TRƯỜNG ĐẠI HỌC ĐIỆN LỰC

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO CHUYÊN ĐỀ HỌC PHẦN

HỌC MÁY NÂNG CAO

ĐỀ TÀI: DỰ ĐOÁN THỜI TIẾT SỬ DỤNG THUẬT TOÁN LOGISTIC REGRESSION VÀ GIẢM CHIỀU DỮ LIỆU PCA

Sinh viên thực hiện : **NGUYỄN THẾ ĐỒNG**

ĐỖ THỊ BÍCH NGỌC

QUÁCH QUANG MINH

Giảng viên hướng dẫn : **ĐÀO NAM ANH**

Ngành : **CÔNG NGHỆ THÔNG TIN**

Chuyên ngành : **CÔNG NGHỆ PHẦN MỀM**

Lớp : **D14CNPM5**

Khóa : **2019-2024**

Hà Nội, tháng 12 năm 2022

PHIẾU CHẤM ĐIỂM

Sinh viên thực hiện:

STT	Họ và tên sinh viên	Nội dung thực hiện	Điểm	Chữ ký
1	Nguyễn Thế Đồng 1910310229			
2	Đỗ Thị Bích Ngọc 19810310223			
3	Quách Quang Minh 19810310365			

Giảng viên chấm

Họ và tên giảng viên	Chữ kí	Ghi chú
Giảng viên chấm 1:		
Giảng viên chấm 2:		

MỤC LỤC

LỜI MỞ ĐẦU

Công nghệ ngày càng phổ biến và không ai có thể phủ nhận được tầm quan trọng và những hiệu quả mà nó đem lại cho cuộc sống chúng ta. Bất kỳ trong lĩnh vực nào, sự góp mặt của trí tuệ nhân tạo sẽ giúp con người làm việc và hoàn thành tốt công việc hơn. Và gần đây, một thuật ngữ "Machine learning" rất được nhiều người quan tâm. Thay vì phải code phần mềm với cách thức thủ công theo một bộ hướng dẫn cụ thể nhằm hoàn thành một nhiệm vụ đề ra thì máy sẽ tự "học hỏi" bằng cách sử dụng một lượng lớn dữ liệu cùng những thuật toán cho phép nó thực hiện các tác vụ.

Đây là một lĩnh vực khoa học tuy không mới, nhưng cho thấy lĩnh vực trí tuệ nhân tạo đang ngày càng phát triển và có thể tiến xa hơn trong tương lai. Đồng thời, thời điểm này nó được xem là một lĩnh vực "nóng" và dành rất nhiều mối quan tâm để phát triển nó một cách mạnh mẽ, bùng nổ hơn.

Hiện nay, việc quan tâm Machine learning càng ngày càng tăng lên là vì nhờ có Machine learning giúp gia tăng dung lượng lưu trữ các loại dữ liệu sẵn, việc xử lý tính toán có chi phí thấp và hiệu quả hơn rất nhiều.

Những điều trên được hiểu là nó có thể thực hiện tự động, nhanh chóng để tạo ra những mô hình cho phép phân tích các dữ liệu có quy mô lớn hơn và phức tạp hơn đồng thời đưa ra những kết quả một cách nhanh và chính xác hơn.

Chính sự hiệu quả trong công việc và các lợi ích vượt bậc mà nó đem lại cho chúng ta khiến machine learning ngày càng được chú trọng và quan tâm nhiều hơn. Vì vậy chúng em đã chọn đề tài "***Dự đoán thời tiết sử dụng thuật toán Logistic Regression và giảm chiều dữ liệu PCA***" để có thể thực hành và nghiên cứu sâu hơn về môn "Học máy nâng cao".

Chúng em xin chân thành cảm ơn thầy Đào Nam Anh đã tận tình giảng dạy, truyền đạt cho chúng em những kiến thức cũng như kinh nghiệm quý báu trong suốt quá trình học.

CHƯƠNG 1: TỔNG QUAN VỀ HỌC MÁY NÂNG CAO

1.1. Học máy là gì ?

Học máy là lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể. Ví dụ: Hãy xem xét một hệ thống lọc thư rác email, thay vì để các lập trình viên xem xét email theo cách thủ công và đưa ra các quy tắc spam. Chúng ta có thể sử dụng thuật toán học máy và cung cấp cho nó dữ liệu đầu vào (email) và nó sẽ tự động khám phá các quy tắc đủ mạnh để phân biệt các emai spam.

