TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO THỰC NGHIỆM

HỌC PHẦN: TRÍ TUỆ NHÂN TẠO

Đề tài: Tìm hiểu thuật toán Naïve Bayes và ứng dụng dự báo thời tiết

Giảng viên hướng dẫn: ThS. Lê Thị Thủy

Lớp học phần: 20224IT6043001

Nhóm 8:

1. **Đinh Hồng Liễu - 2021601817**
2. **Đặng Đức Mạnh - 2020604935**
3. **Lê Thị Ngọc - 2021601311**

Hà Nội, năm 2023

Mục lục

Mục lục **1**
Phụ lục hình ảnh **3**
Danh mục bảng **4**
Lời mở đầu **5**
Chương 1. Tổng quan về trí tuệ nhân tạo **6**
1.1. Khái niệm về trí tuệ nhân tạo 6
1.2. Vai trò của trí tuệ nhân tạo 8
1.3. Kĩ thuật trong TTNT là gì và một số kĩ thuật cơ bản trong TTNT 13
1.4. Lịch sử phát triển của trí tuệ nhân tạo 14
1.5. Các thành phần trong hệ thống của trí tuệ nhân tạo 16
1.6. Các lĩnh vực nghiên cứu và ứng dụng cơ bản 17
Chương 2. Tìm hiểu về thuật toán Naïve Bayes **19**
2.1. Định lý Bayes 19
2.1.1. Xác suất có điều kiện 19
2.1.2. Định lý Bayes 19
2.2. Phân lớp Naïve Bayes 21
2.2.1. Bài toán tổng quát 21
2.2.2. Bài toán với dữ liệu liên tục 23
2.2.3. Bài toán với xác suất điều kiện bằng không 24
2.2.4. Ví dụ minh hoạ 24
2.3. Ưu điểm, hạn chế 27
Chương 3. Ứng dụng Naive Bayes trong bài toán dự báo thời tiết **29**
3.1. Bài toán 29
3.1.1. Phát biểu bài toán 29
3.1.2. Thuật toán áp dụng 29
3.2. Cơ sở dữ liệu 29

3.3. Cài đặt 31
3.3.1. Thuật toán 31
3.3.2. Cài đặt 31
Tổng kết 35
Tài liệu tham khảo 36

Phụ lục hình ảnh

Hinh 1. 1. AI là một bộ phận của khoa học máy tính 6
Hinh 1. 2. Một cảnh trong bộ phim "I, Robot" nói về một AI đã tiến hóa 7
Hinh 1. 3. Áp dụng AI trong y khoa 8
Hinh 1. 4. Áp dụng AI trong lĩnh vực tài chính 9
Hinh 1. 5. Áp dụng AI để sản xuất robot hút bụi 10
Hinh 1. 6. Robot thay thế con người trong một số công việc 12
Hình 3. 1. Đọc dữ liệu từ file CSV 31
Hình 3. 2. Tính P(c1), P(c2) 31
Hình 3. 3. Hàm tính P(x|c1) 32
Hình 3. 4. Hàm tính P(x|c2) 32
Hình 3. 5. Hàm dự đoán phân lớp x 32
Hình 3. 6. Hàm dự đoán thời tiết dựa vào phân lớp cụ thể 33
Hình 3. 7. Hàm dự đoán thời tiết với phân lớp được nhập từ bàn phím 33
Hình 3. 8. Kết quả sau khi chạy 34

Danh mục bảng

Bảng 1. Bảng dữ liệu huấn luyện 24
Bảng 2. Bảng mẫu dữ liệu trong file CSV 30

Lời mở đầu

Ngày nay, Trí tuệ nhân tạo (AI) dần đi vào cuộc sống của chúng ta một cách mạnh mẽ và được xác định là lĩnh vực mũi nhọn trong ngành kinh tế các nước. Cách mạng công nghiệp 4.0 mở ra một thời kỳ mới với việc ứng dụng trí tuệ nhân tạo trong hầu hết các lĩnh vực trong đời sống, mang lại những thay đổi lớn trong xã hội, đặc biệt là trong kinh tế và khoa học ứng dụng.

Nội dung cuốn báo cáo này cung cấp cho người đọc những kiến thức chung nhất về trí tuệ nhân tạo, có một cái nhìn tổng quát về việc áp dụng thuật toán Naive Bayes và Ứng dụng thuật toán Naive Bayes vào dự báo thời tiết. Các vấn đề cụ thể trình bày trong cuốn báo cáo được chia thành 3 chương:

Chương 1: Tổng quan về trí tuệ nhân tạo

Chương 2: Tìm hiểu thuật toán Naive Bayes

Chương 3: Ứng dụng thuật toán Naive Bayes vào dự báo thời tiết

Nhóm em xin gửi lời cảm ơn đến cô Lê Thị Thủy - Giảng viên Trường Đại học Công Nghiệp Hà Nội, cô đã tận tình hướng dẫn và góp ý để hoàn thành bài báo cáo. Mặc dù đã rất cố gắng xong báo cáo này không tránh khỏi những thiếu sót, nhóm em mong nhận những đóng góp ý kiến của cô và bạn đọc để bài báo cáo được hoàn thiện hơn.

Xin chân thành cảm ơn!

Nhóm sinh viên thực hiện

Chương 1. Tổng quan về trí tuệ nhân tạo

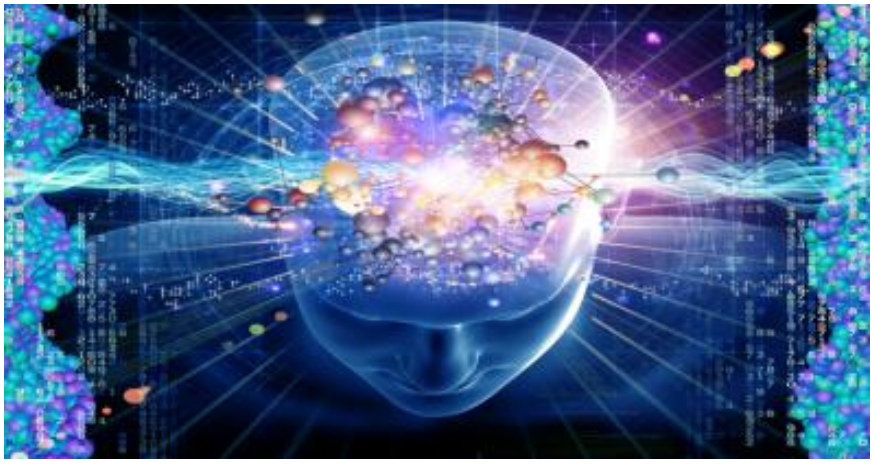
1.1. Khái niệm về trí tuệ nhân tạo

Theo như cha đẻ của trí tuệ nhân tạo, John McCarthy thì nó là "Khoa học và kỹ thuật của việc tạo ra những máy thông minh, đặc biệt là chương trình máy tính thông minh".

Trí tuệ nhân tạo là hướng đi của việc tạo ra máy tính, người máy điều khiển bằng máy tính hay là những phần mềm suy nghĩ thông minh hơn, tương tự như suy nghĩ thông minh của con người.

Trí tuệ nhân tạo được học như bộ não con người, như cách mà con người học, quyết định và làm việc khi giải quyết một vấn đề, và sau đó sử dụng kết quả của quá trình học đó như là nền tảng của việc phát triển phần mềm và hệ thống thông minh.

