TRƯỜNG ĐẠI HỌC ĐIỆN LỰC

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO MÔN HỌC
KHAI PHÁ DỮ LIỆU

ĐỀ TÀI: SỬ DỤNG MÔ HÌNH NAIVE BAYES ĐỂ ĐÁNH GIÁ SẢN PHẨM TRÊN SHOPPEE DỰA TRÊN BÌNH LUẬN CỦA SẢN PHẨM

Sinh viên thực hiện	: TRẦN ĐÌNH MINH VƯƠNG
	NGUYỄN THỊ THÙY TRANG
	LÊ THỊ THÙY DƯƠNG
Giảng viên hướng dẫn	: NGUYỄN THỊ THANH TÂN
Ngành	: CÔNG NGHỆ THÔNG TIN
Chuyên ngành	: CÔNG NGHỆ PHẦN MỀM
Lớp	: D14CNPM6

Hà Nội, tháng 03 năm 2022

PHIẾU CHẤM ĐIỂM

Sinh viên thực hiện:

Họ và tên	Chữ ký	Ghi chú
Trần Đình Minh Vương		
Lê Thị Thùy Dương		
Nguyễn Thị Thùy Trang		

Giảng viên chấm

Họ và tên	Chữ ký	Ghi chú
Giảng viên chấm 1		
Giảng viên chấm 2		

MỤC LỤC

DANH MỤC HÌNH ẢNH

LỜI CẢM ƠN

Qua bài tập lớn này, chúng em xin gửi lời cảm ơn tới thầy cô khoa công nghệ thông tin, đặc biệt là thầy Nguyễn Thị Thanh Tân rất cảm ơn cô đã cho chúng em có cơ hội được tìm hiểu một góc kiến thức mới, hay và bổ ích cùng với đó là sự tận tâm dạy dỗ chúng em, giúp chúng em có thể hoàn thiện đề tài này. Trong quá trình tìm hiểu và hoàn thiện, đề tài sẽ không thể tránh khỏi những sai sót, khuyết điểm. Vì vậy, nhóm thực hiện chúng em hy vọng nhận được sự đánh giá và đóng góp nhiệt tình từ phía thầy và các bạn để bài của nhóm chúng em được hoàn thiện hơn.

Qua bài tập lớn này, chúng em xin cảm ơn các bạn bè lớp D14CNPM6 đã giúp đỡ chúng em trong quá trình học tập và làm bài tập lớn, đã chia sẻ kinh nghiệm kiến thức của các bạn đã tạo nên nền tảng kiến thức cho chúng em.

Cuối cùng, chúng em xin gửi lời cảm ơn gia đình đặc biệt là cha mẹ đã tạo điều kiện tốt nhất cho con có đủ khả năng thực hiện bài tập lớn này, trang trải học phí, động viên tinh thần cho em để học tập trong môi trường đại học tuyệt vời này.

Chúng em xin chân thành cảm ơn!

Nhóm sinh viên thực hiện

Trần Đình Minh Vương

Lê Thị Thùy Dương

Nguyễn Thị Thùy Trang

CHƯƠNG 1 TỔNG QUAN ĐỀ TÀI

1.1Đặt vấn đề

Công nghệ thông tin đang dần chứng tỏ tầm ảnh hưởng rất lớn đến mọi mặt của đời sống xã hội. Các phần mềm tin học ngày càng được sử dụng rộng rãi trong đời sống, nhất là khi ngành thương mại điện tử(e-commerce) ra đời. Sự tăng trưởng nhanh chóng về số lượng người sử dụng internet cùng với sự phát triển kỹ thuật công nghệ hiện đại đã góp phần mang lại những tiện ích tuyện vời cho con người. Ở Việt Nam, tuy hình thức này mới hình thành khoảng những năm gần đây nhưng nó đã cho thấy những lợi ích không tưởng và trở thành một trong những lựa chọn hàng đầu cho các cá nhân, kinh doanh nhỏ lẻ và thậm chí là các doanh nghiệp lớn.

Công nghệ thông tin và thương mại điện tử có mối quan hệ mật thiết, hay nói cách khác công nghệ thông tin là nền tảng phát triển của e-commerce. Các sàn thương mại điện tử lớn ở Việt Nam hiện nay có thể kể đến như: Lazada, Shoppee, Tiki,… Dựa trên những ứng dụng, phương pháp và công cụ kỹ thuật của công nghệ thông tin , những nền tảng trang web, hình thức thanh toán, quy trình mua hàng,… sẽ được thiết lập nhằm mang lại cho khách hàng những trải nghiệm mua sắm trực tuyến tối ưu nhất. Con người mua bán trao đổi hàng qua mạng điện tử, thực hiện các giao dịch thanh toán trực tuyến. Sử dụng nhanh gọn, không tốn nhiều công sức. Bên cạnh đó, khách hàng có thể đánh giá sản phẩm cần mua bằng những bình luận trước đó của những khách hàng trước.