Một chương trình máy tính được xem là học cách thực thi một lớp nhiệm vụ thông qua đó trả nhiệm, đối với các thang đo năng lượng nếu như dùng năng lực ta đo thấy năng lực thực thi của chương trình có tiến bộ sau khi trải qua quá trình trải nghiệm.

Học máy được sử dụng trong nhiều ứng dụng hiện nay như phát hiện thư rác trong email hoặc hệ thống đề xuất phim cho bạn những bộ phim mà bạn có thể thích dựa trên lịch sử xem của minh. Điều thú vị và mạnh mẽ về học máy là: Nó học kho nhận được nhiều dữ liệu hơn và do đó nó càng mạnh mẽ hơn khi chúng ta cung cấp cho họ nhều dữ liệu hơn.

1.2. Các loại học máy

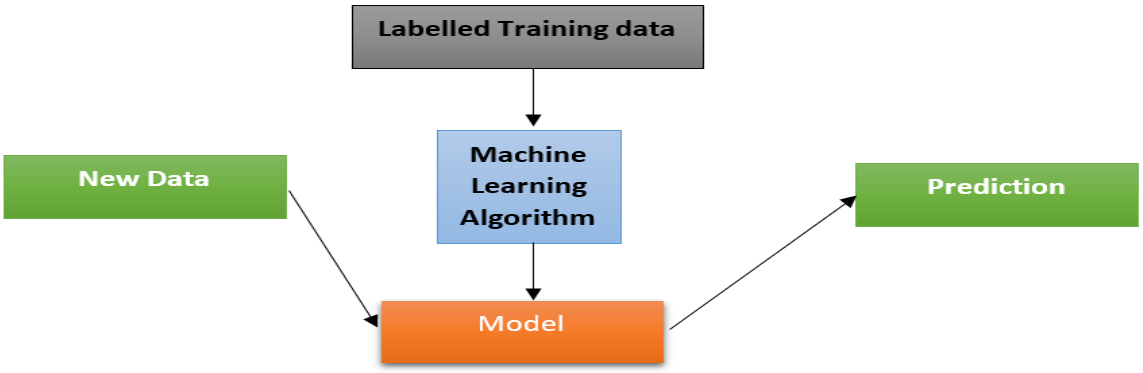
Có ba loại máy học khác nhau: học có giám sát, không giám sát và học tăng cường.

1.2.1. Học máy có giám sát

Mục tiêu của học máy có giám sát là học một mô hình tư dữ liệu đào tạo được gắn nhãn cho phép chúng đưa ra dự đoán về dữ liệu trong tương lai. Để học máy có giám sát hoạt động, chúng ta cần cung cấp cho thuật toán hai thứ: dữ liệu đầu vào và kiến thức của chúng ta về nhãn đó.

Ví dụ về bộ lọc thư rác được đề cập trước đó là một ví dụ điển hình về việc học có giám sát; Chúng có một loạt các email(dữ liệu) và chúng tỗi biết liệu mỗi email có phải là thư rác hay không(nhãn)

Học có giám sát có thể được chia thành hai danh mục phụ:

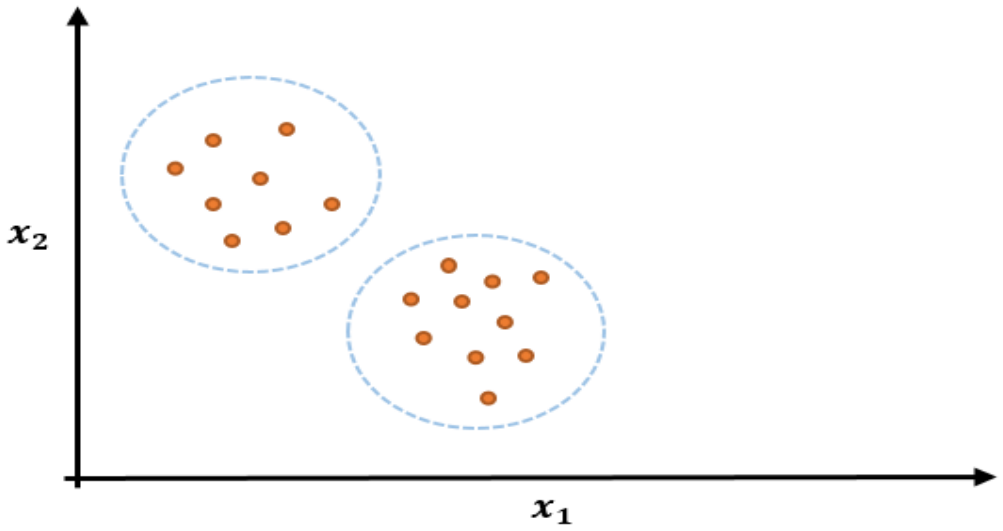


- Phân loại: Nó được sử dụng để dự đoán các danh mục hoặc nhãn lớp dựa trên các quan sát trong quá khứ, tức là chugns ta có biến rời rạc mà bạn muốn phân biệt thành kết quả phân loại rời rạc. Ví dụ: Trong hệ thống lọc thư rác email, đầu ra là "thư rác" rời rạc hoặc "Không phải thư rác".
- Hồi quy: Nó được sử dụng để dự đoán một kết quả liên tục. Ví dụ để xác định giá nhà và nó bị ảnh hưởng như nào bởi só lượng phòng trong ngôi nhà đó. Dữ liệu đầu vào là các tính năng của nôi nhà và đầu ra là giá cả.

1.2.2. Học máy không giám sát

Mục tiêu của việc học máy không giám sát là khám phá cấu trúc hoặc mẫu ẩn trong dữ liệu không được gắn nhãn và nó có thể được chia thnahf hai danh mục phụ

- Phân cụm: Nó được sử dụng để tổ chức thông tin thành các cụm(nhóm con) có ý nghĩa mà không cần biết trước ý nghĩa của chúng



- Giảm kích thước: Nó được sử dụng để giảm dữ liệu kích thước cao hơn thành dữ liệu thứ nguyên thấp hơn.

1.2.3. Học máy tăng cường

Mục tiêu cảu học máy tăng cường là phát triển một hệ thống cải thiện hiệu suất của nó dựa trên tương tác với môi trường năng động và có một phản hồi chậm trễ hoạt động như một phần thưởng. Tức là học máy tăng cường là học bằng cách làm phần thưởng bị trì hoãn. Một ví dụ kinh điển về học máy tăng cường là một trò chơi cờ vua, máy tính đã quyết định một loạt các nước đi và phần thưởng là "thắng" hoặc "thua" ở cuối trò chơi.

1.3.Các bước học máy

1.3.1. Tiền xử lý dữ liệu

Chất lượng của dữ liệu và lượng thông tin hữu ích mà nó chứa ảnh hưởng lớn đến mức độ một thuật toán có thể học được. Do đó, điều quan trọng là phải xử lý trước tập dữ liệu trước khi sử dụng nó. Các bước tiền xử lý phổ biến nhất là: loại bỏ các giá trị bị thiếu, chuyển đổi dữ liệu phân loại thành hình dạng phù hợp với thuật toán học máy và thay đổi quy mô tính năng.

- Thiếu dữ liệu:

Đôi khi các mẫu trong tập dữ liệu bị thiếu một số giá trị và chúng tôi muốn xử lý các giá trị bị thiếu này trước khi chuyển nó vào thuật toán học máy. Có một số chiến lược chúng ta có thể làm theo

+ Xóa các mẫu có giá trị bị thiếu: Cách tiếp cận này cho đến nay là thuận tiện nhất nhưng cuối cùng chúng tôi có thể xóa quá nhiều mẫu và do đó chúng tôi sẽ mất thông tin có giá trị có thể giúp ích cho thuật toán học máy.

+ Gán các giá trị bị thiếu: Thay vì loại bỏ toàn bộ mẫu, chúng tôi sử dụng nội suy để ước tính các giá trị bị thiếu.

- Dữ liệu phân loại

Nói chung, các đối tượng địa lý có thể là số (ví dụ: giá cả, chiều dài, chiều rộng, v.v.) hoặc phân loại (ví dụ: màu sắc, kích thước, v.v.). Các đặc điểm phân loại được chia thành các tính năng danh nghĩa và thứ tự.