Ở thời điểm hiện tại, Thuật ngữ này thường dùng để nói đến các MÁY TÍNH có mục đích không nhất định và ngành khoa học nghiên cứu về các lý thuyết và ứng dụng của trí tuệ nhân tạo. Tức là mỗi loại trí tuệ nhân tạo hiện nay đang dừng lại ở mức độ những máy tính hoặc siêu máy tính dùng để xử lý một loại công việc nào đó như điều khiển một ngôi nhà, nghiên cứu nhận diện hình ảnh, xử lý dữ liệu của bệnh nhân để đưa ra phác đồ điều trị, xử lý dữ liệu để tự học hỏi, khả năng trả lời các câu hỏi về chẩn đoán bệnh, trả lời khách hàng về các sản phẩm của một công ty,...



Hinh 1. 1. AI là một bộ phận của khoa học máy tính

Nói nôm na cho dễ hiểu: đó là trí tuệ của máy móc được tạo ra bởi con người. Trí tuệ này có thể tư duy, suy nghĩ, học hỏi,... như trí tuệ con người. Xử lý dữ liệu ở mức rộng lớn hơn, quy mô hơn, hệ thống, khoa học và nhanh hơn so với con người.

Rất nhiều hãng công nghệ nổi tiếng có tham vọng tạo ra được những AI (trí tuệ nhân tạo) vì giá trị của chúng là vô cùng lớn, giải quyết được rất nhiều vấn đề của con người mà loài người đang chưa giải quyết được.

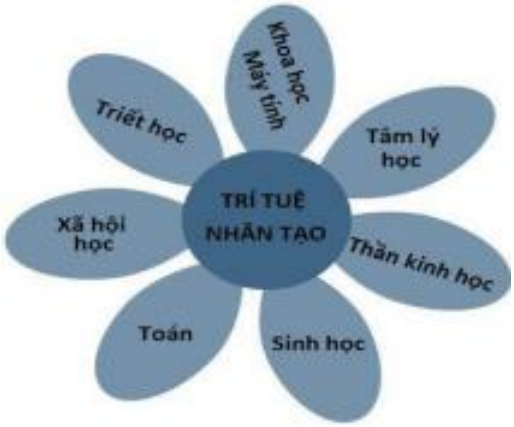
Trí tuệ nhân tạo mang lại rất nhiều giá trị cho cuộc sống loài người, nhưng cũng tiềm ẩn những nguy cơ. Rất nhiều chuyên gia lo lắng rằng khi trí tuệ nhân tạo đạt tới 1 ngưỡng tiến hóa nào đó thì đó cũng là thời điểm loài người bị tận diệt. Rất nhiều các bộ phim đã khai thác đề tài này với nhiều góc nhìn, nhưng qua đó đều muốn cảnh báo loài người về mối nguy đặc biệt này.

Hinh 1. 2. Một cảnh trong bộ phim "I, Robot" nói về một AI đã tiến hóa

Trí tuệ nhân tạo là một ngành khoa học và công nghệ dựa trên nền tảng của Khoa học máy tính, Sinh học, Triết học, Ngôn ngữ học, Toán học và Kỹ thuật. Một chuyên ngành

chính của Trí tuệ nhân tạo là phát triển chức năng của máy tính kết hợp với sự thông minh của con người, chẳng hạn như suy luận, học hỏi và giải quyết vấn đề.

Trong những lĩnh vực dưới đây, một hoặc nhiều lĩnh vực có thể góp thành để xây dựng hệ thống thông minh.



1.2. Vai trò của trí tuệ nhân tạo

Vai trò của AI là vô tận đối với cuộc sống của chúng ta. AI có thể tiếp cận với con người thông qua nhiều lĩnh vực, ngành nghề khác nhau. Ưu điểm của trí tuệ nhân tạo AI là khả năng xử lý dữ liệu khoa học hơn, nhanh hơn, hệ thống hơn so với con người. Việc phát triển và đưa các sản phẩm AI tới tay người dùng đúng cách sẽ thúc đẩy mạnh mẽ sự phát triển của toàn nhân loại. Mở ra một thế giới hoàn toàn mới cùng các giải pháp bù đắp cho những vấn đề mà con người không thể giải quyết.

Vai trò của trí tuệ nhân tạo trong y học

Hinh 1. 3. Áp dụng AI trong y khoa

Công nghệ AI đã mở ra một trang mới cho nền y học thế giới, đặc biệt là nền y học nước nhà. Nó mang đến cho con người những giá trị đáng kinh ngạc trong việc bảo vệ sức khỏe và điều trị bệnh tật. Tại lĩnh vực này, trí tuệ nhân tạo có vai trò quan trọng trong việc hỗ trợ điều trị y tế như định lượng thuốc, các phương pháp điều trị khác nhau cho bệnh nhân và quy trình phẫu thuật trong phòng mổ. Chúng sử dụng những thuật toán phân tích để hỗ trợ bệnh nhân theo dõi kết quả điều trị 24/7.

Vai trò của AI trong tài chính

Hinh 1. 4. Áp dụng AI trong lĩnh vực tài chính

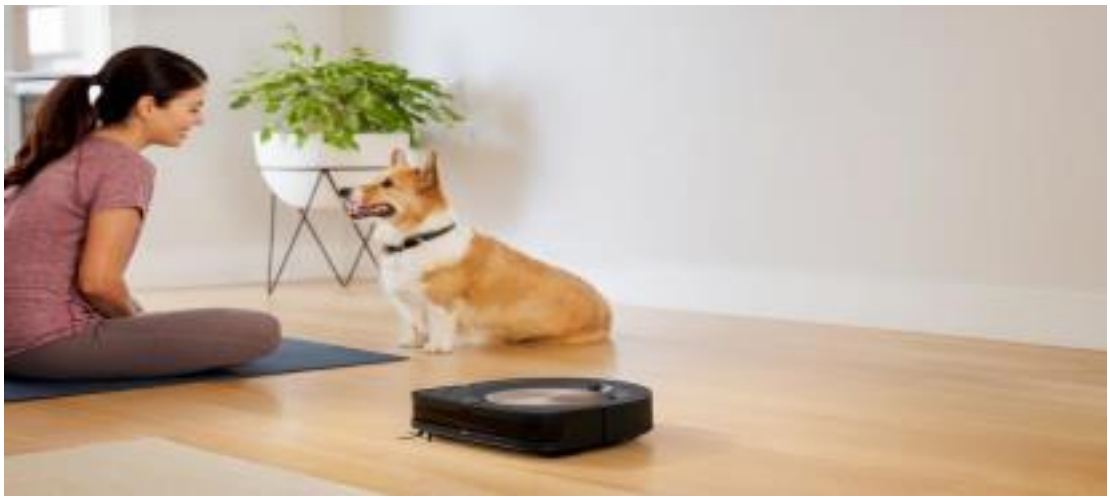
Ngoài việc hỗ trợ con người chăm sóc sức khỏe, AI còn có vai trò quan trọng trong ngành tài chính ngân hàng. AI là công cụ giúp con người xử lý các hoạt động trong ngân hàng như xử lý giao dịch, theo dõi số dư, quản lý tài sản và các tài khoản tiền gửi lớn một cách nhanh chóng và chính xác nhất. Trí tuệ nhân tạo không những giúp các ngân hàng hợp lý hóa giao dịch mà còn có thể ước tính cung, cầu và định giá chứng khoán một cách dễ dàng hơn.