Ứng dụng kỹ thuật phân lớp dữ liệu trong khai phá dữ liệu nhằm xây dựng hệ thống đánh giá sản phẩm là một trong những hướng nghiên cứu chính của đề tài. Sau khi phân tích một số thuật toán cũng như đặc điểm của dữ liệu thu nhập được về các bình luận sản phẩm của Shoppee, đề tài đề xuất ứng dụng mô hình phân lớp bằng cây quyết định với thuật toán Naive bayes để tìm ra qui luật tìm ẩn trong dữ liệu.

1.2 Cơ sở hình thành đề tài

Hiện nay ngành Thương mại điện tử phát triển mạnh mẽ, việc mua bán trên các sàn lớn như Lazada, Shopee rất dễ dàng. Nhưng khi khách hàng mua hàng có thể an tâm về sản phẩm đó được hay không? Sản phẩm có đúng như người bán quảng cáo hay không? Thì khách hàng thường sẽ đọc những bình luận của các khách hàng trước đánh giá. Nhưng khách hàng không thể đọc hết các đánh giá sản phẩm được. Từ ý tưởng đó, nhóm chúng em tạo ra 1 hệ thống sử dụng Al để đánh giá sản phẩm trên Lazada, Tiki dựa trên bình luận. Và từ đấy đưa ra gợi ý cho khách hàng có nên mua sản phẩm không?

1.3 Mục tiêu đề tài

Đề tài tập trung vào nghiên cứu kỹ thuật phân lớp trong khai phá dữ liệu, từ đó nắm bắt được những giải thuật làm tiền đề cho nghiên cứu và xây dựng ứng dụng cụ thể.

1.4 Bố cục đề tài

Đề tài được chia thành các phần:

Chương 1: Tổng quan đề tài

Chương 2: Khai phá dữ liệu

Chương 3: Xây dựng mô hình dữ liệu sử dụng Naïve Bayes

Chương 4: Thực nghiệm và đánh giá

CHƯƠNG 2 KHAI PHÁ DỮ LIỆU

2.1 Tổng quan về kĩ thuật khai phá dữ liệu

2.1.1 Khái niệm

Khai phá dữ liệu (*data mining*) Là quá trình tính toán để tìm ra các mẫu trong các bộ dữ liệu lớn liên quan đến các phương pháp tại giao điểm của máy học, thống kê và các hệ thống cơ sở dữ liệu. Đây là một lĩnh vực liên ngành của khoa học máy tính. Mục tiêu tổng thể của quá trình khai thác dữ liệu là trích xuất thông tin từ một bộ dữ liệu và chuyển nó thành một cấu trúc dễ hiểu để sử dụng tiếp.Ngoài bước phân tích thô, nó còn liên quan tới cơ sở dữ liệu và các khía cạnh quản lý dữ liệu, xử lý dữ liệu trước, suy xét mô hình và suy luận thống kê, các thước đo thú vị, các cân nhắc phức tạp, xuất kết quả về các cấu trúc được phát hiện, hiện hình hóa và cập nhật trực tuyến. Khai thác dữ liệu là bước phân tích của quá trình "khám phá kiến thức trong cơ sở dữ liệu" hoặc KDD.

Khai phá dữ liệu là một bước của quá trình khai thác tri thức (*Knowledge Discovery Proces*s), bao gồm:

- Xác định vấn đề và không gian dữ liệu để giải quyết vấn đề (*Problem understanding and data understanding*).
- Chuẩn bị dữ liệu (*Data preparation*), bao gồm các quá trình làm sạch dữ liệu (*data cleaning*), tích hợp dữ liệu (*data integration*), chọn dữ liệu (*data selection*), biến đổi dữ liệu (*data transformation*).
- Khai thác dữ liệu (*Data mining*): xác định *nhiệm vụ khai thác dữ liệu* và lựa chọn *kỹ thuật khai thác dữ liệu*. Kết quả cho ta một *nguồn tri thức thô*.
- Đánh giá (*Evaluation*): dựa trên một số tiêu chí tiến hành *kiểm tra* và *lọc* nguồn tri thức thu được.
- Triển khai (*Deployment*).

Quá trình khai thác tri thức không chỉ là một quá trình tuần tự từ bước đầu tiên đến bước cuối cùng mà là một quá trình lặp và có quay trở lại các bước đã qua.