Các tính năng thứ tự có thể được sắp xếp và sắp xếp. Ví dụ: kích thước (nhỏ, vừa, lớn), chúng ta có thể đặt hàng các kích thước này lớn > trung bình > nhỏ. Mặc dù các đối tượng địa lý danh nghĩa không có thứ tự chẳng hạn như màu sắc, nhưng sẽ không có ý nghĩa gì khi nói rằng màu đỏ lớn hơn màu xanh lam.

Hầu hết các thuật toán học máy yêu cầu bạn chuyển đổi các tính năng phân loại thành các giá trị số. Một giải pháp sẽ gán cho mỗi giá trị một số khác nhau bắt đầu từ số không. (ví dụ: nhỏ à 0 ,trung bình à 1 ,lớn à 2)

Điều này hoạt động tốt cho các đặc điểm thứ tự nhưng có thể gây ra vấn đề với các tính năng danh nghĩa (ví dụ: màu xanh lam à 0, màu trắng à 1, màu vàng à 2) bởi vì mặc dù màu sắc không được sắp xếp, thuật toán học tập sẽ cho rằng màu trắng lớn hơn màu xanh lam và màu vàng lớn hơn màu trắng và điều này không chính xác.

Để khắc phục vấn đề này là sử dụng mã hóa một nóng, ý tưởng là tạo ra một tính năng mới cho mỗi giá trị duy nhất của tính năng danh nghĩa.

#	Color
0	Red
1	Green
2	Blue
3	Red
4	Blue

#	Red	Green	Blue
0	1	0	0
1	0	1	0
2	0	0	1
3	1	0	0
4	0	0	1

Trong ví dụ trên, chúng tôi đã chuyển đổi tính năng màu sắc thành ba tính năng mới *Đỏ,* Xanh *lá cây, Xanh lam* và chúng tôi đã sử dụng các giá trị nhị phân để biểu thị *màu sắc*. Ví dụ: một mẫu có màu "Đỏ" hiện được mã hóa thành (Đỏ = 1, Xanh lục = 0, Xanh lam = 0)

- Mở rộng quy mô tính năng

Giả sử chúng ta có dữ liệu với hai tính năng, một trên thang điểm từ 1 đến 10 và tính năng còn lại trên thang điểm từ 1 đến 1000. Nếu thuật toán sử dụng khoảng cách như một phần hoạt động của nó hoặc nếu nó đang cố gắng giảm thiểu sai số bình phương trung bình, có thể an toàn khi cho rằng thuật toán sẽ dành phần lớn thời gian để tập trung vào tính năng 2.

Để giải quyết vấn đề này, chúng tôi sẽ sử dụng chuẩn hóa để đưa các tính năng khác nhau lên cùng một quy mô. Chúng tôi mở rộng quy mô các tính năng trong phạm vi [0, 1]

$$x_{norm}^{(i)} = \frac{x^{(i)} - x_{min}}{x_{max} - x_{min}}$$

$x^{(i)}$ is một mẫu cụ thể, x_{min} là giá trị nhỏ nhất trong tính năng và x_{max} là giá trị lớn nhất.

Một cách tiếp cận khác là tiêu chuẩn hóa

$$x_{std}^{(i)} = \frac{x^{(i)} - \mu}{\sigma_x}$$

Trong đó, μ là giá trị trung bình của tất cả các mẫu cho đối tượng địa lý cụ thể đó và σ_x là độ lệch chuẩn của đối tượng địa lý đó.

Bảng sau đây cho thấy kết quả của tiêu chuẩn hóa và chuẩn hóa trên dữ liệu đầu vào mẫu

Input	Min	Max	Normalization	Mean	Std. Dev.	Standardization
89	69	94	0.8	81	10.7	0.74
72			0.12			-0.84
94			1.00			1.21
69			0.00			-1.21

1.3.2. Đào tạo và kiểm tra dữ liệu

Chúng tôi thường chia dữ liệu đầu vào thành các bộ dữ liệu học tập và kiểm tra. Sau đó, chạy thuật toán học máy trên tập dữ liệu học tập để tạo mô hình dự đoán. Sau đó, chúng ta sử dụng tập dữ liệu thử nghiệm để đánh giá mô hình của mình.

