|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**  **KHOA ĐÀO TẠO CHẤT LƯỢNG CAO**  **NGÀNH CÔNG NGHỆ THÔNG TIN**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  http://t0.gstatic.com/images?q=tbn:ANd9GcR7BAz8hld8Rn4YZW5s_LsmbCivU780sKay-OhHY6sCfnmodrnQ  **BÁO CÁO ĐỒ ÁN 3**    **Đề tài:**  **Tìm hiểu mô hình hồi qui tuyến tính**  **(linear regression)**   |  |  | | --- | --- | | MSSV | Họ và tên | | 17110249 | Đào Sơn Tùng | | 17110206 | Trần Thành Phát |   **GVHD: Cô Từ Tuyết Hồng**  **TP. Hồ Chí Minh, ngày 15 tháng 12 năm 2020** |

**Mục Lục**

[**CHƯƠNG I: GIỚI THIỆU** 1](#_Toc58789880)

[**CHƯƠNG II: PHÂN TÍCH MÔ HÌNH** 1](#_Toc58789881)

[**1.** **Định nghĩa** 1](#_Toc58789882)

[**2.** **Mô hình** 2](#_Toc58789883)

[**3.** **Chọn hàm cơ bản** 3](#_Toc58789884)

[**3.1. Giữ nguyên đầu vào** 3](#_Toc58789885)

[**3.2. Chuẩn hoá đầu vào** 3](#_Toc58789886)

[**3.3. Đa thức hoá** 4](#_Toc58789887)

[**3.4. Sử dụng hàm Gaussian** 4](#_Toc58789888)

[**3.5. Sử dụng hàm sigmoid** 4](#_Toc58789889)

[**CHƯƠNG III: BÀI TOÁN** 5](#_Toc58789890)

[**1.** **Giới thiệu** 5](#_Toc58789891)

[**2.** **Xây dựng bộ dữ liệu** 5](#_Toc58789892)

[**3 Áp dụng giải thuật vào bài toán** 7](#_Toc58789893)

[**3.1 Cài đặt bài toán** 7](#_Toc58789894)

[**3.2 Kết quả dự đoán bài toán** 9](#_Toc58789895)

[**CHƯƠNG IV: KẾT LUẬN** 10](#_Toc58789896)

**CHƯƠNG I: GIỚI THIỆU**

Những năm gần đây, AI - Artificial Intelligence (Trí Tuệ Nhân Tạo), và cụ thể hơn là Machine Learning (Học Máy hoặc Máy Học) nổi lên như một bằng chứng của cuộc cách mạng công nghiệp lần thứ tư (1 - động cơ hơi nước, 2 - năng lượng điện, 3 - công nghệ thông tin). Trí Tuệ Nhân Tạo đang len lỏi vào mọi lĩnh vực trong đời sống mà có thể chúng ta không nhận ra. Xe tự hành của Google và Tesla, hệ thống tự tag khuôn mặt trong ảnh của Facebook, trợ lý ảo Siri của Apple, hệ thống gợi ý sản phẩm của Amazon, hệ thống gợi ý phim của Netflix, máy chơi cờ vây AlphaGo của Google DeepMind, …, chỉ là một vài trong vô vàn những ứng dụng của AI/Machine Learning.

"Hồi quy tuyến tính" là một phương pháp thống kê để hồi quy dữ liệu với biến phụ thuộc có giá trị liên tục trong khi các biến độc lập có thể có một trong hai giá trị liên tục hoặc là giá trị phân loại. Nói cách khác "Hồi quy tuyến tính" là một phương pháp để dự đoán biến phụ thuộc (Y) dựa trên giá trị của biến độc lập (X). Nó có thể được sử dụng cho các trường hợp chúng ta muốn dự đoán một số lượng liên tục. Ví dụ, dự đoán giao thông ở một cửa hàng bán lẻ, dự đoán thời gian người dùng dừng lại một trang nào đó hoặc số trang đã truy cập vào một website nào đó v.v...

**CHƯƠNG II: PHÂN TÍCH MÔ HÌNH**

* 1. **Định nghĩa**

Mục tiêu của giải thuật hồi quy tuyến tính là dự đoán giá trị của một hoặc nhiều biến mục tiêu liên tục (continuous target variable) y dựa trên một vector đầu vào x.

Để bắt đầu với Hồi quy tuyến tính, chúng ta hãy đi lướt qua một số khái niệm toán học về thống kê.

* Tương quan (r) - Giải thích mối quan hệ giữa hai biến, giá trị có thể chạy từ -1 đến +1
* Phương sai (σ2) - Đánh giá độ phân tán trong dữ liệu của bạn
* Độ lệch chuẩn (σ) - Đánh giá độ phân tán trong dữ liệu của bạn (căn bậc hai của phương sai)
* Phân phối chuẩn
* Sai số (lỗi) - {giá trị thực tế - giá trị dự đoán}

1. **Mô hình**

Mô hình đơn giản nhất là mô hình kết hợp tuyến tính của các biến đầu vào



trong đó:

: là vector biến đầu vào

: là vector trọng số tương ứng.