2.1.2 Quy trình khai phá dữ liệu

a) Nghiên cứu lĩnh vực

Ta cần nghiên cứu lĩnh vực cần sử dụng Data mining để xác định được những tri thức ta cần chất lọc, từ đó định hướng để tránh tốn thời gian cho những tri thức không cần thiết .

b) Tạo tập tin dữ liễu đầu vào

Ta xây dựng tập tin để lưu trữ các dữ liệu đầu vào để máy tính có thể lưu trữ và xử lý.

c) Tiền xử lý, làm sạch, mã hóa

Ở bước này ta tiến hành bỏ bớt những dữ liệu rườm rà, không cần thiết, tinh chỉnh lại cấu trúc của dữ liệu và mã hóa chúng để tiện cho quá trình xử lý .

d) Rút gọn chiều

Thông thường một tập dữ liệu có chiều khá lớn sẽ sinh ra một lượng dự liệu khổng lồ,ví dụ với n chiều ta sẽ có 2^n nguyên tổ hợp .Do đó , đây là một bước quan trọng giúp giảm đáng kể hao tổn hề tài nguyên trong quá trình xử lý tri thức.Thông thường ta sẽ dung Rough set (http://en.wikipedia.org/wiki/Rough_set) để giảm số chiều.

e) Chọn tác vụ khai thác dữ liệu

Để đạt được mục đích ta cần, ta chọn được tác vụ khai thác dữ liệu sao cho phù hợp.Thông thường có các tác vụ sau:

- Đặc trưng(feature)
- Phân biệt(discrimination)
- Kết hợp(association)
- Phân lớp(classification)
- Gom cụm(clusterity)

- Xu thế(trend analysis)
- Phân tích độ lệch
- Phân tích độ hiếm

f) Chọn các thuật giải khai thác dữ liệu

g) Khai thác dữ liệu: Tìm kiếm tri thức

Sau khi tiến hành các bước trên thì đây là bước chính của cả quá trình , ta sẽ tiến hành khai thác và tìm kiếm tri thức.

h) Đánh giá mẫu tìm được

Ta cần đánh giá lại trong các tri thức tìm được , ta sẽ sử dụng được những tri thức nào , những tri thức nào dư thừa,không cần biết.

j) Biểu diễn tri thức

Ta biểu diễn tri thức vừa thu nhập được dưới dạng ngôn ngữ tự nhiên và hình thức sao cho người dùng có thể hiểu được những tri thức đó.

j) Sử dụng các tri thức vừa khám phá

Các bước quan trọng khi Data Mining bao gồm:

Bước 1: Làm sạch dữ liệu – Trong bước này, dữ liệu được làm sạch sao cho không có tạp âm hay bất thường trong dữ liệu.

Bước 2: Tích hợp dữ liệu – Trong quá trình tích hợp dữ liệu, nhiều nguồn dữ liệu sẽ kết hợp lại thành một.

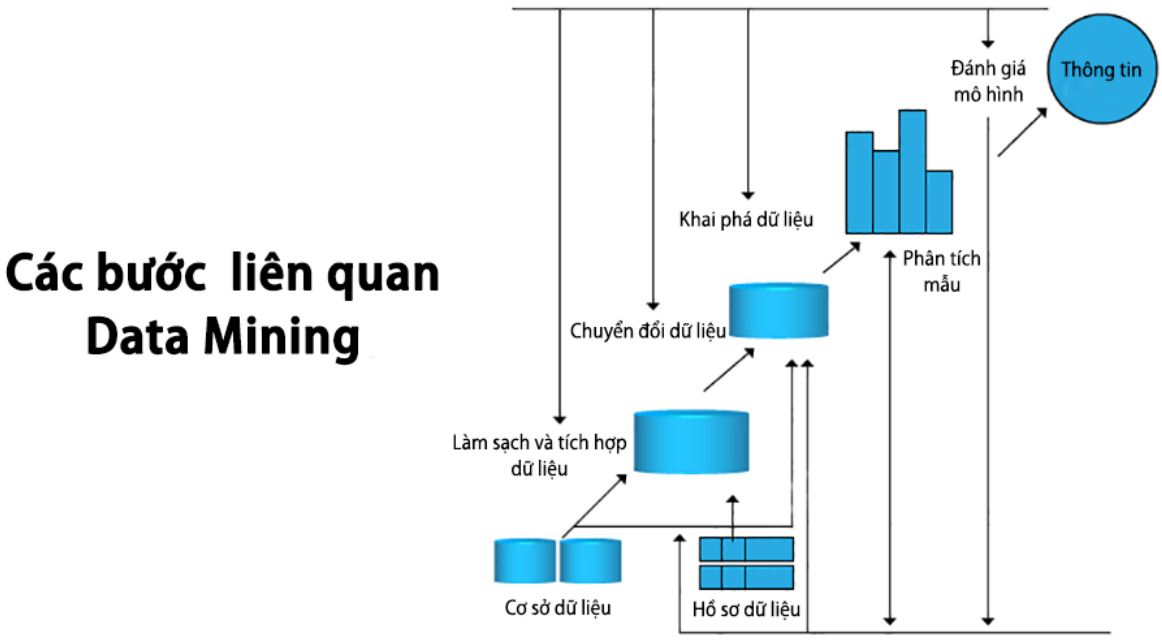
Bước 3: Lựa chọn dữ liệu – Trong bước này, dữ liệu được trích xuất từ cơ sở dữ liệu.