Điều quan trọng là dữ liệu kiểm tra phải tách biệt với dữ liệu được sử dụng trong đào tạo nếu không chúng ta sẽ bị gian lận vì có thể ví dụ như mô hình được tạo sẽ ghi nhớ dữ liệu và do đó nếu dữ liệu kiểm tra cũng là một phần của dữ liệu đào tạo thì điểm đánh giá của chúng ta về mô hình sẽ cao hơn thực tế.

Dữ liệu thường được chia 75% đào tạo và 25% dữ liệu hoặc 2/3 đào tạo và 1/3 thử nghiệm. Điều quan trọng cần lưu ý là: *bộ đào tạo càng nhỏ thì thuật toán càng khó khám phá các quy tắc.*

Ngoài ra, khi tách tập dữ liệu, bạn cần duy trì tỷ lệ lớp và thống kê dân số nếu không chúng ta sẽ có một số lớp được trình bày trong tập dữ liệu đào tạo và được trình bày nhiều hơn trong tập dữ liệu thử nghiệm.

Ví dụ: bạn có thể có 100 mẫu và tổng cộng 80 mẫu được gắn nhãn Class-A và 20 phiên bản còn lại được gắn nhãn Class-B. Bạn muốn đảm bảo khi tách dữ liệu mà bạn duy trì biểu diễn này.

Một cách để tránh vấn đề này và đảm bảo rằng tất cả các lớp được thể hiện trong cả bộ dữ liệu đào tạo và thử nghiệm là phân tầng. Đó là quá trình sắp xếp lại dữ liệu để đảm bảo mỗi bộ là một đại diện tốt của toàn bộ. Trong ví dụ trước của chúng ta, (80/20 mẫu), tốt nhất là sắp xếp dữ liệu sao cho trong mỗi tập hợp, mỗi lớp bao gồm khoảng tỷ lệ 80:20 của hai lớp.

1.3.3. Xác thực chéo

Một bước quan trọng khi xây dựng mô hình máy học của chúng tôi là ước tính hiệu suất của nó trên đó mà mô hình chưa từng thấy trước đây. Chúng tôi muốn đảm bảo rằng mô hình khái quát hóa tốt với dữ liệu vô hình mới.

Một trường hợp, thuật toán học máy có các thông số khác nhau và chúng tôi muốn điều chỉnh các tham số này để đạt được hiệu suất tốt nhất. (Lưu ý: các tham số của thuật toán học máy được gọi là siêu tham số). Một trường hợp khác, đôi khi chúng tôi muốn thử các thuật toán khác nhau và chọn thuật toán hoạt động tốt nhất.

1.4.Các ứng dụng của học máy

Việc học máy nâng cao (Deep Learning) phát triển tạo nên sự chủ động trong mọi việc, con người dần có thể điều khiển cuộc sống của mình. Cùng điểm qua các hình thức mà việc học sâu mang lại

1.4.1 Mô phỏng và nhận diện hình ảnh

Chắc hẳn, chúng ta đều đã từng thấy máy tính tự động nhận diện và phân loại các hình ảnh của bạn. Ví dụ: Facebook có thể tự động gắn thẻ chính bạn và bạn bè của bạn. Tương tự, Google Photos có thể tự động gắn nhãn ảnh của bạn để tìm kiếm dễ dàng hơn.

Và với Deep Learning bạn có thể dễ dàng tìm và phân loại các hình ảnh theo ngày, sự kiện mà không phải dùng thao tác thủ công mất thời gian.

1.4.2. Trợ lý ảo

Ứng dụng phổ biến nhất của Học máy nâng cao (Deep Learning) ngày nay là trợ lý ảo từ Alexa đến Siri, Google Assistant. Mỗi tương tác với các trợ lý này cung cấp cho họ cơ hội tìm hiểu thêm về giọng nói và ngữ điệu của bạn, từ đó cung cấp cho bạn trải nghiệm tương tác như phiên bản thứ 2 của con người.

Trợ lý ảo sử dụng học tập sâu để biết thêm về các chủ đề của họ, từ sở thích ăn tối của bạn đến các điểm truy cập nhiều nhất hoặc các bài hát yêu thích

của bạn. Họ học cách hiểu các mệnh lệnh của bạn bằng cách đánh giá ngôn ngữ tự nhiên của con người để thực hiện chúng.