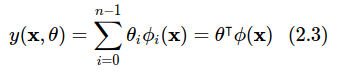
Thường θ được gọi là tham số của mô hình. Giá trị của tham số sẽ được ước lượng bằng cách sử dụng các cặp giá trị



của tập huấn luyện. Thực ra mô hình tuyến tính là chỉ cần ở mức tuyến tính giữa tham số θ và y là đủ. Và mình cho rằng tên gọi tuyến tính là xuất phát giữa θ và y, chứ không phải giữa x và y. Nói cách khác, ta có thể kết hợp các x một cách phi tuyến trước khi hợp với θ để được y. Một cách đơn giản là sử dụng hàm phi tuyến cho x như sau:



được gọi là độ lệch (bias) nhằm cắt đi mức độ chênh lệch giữa mô hình và thực tế. Các hàm phi tuyến  này được gọi là các hàm cơ bản (basic function). Thường người ta sẽ đặt  và viết lại công thức trên như sau:



1. **Chọn hàm cơ bản**

Việc chọn hàm cơ bản ϕ(x) cũng chính là chọn tính năng cho đầu vào rất quan trọng trong học máy. Ngoài ra việc chọn ra sao còn ảnh hưởng tới tốc độ và bộ nhớ để tính toán nữa. Ở đây tôi chỉ để cập tới 1 vài cách đơn giản để chọn hàm cơ bản mà thôi.

**3.1. Giữ nguyên đầu vào**

Giữ nguyên đầu vào có ý là không thay đổi giá trị đầu vào, tức: ϕ(x)=x Thường người ta sẽ gom các đầu vào thành một ma trận



Mỗi hàng của ma trận chứa 1 mẫu và mỗi cột sẽ chứa các thuộc tính đầu vào.

**3.2. Chuẩn hoá đầu vào**

Là phương pháp co giãn các thuộc tính về khoảng [min,max] nào đó (thường là [−1,1] hoặc [−0.5,0.5]) dựa vào kì vọng và độ lệch chuẩn của chúng.



Trong đó,  là trung bình, còn  là độ lệch chuẩn của tính năng i. Đôi lúc người ta cũng có thể lấy  là khoảng rộng chuẩn



Việc này không làm mất tính chất phân phối của chúng nên không ảnh hưởng tới kết quả học. Nhưng lại giúp cho việc học trở lên dễ dàng hơn vì các thuộc tính gần như cùng khoảng nhỏ với nhau. Phương pháp này còn có tên khác là chuẩn hoá trung bình (mean normalization).

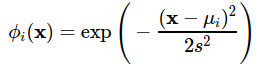
**3.3. Đa thức hoá**

Sử dụng đa thức bậc cao để làm đầu vào:



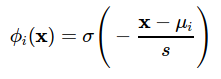
Với các bài toán hồi quy tuyến tính thì phương pháp này rất hay được sử dụng.

**3.4. Sử dụng hàm Gaussian**

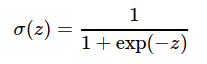


ở đây:sẽ chỉ định vị trí trung bình cho đầu vào còn s sẽ chỉ định độ phân tán cho đầu vào. Việc sử dụng hàm này sẽ giúp ta có được đầu vào theo phân phối chuẩn.

**3.5. Sử dụng hàm sigmoid**

Tương tự như hàm Gaussian, ta có thể sử dụng hàm sigmoid để biến đổi đầu vào:

Hàm sigmoid được sử dụng là sigmoid chuẩn:



Một biến thể khác là sử dụng tanhtanh vì nó khá gần với sigmoidsigmoid:



Ở phần sau chúng ta sẽ tiếp tục tìm hiểu về cách ước lượng tham số trong hàm lỗi, để đánh giá mức độ chênh lệch và thực hành 1 số ví dụ về Hồi quy tuyến tính.

**CHƯƠNG III: BÀI TOÁN**

* 1. **Giới thiệu**

Trong bài toán này sẽ phân tích dữ liệu trong thế giới thực gần đây từ hai trường trung học Bồ Đào Nha. Hai khác nhau-nguồn đã được sử dụng: đánh dấu báo cáo và bảng câu hỏi.