Vai trò của AI trong trò chơi và công nghệ

Hiện nay, những tập đoàn lớn đang ngày càng thúc đẩy việc sử dụng máy móc thông minh vào dây chuyền sản xuất. AI được sử dụng như các robot có thể thay thế một phần công việc của con người. Khối lượng công việc và thời gian hoàn thành sẽ nhanh chóng và nhẹ

nhàng hơn dưới sự hoạt động của máy móc tích hợp trí tuệ nhân tạo. Tiêu biểu là với các sản phẩm như ô tô tự lái và trò chơi điện tử. Trong trò chơi điện tử, trí tuệ nhân tạo AI sẽ tự phân tích các hành vi và đưa ra những đáp án không kém cạnh với trí tuệ con người. Với ô tô tự lái, hệ thống AI tính toán tất cả các dữ liệu bên trong động cơ, tìm hiểu cách đi và ngăn chặn va chạm bởi chướng ngại vật

Sự kết hợp hoàn hảo của AI và robot hút bụi



Hinh 1. 5. Áp dụng AI để sản xuất robot hút bụi

Khi mọi người nghe đến trí tuệ nhân tạo, điều đầu tiên họ thường nghĩ đến là robot. Đối với lĩnh vực dọn dẹp tự động hóa gia đình, AI là điều không thể thiếu. Kết hợp các công nghệ tiên tiến cùng công nghệ AI siêu thông minh, các dòng máy robot hút bụi tự động liên tục được ra mắt trên thị trường. Tiêu biểu là dòng robot hút bụi Roomba của iRobot. Các sản phẩm tích hợp AI thường là những công cụ cao cấp nhất, đem lại hiệu quả cực lớn trong việc làm sạch sàn nhà của các hộ gia đình.

Với thời đại công nghệ 4.0 hiện nay, việc ứng dụng AI không còn xa lạ gì với cuộc sống của chúng ta. Trí tuệ nhân tạo có mặt trong mọi lĩnh vực đời sống từ giải trí cho đến y tế, xã hội. Đây chính là chìa khóa để mở ra một thế hệ mới đầy văn minh, thúc đẩy sự phát triển to lớn của loài người.

So sánh giữa lập trình không có TTNT và lập trình có TTNT

Lập trình không có TTNT	Lập trình có TTNT
Chương trình máy tính mà không có Trí tuệ nhân tạo thì chỉ có thể trả lời những câu hỏi xác định được quy định sẵn để giải quyết vấn đề.	Chương trình máy tính mà có Trí tuệ nhân tạo thì có thể trả lời những câu hỏi chung, cùng loại để giải quyết vấn đề.
Chỉnh sửa chương trình dẫn đến sự thay đổi trong cấu trúc của nó.	Chương trình Trí tuệ nhân tạo có thể tiếp thu sự ập nhật cái mới bằng cách đề cao tính độc lập của những thông tin với nhau. Vì vậy bạn có thể sửa đổi một phần thông tin trong chương trình mà không làm ảnh hưởng đến cấu trúc của nó.
Việc chỉnh sửa thường không nhanh và không dễ dàng. Nó dẫn đến việc ảnh hưởng chương trình của bạn	Chỉnh sửa chương trình nhanh và dễ dàng.

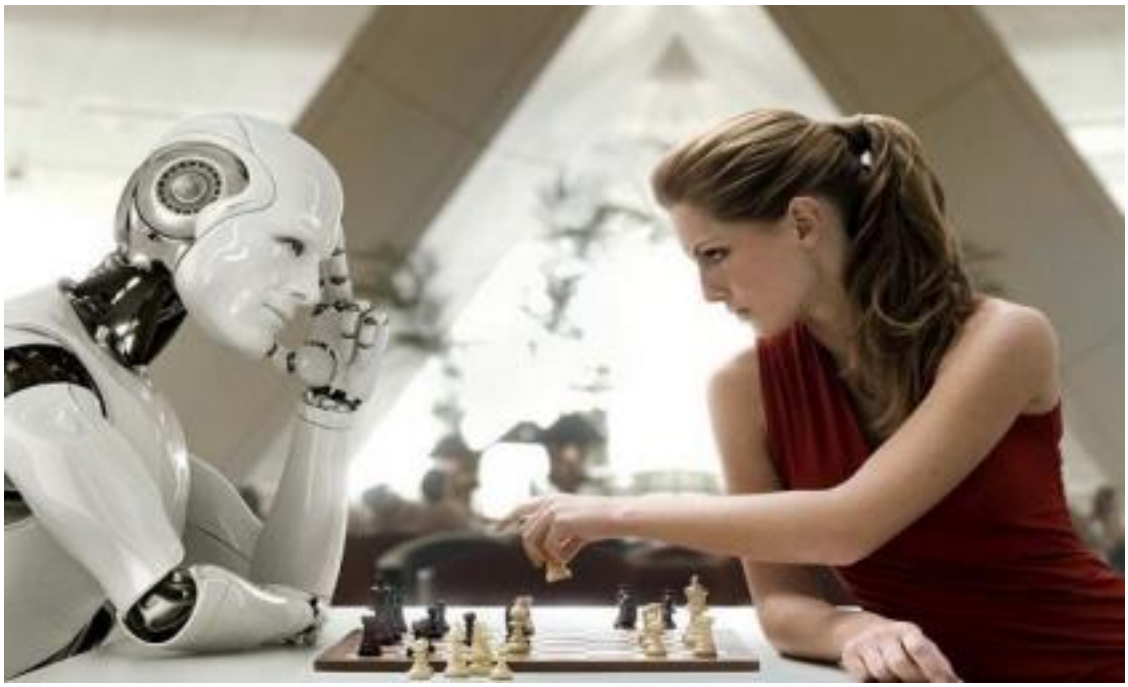
Những tác động của TTNT đến sản xuất trong nền công nghiệp 4.0 như sau:

- Chất lượng – Năng suất dự đoán : Vai trò của trí tuệ nhân tạo đầu tiên là giảm thiểu các hao tổn trong sản xuất và ngăn ngừa các quy trình sản xuất kém hiệu quả. Khi nhu cầu ngày càng tăng để đáp ứng sự cạnh tranh thì trí tuệ nhân tạo là điều vô cùng cần thiết.

- Bảo trì sự dự đoán : Một trong những lợi ích của trí tuệ nhân tạo nữa là bảo trì dự đoán. Thay vì việc bảo trì theo lịch trình định trước thì bảo trì dự đoán sẽ sử dụng

thuật toán để dự đoán lỗi tiếp theo của một bộ phận/máy móc/hệ thống. Nhờ đó có thể cảnh báo nhân viên thực hiện các quy trình bảo trì tập trung để ngăn chặn sự cố. Bảo trì dự đoán có ưu điểm là giảm đáng kể chi phí trong khi loại bỏ nhu cầu về thời gian ngừng hoạt động theo kế hoạch trong nhiều trường hợp. Ngoài ra, nhờ nó mà Tuổi thọ hữu dụng còn lại của máy móc và thiết bị lâu hơn.

● Kết hợp giữa robot và con người



Hinh 1. 6. Robot thay thế con người trong một số công việc

Tính đến năm 2020, ước tính có khoảng 1,64 triệu robot công nghiệp đang hoạt động trên toàn thế giới. Robot sản xuất được chấp thuận làm việc cùng với con người để tăng năng suất công việc.

Khi áp dụng robot ngày càng nhiều thì AI sẽ đóng một vai trò quan trọng trong việc đảm bảo an toàn cho con người. Đồng thời trao cho robot nhiều trách nhiệm hơn trong việc đưa ra các quyết định có thể tối ưu hóa các quy trình dựa trên dữ liệu thời gian thực được thu thập từ sàn sản xuất.