Bước 4: Chuyển đổi dữ liệu – Trong bước này, dữ liệu sẽ được chuyển đổi để thực hiện phân tích tóm tắt cũng như các hoạt động tổng hợp.

Bước 5: Khai phá dữ liệu – Trong bước này, chúng tôi trích xuất dữ liệu hữu ích từ nhóm dữ liệu hiện có.

Bước 6: Đánh giá mẫu – Chúng tôi phân tích một số mẫu có trong dữ liệu.

Bước 7: Trình bày thông tin – Trong bước cuối cùng, thông tin sẽ được thể hiện dưới dạng cây, bảng, biểu đồ và ma trận.



Hình 2. 1 Các bước liên quan Data Mining

2.1.3 Ứng dụng của Data Mining

Có nhiều ứng dụng của Data Mining thường thấy như:

- Phân tích thị trường và chứng khoán
- Phát hiện gian lận
- Quản lý rủi ro và phân tích doanh nghiệp
- Phân tích giá trị trọn đời của khách hàng
- Khám phá thêm 10 ứng dụng khai phá dữ liệu

2.2 Tổng quan về hệ hỗ trợ ra quyết định

Hệ hỗ trợ ra quyết định là một hệ thống thuộc hệ thống thông tin, có nhiệm vụ cung cấp các thông tin hỗ trợ cho việc ra quyết định để tham khảo và giải quyết vấn đề. Hệ hỗ trợ ra quyết định có thể dùng cho cá nhân hay tổ chức và có thể hỗ trợ gián tiếp hoặc trực tiếp.

2.3 Bài toán phân lớp trong khai phá dữ liệu

2.3.1 Khái niệm

Phân lớp là một hình thức phân tích dữ liệu nhằm rút ra những mô hình mô tả những lớp trong dữ liệu. Những mô hình này gọi là mô hình phân lớp (classifier hoặc

classsification) được dùng để dự đoán những nhãn lớp có tính phân loại (categorical), rời rạc và không có thứ tự cho những đối tượng dữ liệu mới.

2.3.2 Quá trình phân lớp dữ liệu

a) Chuẩn bị tập dữ liệu huấn luyện (dataset) và rút trích đặc trưng (feature extraction)

Công đoạn này được xem là công đoạn quan trọng trong các bài toán về Machine Learning. Vì đây là input cho việc học để tìm ra mô hình của bài toán. Chúng ta phải biết cần chọn ra những đặt trưng tốt (good feature) của dữ liệu, lược bỏ những đặc trưng không tốt của dữ liệu, gây nhiễu (noise). Ước lượng số chiều của dữ liệu bao nhiêu là tốt hay nói cách khác là chọn bao nhiêu feature. Nếu số chiều quá lớn gây khó khăn cho việc tính toán thì phải giảm số chiều của dữ liệu nhưng vẫn giữ được độ chính xác của dữ liệu (reduce demension).

Ở bước này chúng ta cũng chuẩn bị bộ dữ liệu để test trên mô hình. Thông thường sẽ sử dụng cross-validation (kiểm tra chéo) để chia tập datasets thành 2 phần, 1 phần phục vụ cho training (training datasets) và phần còn lại phục vụ cho mục đích testing trên mô hình (testing dataset). Có 2 cách thường sử dụng trong cross-validation là splitting và k-fold.

b) Xây dựng mô hình phân lớp (classifier model)

Mục đích của mô hình huấn luyện là tìm ra hàm f(x) và thông qua hàm f tìm được để gán nhãn cho dữ liệu, bước này thường được gọi là learning hay training.

F(x) = y

x: các feature hay input đầu vào ủa dữ liệu

y: nhãn lớp hay output đầu ra

Thông thường để xây dựng mô hình phân lớp cho bài toán này cần sử dụng các thuật toán học giám sát (supervised learning) như k-nearest neighbors, Neural Network, SVM, Decision tree, Naïve Bayes.

c) Kiểm tra dữ liệu với mô hình (making predictions)

Sau khi đã tìm được mô hình phân lớp ở bước 2, thì ở bước này sẽ đưa vào các dữ liệu mới để kiểm tra trên mô hình phân lớp.