Vì trước đây chứa thông tin khan hiếm (tức là chỉ đã có điểm và số lần vắng mặt), nó đã được bổ sung với cái sau, cho phép tập hợp một số cơ sở nhân khẩu học, xã hội và trường học thuộc tính say mê (ví dụ: tuổi học sinh, mức tiêu thụ rượu-tion, giáo dục của mẹ). Mục đích là để dự đoán học sinh thành tích và nếu có thể xác định các biến chính ảnh hưởng đến sự thành công / thất bại trong giáo dục. Hai lõi các lớp học (tức là Toán và tiếng Bồ Đào Nha) sẽ được sửa đổi-đạt dưới ba mục tiêu VCK:

i) phân loại nhị phân (đạt / không đạt)

ii) phân loại với năm cấp độ (từ Tôi rất tốt hoặc xuất sắc đến V - không đủ) và

iii) hồi quy, với đầu ra số có phạm vi là-mười bảy không (0%) và hai mươi (100%).

Nghiên cứu này sẽ xem xét dữ liệu được thu thập trong năm 2005- năm học 2006 từ hai trường công lập, từ Alen-vùng tejo của Bồ Đào Nha. Mặc dù đã có một xu hướng để tăng cường đầu tư vào Công nghệ thông tin từ Chính phủ, phần lớn người Bồ Đào Nha hệ thống thông tin của trường công lập rất kém, chủ yếu là trên các tờ giấy (đó là trường hợp hiện tại).

* 1. **Xây dựng bộ dữ liệu**

Phân loại và hồi quy là hai DM quan trọng bàn thắng. Cả hai đều yêu cầu học tập có giám sát, trong đó một mô hình được điều chỉnh thành tập dữ liệu được tạo thành từ k ∈ {1, ..., N} ví dụ các mẫu, mỗi ánh xạ một vectơ đầu vào (x k 1, ..., x k Tôi ) để một mục tiêu y k cho trước . Sự khác biệt chính được đặt ra về của biểu diễn đầu ra, (tức là rời rạc cho classifi- cation và liên tục đối với hồi quy). Trong phân loại, các mô hình thường được đánh giá bằng cách sử dụng Phần trăm Cor- Phân loại trực tràng (PCC)

|  |  |
| --- | --- |
| Thuộc tính | Mô tả |
| Giới tính | Giới tính của học sinh |
| Tuổi | Tuổi của học sinh |
| Trường học | Trường học của học sinh |
| Địa chỉ | Địa chỉ nhà của sinh viên |
| Pstatus | Tình trạng chung sống của cha mẹ học sinh |
| Medu | Học vấn của mẹ học sinh |
| Mjob | Công việc của mẹ học sinh |
| Fjob | Công việc của cha |
| Người giám hộ | Người giám hộ của học sinh |
| famsize | Quy mô gia đình |
| famrel | Chất lượng mối quan hệ gia đình |
| Thời gian học | Thời gian học hàng tuần |
| Vi phạm | Số lỗi vi phạm của học sinh |
| Hoạt động | Các hoạt động ngoại khóa |
| Internet | Truy cập tại nhà riêng của học sinh |
| Sức khỏe | Tình trạng sức khỏe hiện tại của học sinh |
| Số lần nghỉ học | Số lần nghỉ học của học sinh |
| G1 | Lớp học đầu tiền |
| G2 | Lớp học thứ 2 |
| G3 | Lớp học cuối cùng |

Tất cả các thí nghiệm được báo cáo trong nghiên cứu này đã được thực hiện sử dụng RMiner (Cortez In press), một nguồn mở thư viện cho môi trường R tạo điều kiện cho việc sử dụng Kỹ thuật DM. R là miễn phí và cấp cao ngôn ngữ lập trình ma trận với một bộ các công cụ để phân tích thống kê và dữ liệu. Nó có thể chạy trên nhiều nền tảng (ví dụ: Windows, MacOS hoặc Linux) và các tính năng mới có thể thêm bằng cách tạo các gói. Thư viện RMiner 2 trình bày một tập hợp các func- tions (ví dụ: khai thác, saveMining) để phân loại và nhiệm vụ hồi quy (Cortez Trong báo chí). Đặc biệt, li-brary sử dụng rpart (DT), randomForest (RF), nnet (NN) và gói kernlab (SVM). Ví dụ, mã RMiner / R sau đây đã được sử dụng trong DT các thí nghiệm dự đoán của phân loại nhị phân Bồ Đào Nha

**3 Áp dụng giải thuật vào bài toán**

**3.1 Cài đặt bài toán**

*import numpy as np*

*import pandas as pd*

*import matplotlib.pyplot as plt*

*from sklearn.linear\_model import LinearRegression*

*# doc file*

*d1=pd.read\_csv("student-mat.csv",sep=";")*

*#d2=pd.read\_csv("student-por.csv",sep=";")*

*#d1=pd.concat([d1,d2])*

*# sửa dữ liệu chữ thành dang int*

*d1["school"]=d1["school"].replace({"GP": "0","MS": "1"})*

*d1["sex"]=d1["sex"].replace({"M": "0","F": "1"})*

*d1["address"]=d1["address"].replace({"U": "0","R": "1"})*

*d1["famsize"]=d1["famsize"].replace({"LE3": "0","GT3": "1"})*

*d1["Pstatus"]=d1["Pstatus"].replace({"T": "0","A": "1"})*

*d1["Mjob"]=d1["Mjob"].replace({"at\_home": "0","health": "1","services":"2","teacher":"3","other":"4"})*