● Thiết kế sáng tạo : Nhà sản xuất có thể tận dụng trí tuệ nhân tạo vào giai đoạn thiết kế. Khi có bản tóm tắt thiết kế được xác định rõ ràng làm đầu vào thì các nhà kỹ sư, thiết kế có thể sử dụng thuật toán al. Mục đích để khám phá tất cả các cấu hình có thể có của một giải pháp.

● Nhu cầu cung ứng thị trường : Vai trò của trí tuệ nhân tạo cuối cùng mà chúng tôi muốn nhắc đến là cung ứng thị trường. Hiện nay trí tuệ nhân tạo đang hiện hữu ở mọi nơi trong hệ sinh thái công nghiệp 4.0. Nhà sản xuất có thể sử dụng các thuật toán AI để tối ưu hóa chuỗi cung ứng của các hoạt động sản xuất. Đồng thời giúp họ phản ứng và dự đoán tốt hơn những thay đổi trên thị trường.

1.3. Kĩ thuật trong TTNT là gì và một số kĩ thuật cơ bản trong TTNT

Trong thế giới thực, Tri thức có một vài thuộc tính như sau:

● Dung lượng đồ sộ, phi thường.

● Tổ chức tốt, định dạng tốt.

● Luôn luôn cập nhật sự thay đổi.

Kỹ thuật Trí tuệ nhân tạo là một cách để tổ chức và sử dụng tri thức có hiệu quả trong những cách sau đây:

● Có thể nhận thức được người đã cung cấp cho nó.

● Có thể sửa đổi dễ dàng để sửa lỗi.

● Nó có thể hữu ích trong một số tình huống dù nó chưa hoàn thiện hoặc chưa chính xác lắm.

Kỹ thuật Trí tuệ nhân tạo nâng cao tốc độ thực thi của những chương trình phức tạp.

Một số kỹ thuật Trí tuệ nhân tạo cơ bản :

● Lý thuyết giải bài toán và suy diễn thông minh

- Lý thuyết tìm kiếm may rủi
- Các ngôn ngữ về TTNT
- Lý thuyết thể hiện tri thức và hệ chuyên gia
- Lý thuyết nhận dạng và xử lý tiếng nói
- Người máy
- Tâm lý học xử lý thông tin
- Xử lý danh sách, kỹ thuật đệ quy, kỹ thuật quay lui và xử lý cú pháp hình thức

1.4. Lịch sử phát triển của trí tuệ nhân tạo

Đây là lịch sử của Trí tuệ nhân tạo trong suốt thế kỷ XX.

Năm	Cột mốc/ Phát minh
1923	Vở kịch khoa học viễn tưởng của Karel Capek tên là "Rossum's Universal Robots" (RUR) diễn ra tại Luân Đôn (nước Anh). Lần đầu tiên sử dụng từ "robot" trong tiếng Anh.
1943	Nền tảng của mạng thần kinh được đặt nền móng.
1945	Isaac Asimov, một cựu sinh viên trường Đại học Columbia, đưa ra thuật ngữ "Robotics"
1950	Alan Turing giới thiệu Bài kiểm tra Turing để đánh giá sự thông minh và công bố Máy thông minh và Sự thông minh. Claude Shannon công bố "Phân tích chi tiết của việc chơi cờ".
1956	John McCarthy đưa ra thuật ngữ Trí tuệ nhân tạo. Biểu diễn chạy chương

	trình trí tuệ nhân tạo đầu tiên tại trường Đại học Carnegie Mellon.
1958	John McCarthy sáng tạo ra LISP, ngôn ngữ lập trình cho trí tuệ nhân tạo.
1964	Bài luận văn của Danny Bobrow tại MIT cho thấy máy tính có thể hiểu được ngôn ngữ tự nhiên của con người.
1965	Joseph Weizenbaum tại MIT đã xây dựng ELIZA, một vấn đề tương tác được mang trong đoạn đối thoại Tiếng Anh.
1969	Cá nhà khoa học tại Viện nghiên cứu Stanford đã phát triển Shakey, một robot, được trang bị sự vận động, nhận thức, và giải quyết vấn đề.
1973	Các nhóm hội về người máy tại Đại học Edinburgh đã xây dựng Freddy. Một người máy Scotland nổi tiếng, có khả năng sử dụng thị giác để định vị và lắp ráp mô hình.
1979	Xe tự quản được điều khiển bằng máy tính đầu tiên được xây dựng. Đó là Stanford Cart.
1985	Harold Cohen tạo và trình diễn chương trình đồ họa mang tên Aaron.
1985	Những chuyên đề nâng cao trong tất cả các lĩnh vực của Trí tuệ nhân tạo là: ● Có tính chất quan trọng trong "học máy". ● Suy luận theo tình huống ● Lên lịch trình ● Khai thác dữ liệu, thu thập web ● Hiểu và dịch ngôn ngữ tự nhiên của con người ● Thị giác và thực tế ảo

	● Ứng dụng trong trò chơi
1997	Chương trình "Deep Blue Chess" đánh bại nhà vô địch cờ thế giới, Garry Kasparov.
2000	Những robot thú cưng có sự tương tác đã được thương mại hóa. MIT đã trình diễn Kismet - một robot có khuôn mặt có thể biểu lộ cảm xúc. robot Nomad khám phá những vùng xa xôi hẻo lánh của Nam Cực và xác định thiên thạch.

1.5. Các thành phần trong hệ thống của trí tuệ nhân tạo

Hệ thống trí tuệ nhân tạo bao gồm hai thành phần cơ bản đó là biểu diễn tri thức và tìm kiếm tri thức trong miền biểu diễn:

TTNT = Tri thức + Suy diễn

Tri thức của bài toán có thể được phân ra làm ba loại cơ bản đó là tri thức mô tả, tri thức thủ tục và tri thức điều khiển.

Để biểu diễn tri thức người ta sử dụng các phương pháp sau đây:

- Phương pháp biểu diễn nhờ luật
- Phương pháp biểu diễn nhờ mạng ngữ nghĩa
- Phương pháp biểu diễn nhờ bộ ba liên hợp OAV
- Phương pháp biểu diễn nhờ Frame
- Phương pháp biểu diễn nhờ logic vị tư

Sau khi tri thức của bài toán đã được biểu diễn, kỹ thuật trong lĩnh vực trí tuệ nhân tạo là các phương pháp tìm kiếm trong miền đặc trưng tri thức về bài toán đó. Với mỗi cách biểu diễn sẽ có các giải pháp tương ứng. Các vấn đề này sẽ được đề cập trong chương 3.

1.6. Các lĩnh vực nghiên cứu và ứng dụng cơ bản

Trí tuệ nhân tạo có những ảnh hưởng vượt trội trong nhiều lĩnh vực như:

- Game - Trí tuệ nhân tạo đóng vai trò cốt yếu trong những game chiến lược như cờ, đánh bài, tic-tac-toe (như cờ caro), ... nơi mà máy móc có thể suy nghĩ số lớn những trường hợp có khả năng xảy ra dựa trên tri thức.

- Xử lý ngôn ngữ tự nhiên - Nó có khả năng tương tác với máy tính, hiểu ngôn ngữ tự nhiên mà con người nói.

- Hệ thống chuyên môn hóa - Có một vài ứng dụng mà các máy móc thông minh, phần mềm và những thông tin đặc biệt để suy luận. Nó giải thích và đưa ra lời khuyên cho người dùng hệ thống đó.

- Hệ thống thị giác - Hệ thống có thể hiểu, phân tích và tiếp thu dữ liệu vào thuộc về thị giác ngay trên máy tính. Ví dụ như:

    - Những máy bay do thám chụp lại hình ảnh, sau đó sử dụng kỹ thuật này để mô hình hóa những thông tin không gian hay bản đồ của khu vực.