1.4.3. Thị giác máy tính

Là một lĩnh vực nghiên cứu rất sôi động hiện nay, với các phương pháp dựa trên năng lực tính toán ngày càng mạnh mẽ của hệ thống máy tính với các bài toán ứng dụng thực tiễn có giá trị to lớn. -Phương pháp sinh trắc học để nhận dạng các yếu tố của con người được nghiên cứu mạnh mẽ và ứng dụng vào hệ thống nhận dạng trên cơ sở các đặc điểm thể chất hoặc hành vi của mỗi người

CHƯƠNG 2: PHÂN TÍCH BÀI TOÁN

2.1. Phương pháp giảm chiều dữ liệu PCA

2.1.1. Giới thiệu

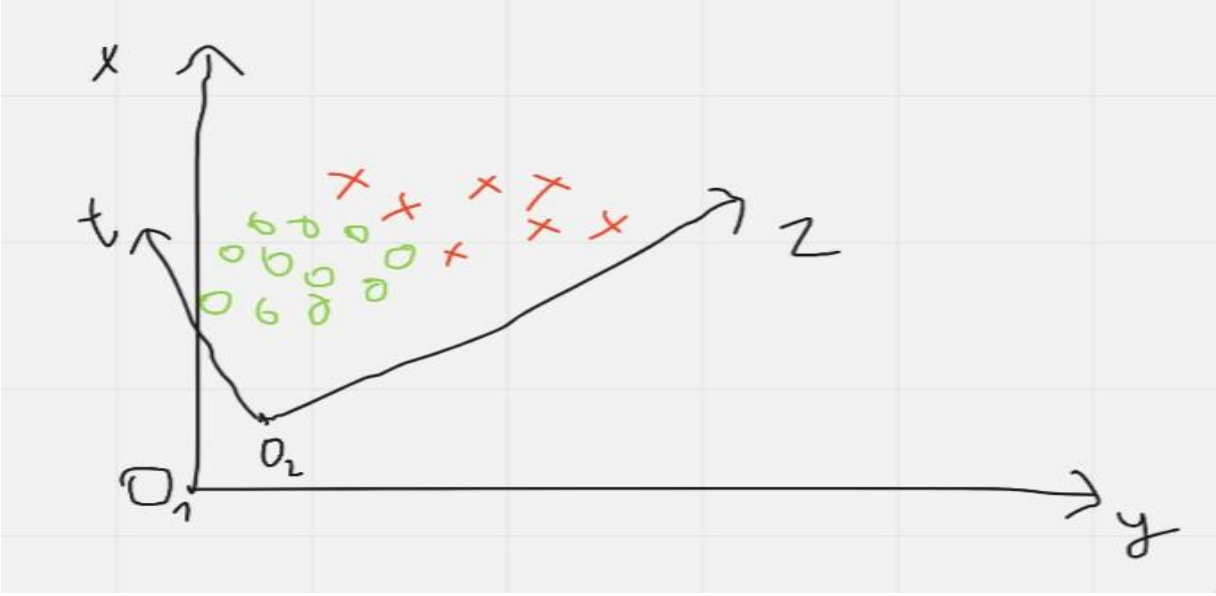
Trong các bài toán học máy thì dữ liệu có kích thước rất lớn. Máy tính có thể hiểu và thực thi các thuật toán trên dữ liệu này, tuy nhiên đối với con người để "nhìn" dữ liệu nhiều chiều thật sự là rất khó. Vì vậy bài toán giảm chiều dữ liệu ra đời giúp đưa ra cái nhìn mới cho con người về dữ liệu nhiều chiều. Ngoài để trực quan dữ liệu, các phương pháp giảm chiều dữ liệu còn giúp đưa dữ liệu về một không gian mới giúp khai phá các thuộc tính ẩn mà trong chiều dữ liệu ban đầu không thể hiện rõ, hoặc đơn giản là giảm kích thước dữ liệu để tăng tốc độ thực thi cho máy tính.