*d1["Fjob"]=d1["Fjob"].replace({"at\_home": "0","health": "1","services":"2","teacher":"3","other":"4"})*

*d1["reason"]=d1["reason"].replace({"course": "0","home": "1","reputation": "2","other": "3"})*

*d1["guardian"]=d1["guardian"].replace({"father": "0","mother": "1","other": "2"})*

*d1["schoolsup"]=d1["schoolsup"].replace({"no": "0","yes": "1"})*

*d1["famsup"]=d1["famsup"].replace({"no": "0","yes": "1"})*

*d1["paid"]=d1["paid"].replace({"no": "0","yes": "1"})*

*d1["activities"]=d1["activities"].replace({"no": "0","yes": "1"})*

*d1["nursery"]=d1["nursery"].replace({"no": "0","yes": "1"})*

*d1["higher"]=d1["higher"].replace({"no": "0","yes": "1"})*

*d1["internet"]=d1["internet"].replace({"no": "0","yes": "1"})*

*d1["romantic"]=d1["romantic"].replace({"no": "0","yes": "1"})*

*###################*

*data = d1.drop(columns=['G3'])*

*#lay 80% train và 20% test*

*train =int(len(data) \* 0.8)*

*test = int(len(data) \* 0.2)*

*data\_train = data.loc[:train]*

*data\_test = data.loc[train+1:]*

*#print(data\_train)*

*#print(data\_test)*

*# gan label train va test*

*labels = d1['G3']*

*labels\_train = labels.loc[:train]*

*labels\_test = labels.loc[train+1:]*

*#print(labels\_train)*

*#print(labels\_test)*

*##################*

*## khoi tao mo hinh LinearRegression*

*model = LinearRegression()*

*#fit mo hinh*

*model.fit(data\_train, labels\_train)*

*#print(model.coef\_)*

*#print(model.intercept\_)*

*#Diem cua model*

*print(model.score(X=data\_train, y=labels\_train))*

*#Ham du doan ket qua*

*def myfuntion(x):*

*x = model.predict(x)*

*df = pd.DataFrame({'KetQua':x})*

*for i in range(df.size):*

*if(df['KetQua'][i]>=10):*

*df['KetQua'][i] = 'dau'*

*else:*

*df['KetQua'][i] = 'rot'*

*return df*

*# test*

*print(myfuntion(data\_test))*

*# voi diem lon hon bang 10 thi dau va nho hon 10 thi rot*

*print(labels\_test)*

**3.2 Kết quả dự đoán bài toán**

*runfile('D:/Hocmay/ML-20201210T144539Z-001/ML/student.py', wdir='D:/Hocmay/ML-20201210T144539Z-001/ML')*

*0.8664680892463477*

*KetQua*

*0 rot*

*1 dau*

*2 dau*

*3 dau*

*4 rot*

*.. ...*

*73 rot*

*74 dau*

*75 rot*

*76 dau*

*77 rot*

*[78 rows x 1 columns]*

*317 9*

*318 10*

*319 11*

*320 13*

*321 9*

*..*

*390 9*

*391 16*

*392 7*

*393 10*

*394 9*

*Name: G3, Length: 78, dtype: int64*

**CHƯƠNG IV: KẾT LUẬN**

Kết quả đạt được: chúng em đã cài đặt được thuật toán và sử dụng dụng thư viện scikit-learn trong quá trình học tập. Nhưng bên cạnh đó thuật toán vẫn còn những ưu nhược điểm như:

* Ưu điểm: Nhanh chóng để mô hình hóa và đặc biệt hữu ích khi mối quan hệ được mô hình hóa không quá phức tạp và nếu bạn không có nhiều dữ liệu. Hồi quy tuyến tính là đơn giản để hiểu, nó rất có giá trị cho các quyết định kinh doanh.
* Nhược điểm: Đối với dữ liệu phi tuyến tính, hồi quy đa thức có thể khá khó khăn để thiết kế, vì người ta phải có một số thông tin về cấu trúc của dữ liệu và mối quan hệ giữa các biến tính năng.

Do thời gian và kiến thức có hạn nên báo cáo chúng em vẫn còn nhiều sai sót, rất mong các thầy cô góp ý giúp chúng em hoàn thiện báo hơn nữa.