    - Bác sĩ sử dụng hệ thống buồng bệnh chuyên môn để chẩn đoán cho bệnh nhân.

    - Cảnh sát có thể sử dụng phần mềm máy tính để nhận diện khuôn mặt của tội phạm từ những hình chân dung được vẽ lại bởi những họa sĩ pháp y.

- Nhận diện lời nói - Một vài hệ thống thông minh có khả năng nghe và tiếp thu ngôn ngữ trong cấu trúc và nghĩa của câu trong khi con người nói. Nó có thể nắm bắt được độ nhấn mạnh khác nhau, từ lóng, tiếng ồn phía sau, sự thay đổi trong âm thanh của con người do trời lạnh, ...

- Nhận diện chữ viết tay - Phần mềm nhận diện chữ viết tay đọc văn bản được viết trên giấy bằng bút hoặc viết trên màn hình bằng bút cảm ứng. Nó nhận dạng được hình dạng của chữ và chuyển nó thành văn bản có thể chỉnh sửa được.

● Người máy thông minh - Người máy có khả năng thực hiện nhiệm vụ mà con người giao cho. Nó có các cảm biến để nhận dạng các dữ liệu vật lý trong thế giới thực như ánh sáng, hơi nóng, nhiệt độ, sự di chuyển, âm thanh, sự va chạm và áp lực. Nó được trang bị bộ xử lý hiệu quả, đa cảm biến và bộ nhớ lớn để thể hiện sự thông minh. Hơn thế nữa, nó có khả năng học từ lỗi sai của nó và thích nghi với môi trường mới.

Chương 2. Tìm hiểu về thuật toán Naïve Bayes

2.1. Định lý Bayes

2.1.1. Xác suất có điều kiện

Xác suất có điều kiện là khả năng đúng của một sự kiện dựa trên một sự kiện nào đó.

Ta có 2 sự kiện ngẫu nhiên A và B.

- Nếu A, B là 2 điều kiện độc lập, ta có xác suất để xảy ra A và B đồng thời là:

$$P(A, B) = P(A)P(B) \quad (1)$$

- Nếu A và B là 2 sự kiện liên quan đến nhau, và xác suất xảy ra sự kiện B lớn hơn 0, ta có thể định nghĩa xác suất xảy ra A khi biết B xảy ra như sau:

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

Ta có thể viết lại thành:

$$P(A,B) = P(A|B)P(B)$$

2.1.2. Định lý Bayes

Định lý Bayes dựa trên định nghĩa về xác suất có điều kiện ở trên, được phát biểu dưới dạng công thức như sau:

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

Trong đó:

+ $\boldsymbol{P(A)}$: Xác suất xảy ra A của riêng nó, không quan tâm đến B. Nó được gọi là "tiên nghiệm" với hàm ý rằng nó không quan tâm tới bất kỳ thông tin nào của B.
+ $\boldsymbol{P(B)}$: Xác suất xảy ra B của riêng nó, không quan tâm đến A.
+ $\boldsymbol{P(A|B)}$: xác suất xảy ra biến cố A với điều kiện xảy ra biến cố B.

+ $\boldsymbol{P(B|A)}$: Xác xuất xảy ra biến cố B với điều kiện xảy ra biến cố A.

- Công thức xác suất toàn phần: Nếu $B_1 + B_2 + \ldots + B_n = \square$ và $B_iB_j = \emptyset\ \forall i \neq j$, khi đó với biến cố A liên quan được tính theo công thức:

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

- Áp dụng công thức trên, ta có:

Giả sử $\{A_1, A_2, A_3, \ldots A_n\}$ là hệ đầy đủ và B là một sự kiện bất kì có thể xảy ra trong phép thử. Khi đó ta có công thức Bayes:

$$P(A_k|B) = \frac{P(A_k)P(B|A_{k)}}{P(B)} = \frac{P(A_k)P(B|A_{k)}}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}, k = 1,2, \ldots, n$$

Ví dụ:

Dây chuyền lắp ráp máy vô tuyến điện gồm các linh kiện là sản phẩm từ 2 nhà máy sản xuất ra. Số linh kiện nhà máy 1 sản xuất chiếm 55%, số linh kiện nhà máy 2 sản xuất chiếm 45%; tỷ lệ sản phẩm đạt tiêu chuẩn của nhà máy 1 là 90%, nhà máy 2 là 87%. Lấy ngẫu nhiên ra 1 linh kiện từ dây chuyền lắp ráp đó ra kiểm tra thì được kết quả linh kiện đạt chuẩn. Tìm xác suất để linh kiện đó do nhà máy 1 sản xuất?

Gọi A_i = "linh kiện do nhà máy thứ i sản xuất", i = 1, 2.

Gọi B = "linh kiện đạt chuẩn".

Ta cần tìm $P(A_1|B)$.

- Xác suất lấy được linh kiện của nhà máy 1:

$$P(A_1) = 0{,}55$$

- Xác suất lấy được linh kiện của nhà máy 2:

$$P(A_2) = 0{,}45$$

- Xác suất lấy được linh kiện đạt chuẩn của nhà máy 1:

$$P(B|A_1) = 0,9$$

- Xác suất lấy được linh kiện đạt chuẩn của nhà máy 2:

$$P(B|A_2) = 0,87$$

- Xác suất để lấy được linh kiện đạt chuẩn:

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) = 0,55.0,9 + 0,45.0,87 = 0,8865.$$

Vậy, xác suất để lấy được linh kiện của nhà máy một với điều kiện đạt chuẩn:

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(B)} = \frac{0,55.0,9}{0,8865} = 0,5583.$$

2.2. Phân lớp Naïve Bayes

- Thuật toán *Naïve Bayes* 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 dựa trên một tập dữ liệu đã được thu thập, vì nó khá dễ hiểu và độ chính xác cao.

- Thuật toán *Naive Bayes* tính xác suất cho các yếu tố, sau đó chọn kết quả với xác suất cao nhất.

2.2.1. Bài toán tổng quát

Xét bài toán phân lớp. Cho một tập dữ liệu huấn luyện $X \epsilon R^{n \times (m+1)}$ gồm n mẫu dữ liệu, mỗi mẫu có m thuộc tính và một thuộc tính lớp. Mỗi mẫu huấn luyện $x \in X$ được biểu diễn là một vectơ $m + 1$ chiều $x(x_1, x_2, \dots x_m, y)$ trong đó gồm m thành phần dữ liệu và y là nhãn lớp. Cho một tập xác định các nhãn lớp $C = \{c_1, c_2, \dots, c_q\}$ gồm q lớp. Có thể thấy $y \in C$. Cho một mẫu dữ liệu mới $z \in R^m$ và z được biểu diễn bằng $z(z_1, z_2, \dots, z_m)$. Hãy xác định lớp của z.

