**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**ĐHQG - TPHCM**





**Khoa công nghệ thông tin**

**Môn: Phân tích thống kê dữ liệu nhiều biến**

**PRACTICE #03**

**Scikit-learn with PCA & LDA**

**Giáo viên hướng dẫn:** Nguyễn Mạnh Hùng

Lý Quốc Ngọc

Phạm Thanh Tùng

**Lớp:** 20TGMT01

**Sinh viên thực hiện:** Phùng Nghĩa Phúc – 20127284

*TP.Hồ Chí Minh, ngày 25 tháng 4 năm 2023*

**MỤC LỤC**

[MỤC LỤC 1](#_Toc133351536)

[GIỚI THIỆU 3](#_Toc133351537)

[BẢNG ĐÁNH GIÁ 4](#_Toc133351538)

[CHƯƠNG 1: Dataset Travel times 5](#_Toc133351539)

[1. Giới thiệu 5](#_Toc133351540)

[2. Mô tả dữ liệu 5](#_Toc133351541)

[3. Tóm tắt dữ liệu 5](#_Toc133351542)

[4. Mô tả tương quan dữ liệu 6](#_Toc133351543)

[CHƯƠNG 2: Trực quan hoá dữ liệu 7](#_Toc133351544)

[1. Biểu đồ dữ liệu đa biến 7](#_Toc133351545)

[2. Biểu đồ hiển thị quảng đường đi được và tốc độ tối đa của người lái xe trong tuần 7](#_Toc133351546)

[3. Biểu đồ hiển thị tốc độ tối đa và trung bình tốc độ di chuyển của người lái xe trong tuần 8](#_Toc133351547)

[4. Biểu đồ đường thể hiện thông số của tài xế 9](#_Toc133351548)

[**4.1 Biểu đồ hiển thị toàn bộ thống số của chuyển đi trong tuần** 9](#_Toc133351549)

[**4.2 Biểu đồ hiển thị Distance, MaxSpeed, AvgSpeed, AvgMovingSpeed** 9](#_Toc133351550)

[**4.3 Biểu đồ hiển thị Distance, MaxSpeed, AvgSpeed, AvgMovingSpeed** 10](#_Toc133351551)

[**4.4 Biểu đồ hiển thị TotalTime, MovingTime** 10](#_Toc133351552)

[5. Thống kê các dữ liệu đa biến tính được trên tập data. 10](#_Toc133351553)

[**5.1 Với tính trung bình các mẫu trên data.** 10](#_Toc133351554)

[**5.2 Với độ lệch chuẩn** 11](#_Toc133351555)

[**5.3 Với số liệu lớn nhất** 11](#_Toc133351556)

[**5.4 Với số liệu nhỏ nhất** 11](#_Toc133351557)

[6. Trung bình mẫu, độ lệch chuẩn và số lượng mẫu của mỗi nhóm data 11](#_Toc133351558)

[**6.1 Trung bình mẫu** 11](#_Toc133351559)

[**6.2. Độ lệch chuẩn** 12](#_Toc133351560)

[**6.3 Số lượng mẫu trong tuần** 12](#_Toc133351561)

[7. Within\_groups và between\_groups với một biến cụ thể: Biến Distance 12](#_Toc133351562)

[8. Sự khả tách của từng biến 12](#_Toc133351563)

[9. Within\_groups và between\_groups với 2 biến cụ thể: Biến AvgSpeed và AvgMovingSpeed 12](#_Toc133351564)

[10. Sự tương quan cho dữ liệu đa biến 13](#_Toc133351565)

[**10.1 Tính sự tương quan** 13](#_Toc133351566)

[**10.2 Ma trận tương quan** 13](#_Toc133351567)

[**10.3 Headmap ma trận tương quan** 13](#_Toc133351568)

[11. Biểu đồ Hinton để trực quan hoá ma trận trọng số. 13](#_Toc133351569)

[12. hệ số tương quan tuyến tính cho từng cặp biến theo thứ tự của hệ số tương quan. 14](#_Toc133351570)

[CHƯƠNG 3: PCA (principal component analysis) 15](#_Toc133351571)

[1. Tổng quan về PCA 15](#_Toc133351572)

[2. Các bước tính của PCA 15](#_Toc133351573)

[3. Thực thi 15](#_Toc133351574)

[**3.1 Tiêu chuẩn hoá các biến** 15](#_Toc133351575)

[**3.2 độ lệch chuẩn của từng thành phần chính** 15](#_Toc133351576)

[**3.3 Tổng phương sai** 16](#_Toc133351577)

[4. Biểu đồ với các thành phần chính của data 16](#_Toc133351578)

[5. Biểu đồ phân tán của các thành phần chính 16](#_Toc133351579)

[CHƯƠNG 4: LDA (Linear Discriminant Analysis) 18](#_Toc133351580)

[1. Tổng quan về LDA 18](#_Toc133351581)

[- Khác với PCA, LDA tìm phép chiếu sao cho tối đa hóa sự khác biệt giữa các lớp để có thể phân lớp hiệu quả. 2. Các bước tính LDA 18](#_Toc133351582)

[3. Thực thi với code 18](#_Toc133351583)

[**3.1 Làm đẹp data bằng tính hệ số phân biệt tuyến tính** 18](#_Toc133351584)

[**3.2 Chuẩn hoá dữ liệu theo nhóm** 19](#_Toc133351585)

[**3.3 Tính sự khả tách đạt được bởi từng hàm phân biệt** 19](#_Toc133351586)