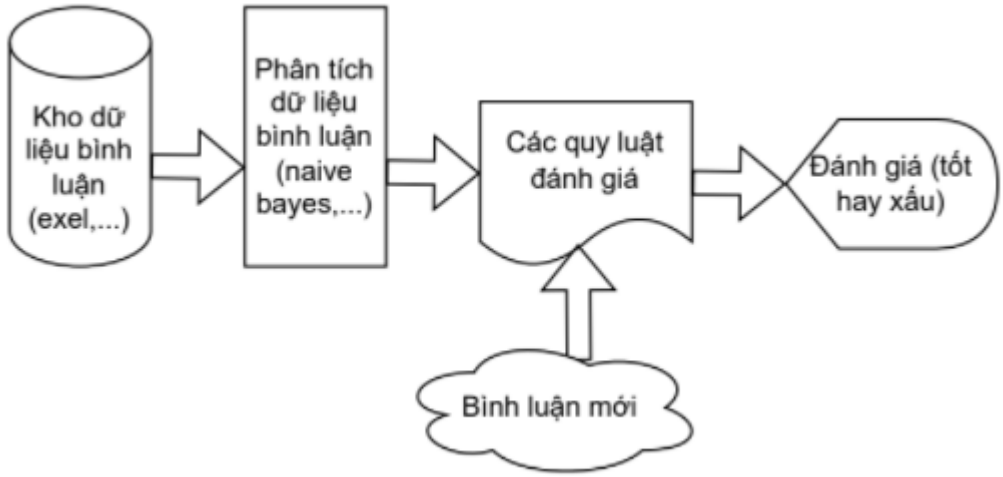
d) Đánh giá mô hình phân lớp và chọn ra mô hình tốt nhất

Bước cuối cùng sẽ đánh giá mô hình bằng cách đánh giá mức độ lỗi của dữ liệu testing và dữ liệu training thông qua mô hình tìm được. Nếu không đạt được kết quả mong muốn thì phải thay đổi các tham số (turning parameters) của các thuật toán học để tìm ra các mô hình tốt hơn và kiểm tra, đánh giá lại mô hình phân lớp, và cuối cùng chọn ra mô hình phân lớp tốt nhất cho bài toán.

CHƯƠNG 3 XÂY DỰNG MÔ HÌNH DỮ LIỆU SỬ DỤNG NAIVE BAYES

3.1 Cơ sở dữ liệu xây dựng mô hình

Sau khi thu thập dữ liệu ta cần xây dựng cơ sở dữ liệu, lưu trữ các thông tin cần thiết cho bộ điều khiển theo mô hình sau:



Hình 3. 1 Cơ sở dữ liệu xây dựng mô hình

3.2 Phương pháp Naive Bayes sử dụng trong khai phá dữ liệu

3.2.1 Định nghĩa thuật toán Naive Bayes

Naïve 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. Naïve Bayes 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.

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 Naïve 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:

$$P(A|B) = \frac{P(B|A)XP(A)}{P(B)}$$

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:

$$P(A|B) = \frac{(P(B_1|A)xP(B_2|A)xP(B_3|A) \dots xP(B_n|A)xP(A))}{P(B_1)xP(B_2)xP(B_3) \dots xP(B_n)}$$

Ví dụ: Ta có một thống kê như sau:

+ 120 thượng nghị sĩ thuộc đảng Dân Chủ.

+ 80 thượng nghĩ sĩ thuộc đảng Cộng Hòa.

+ Số lượng Nữ giới trong đám thượng nghị sĩ là 60 người

+ Còn lại 140 người còn lại là Nam giới (giả dụ chả có ông thượng nghị sĩ nào mới đi Thái về cả).

+ Và số lượng Nữ giới trong đám Dân Dủ là 30 người.

Vậy nếu tôi chọn ngẫu nhiên một người trong đám thượng nghị sĩ thì tỷ lệ thượng nghị sĩ là Nữ giới và thuộc đảng Dân Chủ thì tỷ lệ là bao nhiêu?

Áp dụng công thức Bayes ta có thể tính toán được bằng công thức sau:

$$P(Female|Democrat) = \frac{P(Democrat|Female)X\ P(Female)}{P(Democrat)}$$

- P(Female|Democrat): Chính là tỷ lệ nữ giới thuộc đảng dân chủ trong cả đám thượng nghị sĩ cần tính toán
- P(Demorate|Female): Chính là tỷ lệ nữ giới trong đảng dân chủ
- P(Female): Chính là tỷ lệ nữ giới trong cả đám thượng nghị sĩ
- P(Democrat): Chính là tổng cả đám thượng nghị sĩ.