- Để xác định được lớp của z, ta cần tính được xác suất xảy ra khả năng z được phân vào từng lớp c_i, $i = 1 \dots q$, tức khả năng xảy ra c_i. Mẫu z sẽ được phân vào lớp nào có xác suất xảy ra cao nhất.
- Tuy nhiên, mẫu z là xác định với các thành phần quan sát được là $z_1, z_2, \dots, z_m$. Do đó, xác suất để z thuộc vào lớp c_i phải là xác suất có điều kiện $P(c_i|z_1, z_2, \dots, z_m)$ và được ký hiệu là $P(c_i|z)$. Theo định lý Bayes, xác suất này được tính:

$$P(c_i|z) = \frac{P(z|c_i) \times P(c_i)}{P(z)} \quad (1)$$

- Do các thuộc tính của z là độc lập và có điều kiện đối với các thuộc tính khác (tính chất của Naïve Bayes) nên:

$$P(z|c_i) = \prod_{j=1}^{m} P(z_j|c_i) \quad (2)$$

- Do đó (1) trở thành:

$$P(c_i|z) = \frac{\prod_{j=1}^{m} P(z_j|c_i)\, P(c_i)}{P(z)} \quad (3)$$

- Mẫu dữ liệu z sẽ được phân vào lớp c_k nếu $P(c_k|z)$ là lớn nhất tức:

$$c_k = argmax\, P(c_i|z)\ c_i \in C\ = argmax\ \ c_i \in C\ \frac{\prod_{j=1}^{m} P(z_j|c_i) \times P(c_i)}{P(z)} \quad (4)$$

- Vì $P(z)$ là hằng số đối với các c_i khác nhau, do vậy (4) sẽ tương đương với:

$$c_k = argmax\ c_i \in C\ \prod_{j=1}^{m} P(z_j|c_i) \times\ P(c_i) \quad (5)$$

Nói cách khác, để xác định lớp cho mẫu dữ liệu z, ta lần lượt tính các giá trị của biểu thức $\prod_{j=1}^{m} P(z_j|c_i) \times\ P(c_i)$ với từng lớp $c_i \in \{c_1, c_2, \dots, c_q\}$. Lớp c_i nào cho giá trị biểu

thức lớn nhất sẽ là lớp của z. Quá trình phân lớp sử dụng phương pháp *Naive Bayes* gồm hai bước:

+ **Bước 1:** Đối với mỗi lớp $c_i \in C$, tính giá trị của:
    - Xác suất tiên nghiệm $P(c_i)$. Xác suất này được tính xấp xỉ bằng tổng số mẫu thuộc lớp c_i trên tổng số mẫu của bộ dữ liệu huấn luyện.
    - Đối với mỗi giá trị thuộc tính z_i, tính $P(z_j|c_i)$ là xác suất xảy ra của giá trị đó trong lớp c_i. Giá trị này cũng được tính xấp xỉ bằng tỷ lệ các mẫu có giá trị trên thuộc tính thứ j là z_j trong số các mẫu thuộc lớp c_i.
+ **Bước 2:** Cần xác định lớp cho một mẫu dữ liệu mới z, ta thực hiện:
    - Đối với mỗi lớp $c_i \in C$, ta tính giá trị của biểu thức:

$$\prod_{j=1}^{m} P(z_j|c_i) \times P(c_i)$$

    - Xác định lớp của z là c_k:

$$c_k = argmax\ c_i \in C \prod_{j=1}^{m} P(z_j|c_i) \times P(c_i)$$

2.2.2. Bài toán với dữ liệu liên tục

– Trong trường hợp thuộc tính có giá trị liên tục, ta có thể áp dụng các phương pháp rời rạc hóa. Nếu không rời rạc hóa dữ liệu, thay vì xác suất, ta sử dụng hàm mật độ xác suất. Thông thường, ta hay giả thiết là dữ liệu trong mỗi lớp c_i của các thuộc tính liên tục tuân theo phân bố *Gauss* và phương pháp lúc này được gọi là *Gauss Naive Bayes.*

– Xét thuộc tính A với các giá trị liên tục. Khi đó, ta phân đoạn các giá trị của A theo từng lớp. Với mỗi lớp c_i, ta tính μ_i là giá trị trung bình và σ_i^2 là phương sai của các giá trị của A trong lớp c_i (với N_i là số mẫu thuộc lớp c_i và lớp y_i là lớp của mẫu x_i):

(1)

$$\mu_i = \frac{1}{N_i} \sum_{x_i : y_i = c_i} x_i$$

$$\sigma_i^2 = \frac{1}{N_i - 1} \sum_{x_i : y_i = c_i} (x_i - \mu_i)^2$$

– Giá trị $P(x|c_i)$ khi đó gọi là phân bố xác suất của x vào lớp c_i và được tính bằng:

$$P(x|c_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \quad (2)$$

2.2.3. Bài toán với xác suất điều kiện bằng không

−Xét một mẫu dữ liệu cần phân lớp $x(x_1, x_2, \dots, x_m)$. Xét giá trị x_j trên thuộc tính j. Nếu không có mẫu dữ liệu nào trong lớp c_i có giá trị trên thuộc tính j là x_j thì hiển nhiên $P(c_i) = 0$. Điều này kéo theo $P(c_i) \times \prod_{j=1}^{m} P(x_i|c_i) = 0$.

−Giải pháp được đưa ra là sử dụng ước lượng *Laplace* để ước lượng $P(x_i|c_i)$ thay cho giá trị 0 đã tính được ở trên.

−Ta giả sử số mẫu dữ liệu trong lớp cần xét là lớn và do đó nếu ta bổ sung 1 mẫu dữ liệu cho mỗi tập thuộc tính thì việc này sẽ không ảnh hưởng nhiều tới xác suất đã tính.

2.2.4. Ví dụ minh hoạ

Để mọi thứ được rõ rang hơn, chúng ta cùng xem ví dụ với dữ liệu huấn luyện được cho trong Bảng dưới đây. Bảng dữ liệu này được lấy từ cuốn sách Data Mining: Practical Machine Learning Tools and Techniques, trang 11. Bảng dữ liệu này mô tả mối quan hệ giữa thời tiết trong 14 ngày (bốn cột đầu, không tính cột id) và việc một đội bóng có chơi bóng hay không (cột cuối cùng). Nói cách khác, ta phải dự đoán giá trị ở cột cuối cùng nếu biết giá trị của bốn cột còn lại.

Id	*Outlook*	*Temperature*	*Humidity*	*Wind*	*Play*

1	*Sunny*	*Hot*	*High*	*Weak*	*No*
2	*Sunny*	*Hot*	*High*	*Strong*	*No*
3	*Overcast*	*Hot*	*High*	*Weak*	*Yes*
4	*Rainy*	*Mild*	*High*	*Weak*	*Yes*
5	*Rainy*	*Cool*	*Normal*	*Weak*	*Yes*
6	*Rainy*	*Cool*	*Normal*	*Strong*	*No*
7	*Overcast*	*Cool*	*Normal*	*Strong*	*Yes*
8	*Sunny*	*Mild*	*High*	*Weak*	*No*
9	*Sunny*	*Cool*	*Normal*	*Weak*	*Yes*
10	*Rainy*	*Mild*	*Normal*	*Weak*	*Yes*
11	*Sunny*	*Mild*	*Normal*	*Strong*	*Yes*
12	*Overcast*	*Mild*	*High*	*Strong*	*Yes*
13	*Overcast*	*Hot*	*Normal*	*Weak*	*Yes*
14	*Rainy*	*Mild*	*High*	*Strong*	*No*
15	*Rainy*	*Cool*	*High*	*Strong*	*?*

Bảng 1. Bảng dữ liệu huấn luyện

Có bốn thuộc tính thời tiết:

1. *Outlook* nhận một trong ba giá trị: sunny, overcast, rainy.
2. *Temperature* nhận một trong ba giá trị: hot, cool, mild.
3. *Humidity* nhận một trong hai giá trị: high, normal.
4. *Wind* nhận một trong hai giá trị: weak, strong.

(Tổng cộng có 3×3×2×2 = 36 loại thời tiết khác nhau, trong đó 14 loại được thể hiện trong bảng.)

Chúng ta sẽ training cho bộ dữ liệu này bằng thuật toán Naïve Bayes.

