trong bộ phân loại Naïve Bayes, mọi tính năng đều có tiếng nói trong việc xác định nhãn nào sẽ được gán cho một giá trị đầu vào nhất định

Cách Naïve Bayes chọn nhãn:

▪ Tính xác suất của từng nhãn: kiểm tra tần suất của nhãn trong tập huấn luyện

* Chọn nhãn có khả năng ước tính cao nhất gán cho giá trị đầu vào

Diagram

Description automatically generated

Hình 6. 5 Minh họa quy trình sử dụng bộ phân loại Naïve Bayes

▪ Các tính năng riêng lẻ đóng góp vào quyết định chung bằng cách "bỏ phiếu chống" các nhãn không thường xuyên xảy ra với tính năng đó

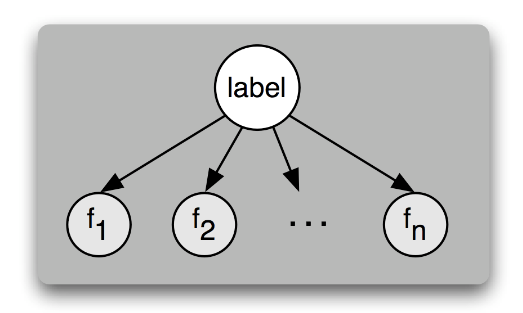
Diagram, schematic

Description automatically generated

Hình 6. 6 Tính khả năng nhãn với Naïve Bayes

1. ***Mô hình xác suất cơ bản***

❖ Naive Bayes giúp việc kết hợp đóng góp của các tính năng khác nhau trở nên dễ dàng hơn vì cách chúng tương tác với nhau



Hình 6. 7 Biểu đồ mạng Bayes minh họa quá trình tổng quát được giả định bởi Naive Bayes

❖ Từ giá định ta có: P(nhãn|tính năng) xác suất đầu vào với tính năng cụ thể với điều kiện có tập hợp các tính năng cụ thể:

⬥ Chọn nhãn cho đầu vào mới: chọn nhãn l tối đa hóa P(l|features)

❖ Ta có:

*P(nhãn|tính năng) = P(tính năng, nhãn)/P(tính năng)*

❖ Tính khả năng xảy ra: *P(features, label)*

⬥ Mở rộng khả năng của nhãn với xác suất của tính năng được cung cấp nhãn

*P(tính năng, nhãn) = P(nhãn) × P(tính năng|nhãn)*

⬥ Hoặc tách xác suất của từng tính năng riêng lẻ:

*P(tính năng, nhãn) = P(nhãn) × Sản phẩm f in| tính năng P(f|nhãn)*

1. ***Đếm zero và làm mịn***

❖ Cách đơn giản nhất để tính P(f|label), đóng góp của một tính năng f đối với khả năng có nhãn của một nhãn label, là lấy tỷ lệ phần trăm các trường hợp đào tạo có nhãn đã cho cũng có tính năng đã cho:

*P(f|nhãn) = đếm(f, nhãn) / đếm(nhãn)*

❖ Tuy nhiên có thể xảy ra trường hợp giá trị tính toán của chúng tôi cho P(f|nhãn) sẽ bằng không, điều này sẽ khiến khả năng nhãn của nhãn đã cho bằng không.

* Khi xây dựng Naive Bayes, kỹ thuật làm mịn được sử dụng để tính *P(f|label), xác suất của một tính năng được gán nhãn*

1. ***Tính năng phi nhị phân***

❖ Tính năng nhị phân nghĩa là đầu vào có hoặc không có tính năng. Các tính năng có giá trị nhãn có thể được chuyển đổi thành các tính năng nhị phân bằng cách:

⬥ Thay thế bằng cách tính năng nhị phân

⬥ Tạo chuỗi

⬥ Dùng phương pháp hồi quy để lập mô hình xác suất các đặc trưng số

1. ***The Naivete of Independence***

❖ Naive Bayes được gọi là “ Naive – ngây thơ” vì thật vô lý khi cho rằng tất cả các tính năng đều độc lập với nhau, bởi chúng đều chứa các đặc trưng với mức độ phụ thuộc lẫn nhau khác nhau

❖ Một vấn đề phát sinh khi bỏ qua giả định độc lập là bộ phân loại có thể kết thúc "tính hai lần" hiệu ứng của các tính năng có mối tương quan cao, đẩy bộ phân loại đến gần một nhãn nhất định hơn là hợp lý.

❖ Để xem điều này có thể xảy ra như thế nào, hãy xem xét một bộ phân loại giới tính tên có chứa hai tính năng giống hệt nhau, f 1 và f 2 . Nói cách khác, f 2 là bản sao chính xác của f 1 và không chứa thông tin mới. Khi bộ phân loại đang xem xét một đầu vào, nó sẽ bao gồm sự đóng góp của cả f 1 và f 2 khi quyết định chọn nhãn nào. Như vậy, nội dung thông tin của hai tính năng này sẽ được coi trọng hơn mức xứng đáng.

1. ***Nguyên nhân của việc đếm hai lần***

❖ Trong quá trình đào tạo, các đóng góp của tính năng được tính riêng; nhưng khi sử dụng bộ phân loại để chọn nhãn cho đầu vào mới, những đóng góp tính năng đó được kết hợp.