Ở đây với dữ liệu cho bên trên ta có thể tính toán được

- P(Democrat|Female) = Số nữ giới trong đám dân chủ / Tổng đám thượng nghị đảng dân chủ
- P(Democrat|Female) = 30/ 120 = 0.25
- P(Female) = Số nữ giới trong cả đám thượng nghị sĩ / Tổng đám thượng nghị sĩ
- P(Female) = 60/200 = 0.3
- P(Democrat) = Tổng đám thượng nghĩ sĩ
- P(Democrat) = 1

Vậy ta có thể tín ra P(Female|Democrat) theo công thức Bayes như sau:

- P(Female|Democrat) = (0.25 * 0.3) / 1 = 0.075

Có nghĩa là nếu tôi chọn chọn ngẫu nhiên một người trong đám thượng nghị sĩ thì tỷ lệ thượng nghị sĩ là Nữ giới và thuộc đảng Dân Chủ thì tỷ lệ sẽ là "7,5%".

Trên đây là một ví dụ rất đơn giản được tính toán bằng định lý Bayes mà thật ra nếu bạn nào giỏi có thể tự tính nhẩm ra mà ko cần sử dụng định lý trên.

Nguyên tắc hoạt động của bộ phân lớp Naive Bayes

1. Cho D là tập dữ liệu huấn luyện cùng với các nhãn lớp tương ứng. Như thường lệ, mỗi bộ dữ liệu được mô tả bởi n thuộc tính và được diễn đạt dưới dạng vector n chiều X = (x1, x2,x3,…,xn).
2. Giả sử rằng có m nhãn lớp khác nhau gồm C1, C2,…, Cm. Cho một bộ dữ liệu X, bộ phân lớp sẽ dự đoán X thuộc về phân lớp có xác suất hậu nghiệm cao nhất.

$$P(C_i|X) > P(C_j|X) với\ 1 \le j \le m, j \ne i$$
$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

3. Do P(X) không đổi, nên ta chỉ cần cực đại hóa giá trị P(X|Ci)P(Ci)

Ví dụ:

Dữ liệu được minh họa như hình:

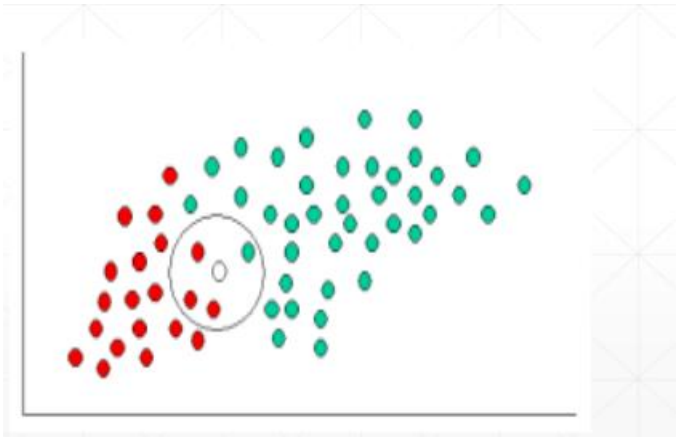
Hình 3. 2 Dữ liệu minh họa

Có 2 lớp: xanh và đỏ; N: tổng số đối tượng

P(xanh) = |xanh|/N = 40/60

P(đỏ) = |đỏ|/N = 20/60

Với các xác suất tiên nghiệm đã xác định ở trên: P(xanh) và P(đỏ) hãy xác định nhãn lớp cho các đối tượng x mới trên hình.



Hình 3. 3 Dữ liệu minh họa

Lấy x làm tâm, vẽ vòng tròn giới hạn các đối tượng lân cận với x, tính:

P(x|xanh) = |xanh lân cận|/|xanh| = 1/40

P(x|đỏ) = |đỏ lân cận|/|đỏ| = 3/20

P(xanh|x) = P(x|xanh).P(xanh) = (1/40 *40/40) = 1/60

P(đỏ|x) = P(x|đỏ).P(đỏ) = (3/20 * 20/60) = 1/20

x được gán nhãn đỏ.

Ví dụ:

Cơ sở dữ liệu khách hàng:

ID	Tuổi	Thu nhập	Sinh viên	Đánh giá tín dụng	Mua máy tính
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle	low	yes	excellent	yes
8	youth	medum	no	fair	yes
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle	medium	no	excellent	yes
13	middle	high	yes	fair	yes
14	senior	medium	no	excellent	no

Giả sử ta có một khách hàng mới X có các thuộc tính

X = (age = youth, income = medium, student = yes, credit_rating = fair)

Bây giờ cần xác định xem khách hàng X có thuộc lớp C_{yes} (mua máy tính) hay

không, ta tính toán như sau:

$P(C_{yes}) = 10/14 = 0.714$

$P(C_{no}) = 4/14 = 0.286$

Các xác suất thành phần:

$P(age = youth|C_{yes}) = 3/10 = 0.3$

$P(age = youth|C_{no}) = 2/4 = 0.5$

$P(income = medium|C_{yes}) = 4/10 = 0.4$

$P(income = medium|C_{no}) = 1/4 = 0.25$

$P(student = yes|C_{yes}) = 6/10 = 0.6$

$P(student = yes|C_{no}) = 1/4 = 0.25$

$P(credit\_rating = fair|C_{yes}) = 7/10 = 0.7$

$P(credit\_rating = fair|C_{no}) = 1/4 = 0.25$

Cuối cùng:

$P(X|C_{yes}) = 0.3 * 0.4 * 0.6* 0.7 = 0.05$

$P(X|C_{no}) = 0.5 * 0.25 * 0.25*0.25 = 0.007$

$P(X|C_{yes})*P(C_{yes}) = 0.05 * 0.714 = 0,0357$

$P(X|C_{no})*P(C_{no}) =0.007 * 0.286 = 0,002$

Từ kết quả này ta thấy $P(X|C_{yes})P(C_{yes})$ có giá trị lớn nhất, do đó thuật toán Bayes sẽ kết luận là khách hàng X sẽ mua máy tính.

3.2.2 Đặc điểm của thuật toán Navie Bayes

- Ưu điểm của thuật toán Naive Bayes

+ Cho phép kết hợp tri thức tiền nghiệm (prior knowledge) và dữ liệu quan sát được (obserwed data).

+ Giả định độc lập: hoạt động tốt cho nhiều bài toán/miền sữ liệu và ứng dụng.

+ Đơn giản nhưng đủ tốt để giải quyết nhiều bài toán như phân lớp văn bản, lọc spam,..

+ Tốt khi có sự chệnh lệch số lượng giữa các lớp phân loại.

+ Huấn luyện mô hình (ước lượng tham số) dễ và nhanh

- Nhược điểm của thuật toán Naive Bayes

+ Giả định độc lập (ưu điểm cũng chính là nhược điểm) hầu hết các trường hợp thực tế trong đó có các thuộc tính trong các đối tượng thường phụ thuộc lẫn nhau.

+ Vấn đề zero (đã nêu cách giải quyết ở phía trên)

+ Mô hình không được huấn luyện bằng phượng pháp tối ưu mạnh và chặt chẽ. Tham số mủa mô hình là các ước lượng xác suất điều kiện đơn lẻ. Không tính đến sự tương tác giữa các ước lượng này.

3.2.3 Ứng dụng của thuật toán Navie Bayes

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

- **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:** Naive Bayes 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 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 …

3.3 Thuật toán Naive Bayes trong giải quyết bài toán đánh giá sản phẩm trên Shoppee dựa trên bình luận của sản phẩm

3.3.1 Phương pháp Naive Bayes

Lý thuyết Bayes thì có lẽ không còn quá xa lạ với chúng ta nữa rồi. Nó chính là sự liên hệ giữa các xác suất có điều kiện. Điều đó gợi ý cho chúng ta rằng chúng ta có thể tính toán một xác suất chưa biết dựa vào các xác suất có điều kiện khác. Thuật toán ***Naive Bayes*** cũng dựa trên việc tính toán các xác suất có điều kiện đó. Nghe tên thuật toán là đã thấy gì đó ngây ngô rồi. Tại sao lại là **Naive** nhỉ. Không phải ngẫu nhiên mà người ta đặt tên thuật toán này như thế. Tên gọi này dựa trên một giả thuyết rằng các chiều của dữ liệu X=(x_1, x_2, ..., x_n)X=(x1,x2,....,xn) là độc lập về mặt xác suất với nhau.

$$p(\mathbf{x}|c) = p(x_1, x_2, \ldots, x_d|c) = \prod_{i=1}^{d} p(x_i|c)$$

Chúng ta có thể thấy rằng giả thuyết này có vẻ khá ***ngây thơ*** vì trên thực tế điều này có thể nói là không thể xảy ra tức là chúng ta rất ít khi tìm được một tập dữ liệu mà các thành phần của nó không liên quan gì đến nhau. Tuy nhiên, giả thiết ngây ngô này lại mang lại những kết quả tốt bất ngờ. Giả thiết về sự độc lập của các chiều dữ liệu này được gọi là Naive Bayes (xin phép không dịch). Cách xác định class của dữ liệu dựa trên giả thiết này có tên là **Naive Bayes Classifier (NBC)**. Tuy nhiên dựa vào giả thuyết này mà bước training và testing trở nên vô cùng nhanh chóng và đơn giản. Chúng ta có thể sử dụng nó cho các bài toán large-scale. Trên thực tế, **NBC** hoạt động khá hiệu quả trong nhiều bài toán thực tế, đặc biệt là trong các bài toán phân loại văn bản, ví dụ như lọc tin nhắn rác hay lọc email spam. Trong bài tiểu luận sẽ áp dụng lý thuyết để giải quyết bài toán mới đó là bài toán đánh giá sản phẩm trên shopee dựa trên bình luận của sản phẩm.