2.1.2. Định nghĩa

PCA là viết tắt của cụm từ *principal component analysis*. Thuật ngữ Tiếng Việt còn gọi là *phân tích thành phần chính*. Đây là một phương pháp giảm chiều dữ liệu (*dimensionality reduction*) tương đối hiệu quả dựa trên phép phân tích suy biến (*singular decomposition*) mà ở đó chúng ta sẽ chiếu các điểm dữ liệu trong không gian **cao chiều** xuống một số ít những véc tơ thành phần chính trong không gian **thấp chiều** mà đồng thời vẫn bảo toàn tối đa **độ biến động** của dữ liệu sau biến đổi. Ưu điểm của PCA đó là **sử dụng tất cả** các biến đầu vào nên phương pháp này không bỏ sót những biến quan trọng.

2.1.3. Nội dung của phương pháp

Về mặt ý tưởng, thuật toán PCA tìm một hệ không gian mới và tối đa hóa phương sai dữ liệu của không gian mới đó. Sau đó lựa chọn ra n chiều có phương sai lớn nhất (giả thuyết rằng dữ liệu càng phân tán, phương sai càng lớn thì càng có giá trị)



Hình 2.1. Hình minh họa cho PCA trong không gian (O_1xy)

Hình trên đây thể hiện được giá trị của phương sai, khi mà đối với không gian ban đầu (O_1xy) thì phần overlape của 2 lớp khi ánh xạ lên mỗi trục là khá lớn. Khi đó không gian mới (O_2zt) được cực đại hóa phương sai cho trục O_2z nên khi ánh xạ lên đây các lớp sẽ tách biệt với nhau khá rõ.

Để tìm được không gian mới, PCA đi tìm các trị riêng của ma trận hiệp phương sai của dữ liệu đầu vào. Các trị riêng thể hiện phương sai của chiều dữ liệu mới, các vector riêng ứng với trị riêng đó tương ứng với một không gian dữ liệu mới. Vậy nên sau bước này chúng ta chọn các vector riêng ứng với các trị riêng có giá trị lớn nhất để được một không gian mới được cực đại hóa phương sai.

2.1.4. Các bước cần thực hiện của thuật toán PCA

- *Bước 1*: Chuẩn bị dữ liệu cần giảm chiều là X với kích thước (n_sample, n_feature), tương ứng mỗi hàng là 1 mẫu dữ liệu có n_feature thuộc tính
- *Bước 2*: Trừ mỗi điểm dữ liệu cho vector kỳ vọng: $X_k = X_k - X_{mean}$ với k = 1..n_sample và X_{mean} là vector trung bình của tất cả các điểm dữ liệu
- *Bước 3*: Tính ma trận hiệu phương sai : $S = \frac{1}{n-sample} * X^T * X$
- *Bước 4*: Tìm trị riêng, vector riêng của ma trận S
- *Bước 5*: Lấy k trị riêng có giá trị lớn nhất, tạo ma trận U với các hàng là các vector riêng ứng với k trị riêng đã chọn
- *Bước 6*: Ánh xạ không gian ban đầu sang không gian k chiều: $X_{new} = X*U$
- *Ghi chú*: Nếu không hiểu phép nhân ở Bước 6 bạn có thể lấy từng mẫu dữ liệu nhân với từng vector riêng, khi đó mỗi mẫu dữ liệu ban đầu sẽ được nhân với k vector nên sẽ có k chiều.

2.2. Sử dụng phương pháp PCA để dự đoán thời tiết

2.2.1. Phát biểu bài toán

Dự báo thời tiết là bài toán có tính thực tiễn và có ý nghĩa quan trọng đối với ngành nông nghiệp, công nghiệp và dịch vụ. Ngày nay việc sử dụng phương pháp giảm chiều dữ liệu PCA để dự báo thời tiết (nhiệt độ lớn nhất và nhỏ nhất) trong ngày đang dần được phát triển

- Giá trị input: dữ liệu về thời tiết 4999 ngày vừa qua ở 1 địa điểm
- Giá trị ouput: có mưa hay không

2.2.2. Xây dựng bô dữ liệu

Giả sử một đài khí tượng thu thập được một tập dữ liệu lớn gồm báo cáo thời tiết 4999 ngày khác nhau, sau một thời gian thì xác định được gồm 3244 ngày không mưa và 1755 ngày có mưa

Dữ liệu gồm các đầu vào:

- Mintemp: nhiệt độ thấp nhất trong ngày
- Maxtemp: nhiệt độ cao nhất trong ngày

- Rainfall: lượng mưa
- Wind Gust Speed: tốc độ gió giật
- Wind Speed: tốc độ gió
- Humidity: độ ẩm
- Pressure: áp suất
- Cloud: mây
- Temp 9 am: nhiệt độ 9 giờ sang
- Temp 3 pm: nhiệt độ 3 giờ chiều

2.2.3. Các thư viện cần cài đặt

- Scikit-learn
- Numpy
- SciPy
- Matplotlib
- Ipython
- Sympy
- Pandas

CHƯƠNG 3: CÀI ĐẶT CHƯƠNG TRÌNH

Kết nối thư viện sử dụng

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import
accuracy_score,confusion_matrix,classification_report
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.decomposition import  PCA
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import LabelBinarizer
from sklearn.impute import SimpleImputer
from sklearn.impute import KNNImputer
```

Hình 3.1. Thêm thư viện

Đọc file csv

```
df = pd.read_csv('weather.csv')
df.head()
```

✓ 0.5s Python

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	...	Humidity3pm	Pressure9am
0	8.0	24.3	0.0	3.4	6.3	NW	30.0	SW	NW	6.0	...	29	1019.7
1	14.0	26.9	3.6	4.4	9.7	ENE	39.0	E	W	4.0	...	36	1012.4
2	13.7	23.4	3.6	5.8	3.3	NW	85.0	N	NNE	6.0	...	69	1009.5
3	13.3	15.5	39.8	7.2	9.1	NW	54.0	WNW	W	30.0	...	56	1005.5
4	7.6	16.1	2.8	5.6	10.6	SSE	50.0	SSE	ESE	20.0	...	49	1018.3

5 rows × 22 columns

Hình 3.2. Đọc file dữ liệu

Sử dụng PCA và đánh giá mô hình huấn luyện

```
from sklearn.decomposition import PCA
pc = PCA(n_components=2)
X2 = pc.fit_transform(X)
```

✓ 0.3s Python

```
X_train, X_test, y_train, y_test = train_test_split(X2, y, test_size=0.33, random_state=101)
```

✓ 0.7s Python

Hình 3.3. Đánh giá mô hình huấn luyện

Tính độ chính xác của trương trình

```
print(classification_report(y_test,predict_LR))
print(confusion_matrix(y_test,predict_LR))
print('\n')
print('Độ chính xác:', np.round(accuracy_score(y_test,predict_LR),3)*100,'%')
```

✓ 0.1s Python

Hình 3.4. Tính độ chính xác của chương trình

Kết quả chương trình

```
              precision    recall  f1-score   support

           0       0.86      0.98      0.92       103
           1       0.50      0.11      0.18        18

    accuracy                           0.85       121
   macro avg       0.68      0.55      0.55       121
weighted avg       0.81      0.85      0.81       121

[[101   2]
 [ 16   2]]


Độ chính xác: 85.1 %
gốc 700:      Humidity3pm  Rainfall  RainToday  RainTomorrow
200           36       0.0         0            0
201           52       0.0         0            0
202           47       0.0         0            0
203           43       0.0         0            0
204           49       0.0         0            0
..           ...       ...       ...          ...
325           25       0.0         0            0
326           50       0.0         0            1
327           44      17.4         1            0
328           33       0.0         0            0
329           32       0.0         0            0

[130 rows x 4 columns]
```

Hình 3.5. Kết quả của chương trình

KẾT LUẬN

Nhóm em đã hoàn thành ứng dụng thuật toán giảm chiều dữ liệu PCA trong dự báo ngày mưa. Tuy trong thực tế, còn phải xét rất nhiều yếu tố tác động lên kết quả, phân tích một khối lượng lớn dữ liệu, sử dụng nhiều phương pháp khác nhau để có được kết quả chính xác

Kết quả đạt được:

- ✓ Hiểu biết thêm về ngôn ngữ Python
- ✓ Hiểu biết thêm về thuật toán giảm chiều dữ liệu PCA, Logistic Regression và ứng dụng trong thực tế

TÀI LIỆU THAM KHẢO

1.Slide giảng dạy Ngôn ngữ lập trình Python – Thầy Đào Nam Anh
2. Giáo trình Machine Learning cơ bản-Vũ Hữu Tập, Nhà xuất bản khoa học và kỹ thuật