Cho dữ liệu cần phân lớp là Z. Xác định xem với trạng thái thời tiết Z có dữ liệu như trên thì đội bóng có chơi bóng hay không.

Thấy số mẫu dữ liệu n = 14, số thuộc tính của dữ liệu m = 4, thuộc tính lớp là Class với các lớp C = {c_1 = "yes", c_2 = "no"} (2 lớp). Quá trình xác định lớp cho dữ liệu Z = {z_1 = "rainy", z_2="$cool$", z_3= "$high$", z_4 = "$strong$"} trải qua hai bước:

- **Bước 1:** Đối với mỗi lớp $c_i \epsilon C$. Ta có:
  - Xác xuất tiên nghiệm của $\boldsymbol{P}(c_i)$:
    - $\boldsymbol{P}(c_1) = \frac{9}{14}$
    - $\boldsymbol{P}(c_2) = \frac{5}{14}$
  - Đối với mỗi thuộc tính z_j, tính $P(z_j|c_i)$, j = 1...4
    - $\boldsymbol{P}(z_j|c_1)$

$$P(z_1|c_1) = \frac{1}{3}; P(z_2|c_1) = \frac{1}{3};$$

$$P(z_3|c_1) = \frac{1}{3}; P(z_4|c_1) = \frac{1}{3};$$

    - $\boldsymbol{P}(z_j|c_2)$

$$P(z_1|c_2) = \frac{2}{5}; P(z_2|c_2) = \frac{1}{5};$$

$$P(z_3|c_2) = \frac{4}{5}; P(z_4|c_2) = \frac{3}{5};$$

- **Bước 2:** Thực hiện đối với mỗi lớp c_i, tính biểu thức:

$$\prod_{j=1}^{m} P(z_j|c_i) \times P(c_i)$$

  - Với $C = c_1$:

$$P(z_1|c_1) \times P(z_2|c_1) \times P(z_3|c_1) \times P(z_4|c_1) \times P(c_1)$$

$$= \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \times \frac{9}{14} = \frac{1}{126} \approx 0{,}0079 \quad (1)$$

- Với $C = c_2$:

$$P(z_1|c_2) \times P(z_2|c_2) \times P(z_3|c_2) \times P(z_4|c_2) \times P(c_2)$$

$$= \frac{2}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14} = \frac{12}{875} \approx 0{,}0137 \quad (2)$$

- Từ (1) và (2) => Z sẽ thuộc lớp c_2

Vậy mẫu dữ liệu Z sẽ thuộc vào lớp c_2 là lớp đội bóng sẽ không chơi.

2.3. Ưu điểm, hạn chế

- Ưu điểm:

- Đơn giản và dễ triển khai: Navie Bayes là một thuật toán đơn giản và dễ hiểu. Nó yêu cầu ít tham số và dễ dàng triển khai trong các bài toán phân loại.
- Xử lý hiệu quả với dữ liệu lớn: Dựa trên giả thuyết độc lập giữa các đặc trưng, Navie Bayes có thể xử lý hiệu quả dữ liệu lớn và tốn ít tài nguyên tính toán. Thuật toán hoạt động nhanh chóng và thích hợp cho các tập dữ liệu có số lượng lớn các đặc trưng.
- Độc lập với tỷ lệ: Navie Bayes không bị ảnh hưởng bởi tỷ lệ của các lớp trong dữ liệu huấn luyện. Ngay cả khi tỷ lệ các lớp không cân bằng, thuật toán vẫn có thể đưa ra các dự đoán chính xác.
- Hoạt động tốt với dữ liệu thiếu: Naïve Bayes có khả năng xử lý dữ liệu thiếu một cashc linh hoạt. Nếu một mẫu thiếu một giá trị đặc trưng, thuật toán sẽ bỏ qua giá trị đó và tiếp tục tính toán dựa trên các giá trị còn lại.
- Naïve Bayes phù hợp với các biến đầu vào phân loại hơn là các biến số.

- Hạn chế:

- Giả định độc lập: Naïve Bayes giả định rằng các đặc trưng đầu vào là độc lập với nhau. Tuy nhiên, trong thực tế, có rất ít các bài toán mà các đặc trưng thực sự độc lập. Giả định này có thể làm giảm độ chính xác của thuật toán trong một số trường hợp.

- Không xử lý được các đặc trưng mới: Naïve Bayes không thể xử lý các đặc trưng trong tập kiểm tra mà không xuất hiện trong tập huấn luyện. Nếu một đặc trưng mới xuất hiện, thuật toán sẽ gán xác suất 0 cho các lớp liên quan đến đặc trưng đó.
- Nhạy cảm với dữ liệu nhiễu: Naïve Bayes có thể nhạy cảm với dữ liệu nhiễu hoặc các đặc trưng không phù hợp. Các đặc trưng không phù hợp có thể làm giảm độ chính xác của thuật toán.
- Giới hạn trong bài toán phân loại: Naïve Bayes thường được sử dụng trong bài toán phân loiaj và đòi hỏi các lớp phải có các đặc trưng rời rạc hoặc liên tục nhất định. Nếu có một số đặc trưng không phù hợp với giả định này, Naïve Bayes có thể không cho kết quả tốt.
- Ước tính của nó có thể sai trong một số trường hợp, vì vậy bạn không nên quá coi trọng kết quả xác suất của nó.

Chương 3. Ứng dụng Naive Bayes trong bài toán dự báo thời tiết

3.1. Bài toán

3.1.1. Phát biểu bài toán

Bài toán dự báo thời tiết (dự đoán xem trời có mưa hay không) sử dụng phương pháp Naive Bayes dựa vào dữ liệu Nhiệt độ, Sức gió, lượng mưa để phân ra hai lớp là trời có mưa hay không mưa.

3.1.2. Thuật toán áp dụng

Bài toán áp dụng thuật toán Naïve Bayes trong trường hợp tổng quát. Bằng cách tính xác suất $P(z_j|c_i)$ của mỗi thuộc tính z_j trong từng class c_i. Sau đó, tính $P(c_i|z)$ là xác suất có thể dự đoán ra class c_i với điều kiện là bộ dữ liệu z và chọn phân bố vào class có xác suất này lớn nhất.

Bài toán được làm bằng ngôn ngữ lập trình Python – một ngôn ngữ lập trình bậc cao cho các mục đích lập trình đa năng. Python có bộ thư viện chuẩn rộng lớn – đó là một trong những điểm lớn mạnh của ngôn ngữ lập trình này. Trong bài toán này chúng em thực hiện dựa theo công thức tổng quát của Naïve Bayes để giải quyết.

3.2. Cơ sở dữ liệu

Cơ sở dữ liệu được lưu trong file .csv do nhóm tự sinh dựa vào một bảng ví dụ đã có trong giáo trình.

Thông tin, ký hiệu các thuộc tính của các bản ghi trong cơ sở dữ liệu:

1. ID trạng thái thời tiết tại 1 thời điểm (integer)
2. Nhiệt độ (Hot, Cold, Cool)
3. Sức gió (Strong, Weak, No)
4. Độ ẩm (High, low)

5. Mưa (Yes, No)

Tập dữ liệu được lưu dưới dạng file CSV trong đó các cột tương ứng với các thuộc tính trên.