3.3.2 Tập dữ liệu bình luận

Sau khi thu thập data sẽ tiến hành gán nhãn cho chúng, ví dụ ở đây chỉ có 2 trạng thái comment : tốt-tích cực- trung gian :0, xấu- không tốt- tiêu cực : 1 (nếu có nhiều hơn 2 trạng thái thì gán 0,1,2 ,3 ,..).

Có một điều nhận thấy rằng giá trị của các chỉ số là một biến liên tục chứ không phải một giá trị rời rạc chính vì thế nên khi áp dụng thuật toán Naive Bayes chúng ta cần phải áp dụng một phân phối xác suất cho nó. Một trong những phân phối xác suất phổ biến được sử dụng trong phần này đó chính là phân phối Gaussian.

- **Qúa trình train (huấn luyện):**

- Bước 1: Crawl dữ liệu bình luận trên Shoppee và gán nhán cho các bình luận. Ví dụ để đơn giản nhãn 0: Xấu hoặc bình thường, 1: Tốt
- Bước 2: Tiền xử lý dữ liệu như: bỏ các dấu chấm, dấu phẩy…
- Bước 3: Tokenize văn bản bằng thư viện Underthesea (vì đây là văn bản tiếng Việt)
- Bước 4: Embedding văn bản bằng **TF-IDF** (Term Frequency – Inverse Document Frequency) – một thuật toán embbeding có sử dụng việc đánh giá tầm quan trọng của một từ trong một văn bản trong quá trình xây dựng vocabulary. Những từ có giá trị TF-IDF cao là những từ xuất hiện nhiều trong văn bản ta đang dánh giá, và xuất hiện ít trong các văn bản còn lại. Việc này giúp loại bỏ những từ phổ biến và giữ lại những từ có giá trị cao (từ khoá của văn bản đó).
- Bước 5: Train một model SVM để classify dựa trên input là các vector embedding nói trên và output là các nhãn đã gán. Sau khi train xong lưu model vào file.

- **Quá trình test:**

- Bước 1: Đọc một URL về một sản phẩm trên Shoppee và crawl nội dung comment của sản phẩm đó.
- Bước 2: Chúng ta cũng thực hiện tiền xử lý các comment, tokenize sử dụng UndertheSea
- Bước 3: Sau đó thực hiện Embedding bằng TFIDF.
- Bước 4: Lần lượt đưa các comment vào model để classify 0/1 (Xấu/Tốt)
- Bước 5: Nếu các sản phẩm có số comment Tốt > Xấu thì hiện ra recommend nên mua và ngược lại.

3.3.3 Phân phối Gaussian

Với một một dữ liệu x_i thuộc một class c_i chúng ta thấy x_i tuân theo một phân phối chuẩn với kì vọng δ và độ lệch chuẩn μ. Khi đó hàm xác suất của x_i được xác định như sau:

$$p(x_i|c_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

Đây chính là cách tính của thư viện ***sklearn*** tuy nhiên trong bài tiểu luận sẽ cài đặt thủ công vì giúp chúng ta hiểu hơn về bài toán. Phần cài đặt sẽ được trình bày ở chương 4.

CHƯƠNG 4 THỰC NGHIỆM VÀ ĐÁNH GIÁ

KẾT LUẬN

1- Kết quả đạt được

- Thuật toán Naive Bayes đem lại kết quả bất ngờ, tỉ lệ nhận dạng cao. Kết quả phân loại cao nhưng chỉ nằm trong một điều kiện, bài toán cụ thể chứ không phải đúng trong tất cả mô hình khác.
- Một trong những lý do nghĩ đến có thể là trong bình luận có những từ hay xuất hiện ở đánh giá xấu và những từ hay xuất hiện ở đánh giá tốt có tỷ lệ gần bằng nhau dẫn đến việc tính xác suất bị sai. Khi đó mô hình sẽ bị phân loại sai.

2- Hướng phát triển

- Mô hình phân loại bình luận sản phẩm trên chỉ sử dụng thuật toán Naive Bayes, chúng ta còn có thể sử dụng các thuật toán khác như SVM, mạng Nơron… cũng cho ta kết quả khá tốt. Không có thuật toán nào là tối ưu nên việc chọn thuật toán cho phù hợp với bài toán là rất quan trọng.
- Việc thu thập thêm dữ liệu từ cơ sở dữ liệu trên cũng là vấn đề em quan tâm. Dựa vào cơ sở dữ liệu trên ta có thể tìm thêm các dữ liệu rồi gán nhãn cho chúng là bình luận tốt hay xấu. Việc dữ liệu càng nhiều thì mô hình phân loại sẽ chính xác hơn.
- Xây dựng một mô hình phân loại bình luận, hiệu quả cao, tạo ra một ứng dụng để khách hàng có thể sử dụng nó.

TÀI LIỆU THAM KHẢO