❖ Giải pháp:

⬥ Xem xét các tương tác có thể có giữa các đóng góp tính năng trong quá trình đào tạo

⬥ Sử dụng những tương tác đó để điều chỉnh những đóng góp mà các tính năng riêng lẻ tạo ra

⬥ Để chính xác hơn, tách riêng phần đóng góp của từng tính năng (hoặc nhãn)

P(tính năng, nhãn) = w[nhãn] × Sản phẩm f |in| tính năng w[f, nhãn]

<*w[nhãn]* và *w[f, nhãn]* là các tham số hoặc trọng số cho mô hình>

1. **Maximum entropy classifiers**

Bộ phân loại Entropy cực đại sử dụng một mô hình rất giống với mô hình được sử dụng bởi bộ phân loại naive Bayes. Nhưng thay vì sử dụng xác suất để thiết lập các tham số của mô hình, nó sử dụng các kỹ thuật tìm kiếm để tìm một tập hợp các tham số sẽ tối đa hóa hiệu suất của bộ phân loại. Cụ thể, nó tìm kiếm tập hợp các tham số tối đa hóa tổng khả năng xảy ra của tập dữ liệu đào tạo, được định nghĩa là:

Trong đó:

1. **The maximum entropy model**

❖ Giống như mô hình naive Bayes, bộ phân loại Entropy cực đại tính toán khả năng xảy ra của từng nhãn đối với một giá trị đầu vào nhất định bằng cách nhân các tham số có thể áp dụng cho giá trị đầu vào và nhãn với nhau.

❖ Mô hình phân loại Entropy cực đại để người dùng quyết định tổ hợp nhãn và tính năng nào sẽ nhận được tham số riêng của chúng. Mỗi sự kết hợp của nhãn và tính năng nhận tham số riêng của nó được gọi là tính năng chung.

❖ Với các tính năng chung cho mô hình Entropy cực đại, điểm được gán cho nhãn cho một đầu vào nhất định chỉ đơn giản là sản phẩm của các tham số được liên kết với các tính năng chung áp dụng cho đầu vào và nhãn đó:

1. **Tối đa hóa Entropy**

❖ Chúng ta nên xây dựng một mô hình nắm bắt tần số của các đặc điểm chung riêng lẻ mà không đưa ra bất kì giả định không chính đáng nào

❖ Giả sử chúng ta được giao nhiệm vụ chọn nghĩa chính xác cho một từ nhất định, từ danh sách 10 nghĩa có thể có ( được gán nhãn A – J)

|  | A | B | C | D | E | F | J | H | I | J |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| (i) | 10% | 10% | 10% | 10% | 10% | 10% | 10% | 10% | 10% | 10% |
| (ii) | 5% | 15% | 0% | 30% | 0% | 8% | 12% | 0% | 6% | 24% |
| (iii) | 0% | 100% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |

Bảng 6. 1 Tỉ lệ chọn nghĩa chính xác của một từ nhất định

❖ Tiếp theo, chúng ta được thông báo rằng A xuất hiện 55%

|  | A | B | C | D | E | F | J | H | I | J |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| (iv) | 55% | 45% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| (v) | 55% | 5% | 5% | 5% | 5% | 5% | 5% | 5% | 5% | 5% |
| (vi) | 55% | 3% | 1% | 2% | 9% | 5% | 0% | 25% | 0% | 0% |

Bảng 6. 2 Tỉ lệ A xuất hiện

❖ Cuối cùng, giả sử rằng chúng ta được thông báo rằng từ up xuất hiện trong ngữ cảnh gần đó 10% thời gian và khi nó xuất hiện trong ngữ cảnh đó thì có 80% khả năng nghĩa A hoặc C sẽ được sử dụng

|  | A | B | C | D | E | F | J | H | I | J |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| (vii) | 5.1% | 0.25% | 2.9% | 0.25% | 0.25% | 0.25% | 0.25% | 0.25% | 0.25% | 0.25% |
|  | 49.9% | 4.46% | 4.46% | 4.46% | 4.46% | 4.46% | 4.46% | 4.46% | 4.46% | 4.46% |

Bảng 6. 3 kết quả

1. **Mô hình hóa các mẫu ngôn ngữ**

❖ Bộ phân loại có thể giúp chúng ta hiểu các mẫu ngôn ngữ xuất hiện trong ngôn ngữ tự nhiên, bằng cách cho phép chúng ta tạo các mô hình rõ ràng nắm bắt các mẫu đó

❖ Một số mô hình, chẳng hạn như cây quyết định, tương đối minh bạch và cung cấp cho chúng tôi thông tin trực tiếp về yếu tố nào là quan trọng trong việc đưa ra quyết định và yếu tố nào có liên quan với nhau. Các mô hình khác, chẳng hạn như mạng lưới thần kinh đa cấp, mặc dù có thể đạt được cái nhìn sâu sắc bằng cách nghiên cứu chúng, nhưng nó thường mất nhiều công sức hơn

❖ Nhưng tất cả các mô hình rõ ràng đều có thể đưa ra dự đoán về dữ liệu ngôn ngữ mới chưa được nhìn thấy không có trong kho ngữ liệu được sử dụng để xây dựng mô hình. Những dự đoán này có thể được đánh giá để đánh giá độ chính xác của mô hình. Khi một mô hình được coi là đủ chính xác, nó có thể được sử dụng để tự động dự đoán thông tin về dữ liệu ngôn ngữ mới. Các mô hình dự đoán này có thể được kết hợp thành các hệ thống thực hiện nhiều tác vụ xử lý ngôn ngữ hữu ích, chẳng hạn như phân loại tài liệu, dịch tự động và trả lời câu hỏi.

1. **Bài tập**

Link bài tập: <https://colab.research.google.com/drive/1Ktt3qYADvUkM1PDdR524ZoRgy-UeDadJ?usp=sharing>