	ID	Temp	Wind	Moisture	Rain
2	1	Hot	Strong	High	Yes
3	2	Cold	Weak	High	Yes
4	3	Cool	No	Low	No
5	4	Hot	Strong	High	Yes
6	5	Hot	Strong	Low	No
7	6	Hot	Strong	Low	No
8	7	Hot	Strong	Low	No
9	8	Cold	Strong	Low	Yes
10	9	Cold	Strong	Low	No
11	10	Cold	Weak	Low	No
12	11	Cold	Weak	Low	No
13	12	Cold	Weak	Low	Yes
14	13	Cold	Weak	Low	No
15	14	Hot	Weak	High	Yes
16	15	Hot	Weak	High	Yes
17	16	Hot	Weak	High	Yes
18	17	Hot	Weak	High	No
19	18	Hot	Strong	High	No
20	19	Hot	Strong	Low	Yes
21	20	Hot	Strong	Low	No

Bảng 2. Bảng mẫu dữ liệu trong file CSV

3.3. Cài đặt

3.3.1. Thuật toán

- Đầu vào:
  + Bộ dữ liệu để dự báo mưa
  + Các trạng thái thời tiết
- Đầu ra:
  + Tỷ lệ có mưa hoặc không có mưa
  + Kết luận dự đoán

3.3.2. Cài đặt

- Đọc dữ liệu từ file CSV.

```
#import thu vien
import pandas as pd

#doc du lieu
df = pd.read_csv('data_rain.csv')
temp_list = df['Temp']
wind_list = df['Wind']
moisture_list = df['Moisture']
rain_list = df['Rain']
```

Hình 3. 1. Đọc dữ liệu từ file CSV

- Tính toán các P(c1), P(c2), và các phân lớp t,w,m

```
#xet phan lop (t,w,m)
#Buoc 1, xet C1 = yes, tinh P(Yes) va P(x|c1)
#tinh P(Yes)
p_yes = len(df[df['Rain']=='Yes'])/len(df)
#tinh P(No)
p_no = len(df[df['Rain']=='No'])/len(df)
```

Hình 3. 2. Tính P(c1), P(c2)

- Tính các tích P(x|c1) tương ứng với trạng thái có mưa

```
#tinh cac P(x|c1)
def p_x_c_yes(t,w,m):
    res=[]
    k=1
    res.append(len(df[df['Rain']=='Yes'][df['Temp']==t])/len(df[df['Rain']=='Yes']))
    res.append(len(df[df['Rain']=='Yes'][df['Wind']==w])/len(df[df['Rain']=='Yes']))
    res.append(len(df[df['Rain']=='Yes'][df['Moisture']==m])/len(df[df['Rain']=='Yes']))
    for i in res:
        k=k*i
    return k
```

Hình 3. 3. Hàm tính P(x|c1)

- Tương tự hàm tính tích P(x|c2) tương ứng với trạng thái không mưa

```
def p_x_c_no(t,w,m):
    res=[]
    k=1
    res.append(len(df[df['Rain']=='No'][df['Temp']==t])/len(df[df['Rain']=='No']))
    res.append(len(df[df['Rain']=='No'][df['Wind']==w])/len(df[df['Rain']=='No']))
    res.append(len(df[df['Rain']=='No'][df['Moisture']==m])/len(df[df['Rain']=='No']))
    for i in res:
        k=k*i
    return k
```

Hình 3. 4. Hàm tính P(x|c2)

- Hàm đưa ra giá trị lớn hơn và kết luận x thuộc phân lớp nào

```
#tinh P(c1)* tich cac P(xi|c1) va P(c2)* tich cac P(xi|c2), dua ra gia tri lon hon va ket luan
def decision(t,w,m):
    res= max(p_x_c_yes(t,w,m)*p_yes,p_x_c_no(t,w,m)*p_no)
    if res == p_x_c_yes(t,w,m)*p_yes:
        print("Du doan se co mua")
    else:
        print("Du doan se khong mua")
```

Hình 3. 5. Hàm dự đoán phân lớp x

- Dự đoán thời tiết với các phân lớp cụ thể

```
#ket qua du doan voi cac phan lop
print('BAI TOAN DU BAO THOI TIET SU DUNG NAVIE BAYES')
print('--------------------')
print("Kha nang co mua khi troi nong, gio to, do am cao: ",p_x_c_yes('Hot','Strong','High')*p_yes)
print("Kha nang co mua khi troi nong, gio to, do am cao: ",p_x_c_no('Hot','Strong','High')*p_no)
decision('Hot','Strong','High')
print('--------------------')
print("Kha nang co mua khi troi lanh, gio to, do am cao: ",p_x_c_yes('Cold','Strong','High')*p_yes)
print("Kha nang co mua khi troi lanh, gio to, do am cao: ",p_x_c_no('Cold','Strong','High')*p_no)
decision('Cold','Strong','High')
print('--------------------')
print("Kha nang co mua khi troi mat, khong gio, do am thap: ",p_x_c_yes('Cool','No','Low')*p_yes)
print("Kha nang co mua khi troi mat, khong gio, do am thap: ",p_x_c_no('Cool','No','Low')*p_no)
decision('Cool','No','Low')
```

Hình 3. 6. Hàm dự đoán thời tiết dựa vào phân lớp cụ thể

- Dự đoán thời tiết với các phân lớp được nhập từ bàn phím

```
print('--------------------')
print("Nhap nhiet do: ")
t=input()
print("Nhap suc gio: ")
w=input()
print("Nhap do am: ")
m=input()

def prediction(t,w,m):
    return decision(t,w,m)

prediction(t,w,m)
```

s

Hình 3. 7. Hàm dự đoán thời tiết với phân lớp được nhập từ bàn phím

- Kết quả sau khi chạy.

```
BAI TOAN DU BAO THOI TIET SU DUNG NAVIE BAYES
--------------------
Kha nang co mua khi troi nong, gio to, do am cao:  0.04248813068318983
Kha nang khong mua khi troi nong, gio to, do am cao:  0.01892115040594468
Du doan se co mua
--------------------
Kha nang co mua khi troi lanh, gio to, do am cao:  0.015933049006196187
Kha nang khong mua khi troi lanh, gio to, do am cao:  0.007568460162377872
Du doan se co mua
--------------------
Kha nang co mua khi troi mat, khong gio, do am thap:  0.01158767200450632
Kha nang khong mua khi troi mat, khong gio, do am thap:  0.07802394385578643
Du doan se khong mua
```

```
--------------------
Nhap nhiet do:
Hot
Nhap suc gio:
Strong
Nhap do am:
High
Du doan se co mua
```

Hình 3. 8. Kết quả sau khi chạy

Tổng kết

Naive Bayes 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 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 dựa trên một tập dữ liệu đã được thu thập, vì nó khá dễ hiểu và độ chính xác cao.Thuật toán Naïve Bayes sử dụng xác suất để dự đoán lớp hoặc nhãn của một mẫu dựa trên các đặc trưng đầu vào. Ý tưởng chính của một mẫu dựa trên các đặc trưng đầu vào, tính toán xác suất có điều kiện của một lớp dựa trên đặc trưng đầu vào và áp dụng định lý Bayes để tính toán xác xuất đồng thời

Chúng ta có thể ứng dụng Naive Bayes 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.

Tài liệu tham khảo

[1]. Trường Đại học Công Nghiệp Hà Nội, Giáo trình trí tuệ nhân tạo

[2]. Nam Doan, *Định lý Bayes*, 1upnote.me/post/2018/11/ds-ml-bayes-theorem/

[3]. Thân Ngọc Thiện, Kiều Đức Anh, Phan Tiến Phương, *Tổng quan về trí tuệ nhân tạo, "ID3 và ứng dụng dự đoán nhu cầu mở tài khoản ký quỹ"*, 2022

[4]. Đặng Đức Mạnh, *Cơ sở dữ liệu*, github.com/leomanhlol/ttnt/blob/master/data_rain.csv, 2023