[**3.4 Tính khoảng cách dưới dạng tỷ lệ của phương sai giữa các nhóm với phương sai bên trong các nhóm:** 20](#_Toc133351587)

[**3.5 Tính tỷ lệ dấu vết cho từng phân biệt tuyến tính** 20](#_Toc133351588)

[4. Biểu đồ xếp chống của các giá trị LDA với hàm phân biệt đầu tiên : LD1 20](#_Toc133351589)

[4. Biểu đồ xếp chống của các giá trị LDA với hàm phân biệt thứ 2 : LD2 21](#_Toc133351590)

[5. Biểu đồ phân tán của LDA 22](#_Toc133351591)

[CHƯƠNG 5: Bonus : Trực quan hoá dữ liệu với ICA và FA 24](#_Toc133351592)

[1. ICA (Independent Component Analysis) 24](#_Toc133351593)

[**1.1 Tương quan về ICA** 24](#_Toc133351594)

[**1.2 Biểu đồ** 24](#_Toc133351595)

[2. FA (Factor Analysis) 24](#_Toc133351596)

[**2.1 Tương quan về FA** 24](#_Toc133351597)

[**2.2 Biểu đồ** 25](#_Toc133351598)

[CHƯƠNG 6: So sánh PCA và LDA 26](#_Toc133351599)

[1. Biểu đồ của PCA và LDA 26](#_Toc133351600)

[2. So sánh 26](#_Toc133351601)

**GIỚI THIỆU**

1. **Sinh viên**

* Họ và tên: Phùng Nghĩa Phúc
* MSV: 20127284
* Lớp: 20TGMT01
* Môn học: Phân tích thống kê dữ liệu nhiều biến

1. **Chủ đề**

* Sử dụng Scikit-learn để visualization và áp dụng PCA và PDA để đưa ra đánh giá

**BẢNG ĐÁNH GIÁ**

|  |  |  |
| --- | --- | --- |
| STT | Loại | Đánh giá |
| 1 | **Mô tả dữ liệu** | 100% |
| 2 | **Trực quan hoá dữ liệu với các phép tính** | 100% |
| 3 | **PCA** | 100% |
| 4 | **LDA** | 100% |
| 5 | **Mở rộng PCA** | 100% |
| 6 | **Mở rộng LDA** | 100% |
| 7 | **Bonus: ICA và FA** | 100% |

**CHƯƠNG 1: Dataset Travel times**

- Link dataset: <https://openmv.net/info/travel-times>

**1. Giới thiệu**

- Một người lái xe sử dụng một ứng dụng để theo dõi tọa độ GPS khi anh ta lái xe đi làm và quay về mỗi ngày. Ứng dụng thu thập dữ liệu vị trí và độ cao. Dữ liệu cho khoảng 200 chuyến đi được tóm tắt trong bộ dữ liệu này.

**2. Mô tả dữ liệu**

- Dữ liệu có 205 hàng và 13 cột

- Trong dữ liệu chứa các thuộc tính:

- Date: ngày

- StartTime: Thời gian bắt đầu khi lên xe

- DayOfWeek: khi lên xe

- GoingTo: hướng di chuyển

- Distance: Quảng đường đi được

- MaxSpeed: tốc độ nhanh nhất được ghi nhận (tất cả các chuyến đi đều trên đường cao tốc 407 trong một số đoạn)

- AvgSpeed: tốc độ trung bình được ghi lại chỉ trong khi chiếc xe đang di chuyển

- AvgMovingSpeed: the average speed recorded only while the car is moving

- FuelEconomy: một ước tính sơ bộ về tiết kiệm nhiên liệu (nó không chính xác)

- TotalTime: thời lượng của toàn bộ chuyến đi, tính bằng phút

- MovingTime: khoảng thời gian khi ô tô được coi là đang di chuyển (nghĩa là không tính đến tắc nghẽn giao thông, tai nạn hoặc thời gian ô tô đứng yên)

- Take407All: là Yes nếu đường cao tốc thu phí 407 đã được sử dụng cho toàn bộ chuyến đi. Tôi cố gắng tránh đi 407, chọn các tuyến đường quay lại chậm hơn để tiết kiệm chi phí. Nhưng một số ngày tôi đến muộn, hoặc chỉ lười biếng và làm mọi cách.

- Comments

**3. Tóm tắt dữ liệu**

**Graphical user interface, text

Description automatically generated with medium confidence**

**4. Mô tả tương quan dữ liệu**

A picture containing text, window, black, scoreboard

Description automatically generated

**CHƯƠNG 2: Trực quan hoá dữ liệu**

**1. Biểu đồ dữ liệu đa biến**

Chart, line chart

Description automatically generated

**2. Biểu đồ hiển thị quảng đường đi được và tốc độ tối đa của người lái xe trong tuần**

**Chart, scatter chart

Description automatically generated**

- Sơ lược vào biểu đồ có thể thấy được tốc độ lớn nhất mà tái xe đi được nằm trong khoảng 125 -> 130. Tốc độ này được duy trì trên tổng quảng đường di chuyển được là 50 đến 52

**3. Biểu đồ hiển thị tốc độ tối đa và trung bình tốc độ di chuyển của người lái xe trong tuần**

**Chart, scatter chart

Description automatically generated**

**-** Dựa vào biểu đồ, thấy được tốc độ trung bình và tốc độ lớn nhất mà tài xế này di chuyển khá tương đồng nhau

**4. Biểu đồ đường thể hiện thông số của tài xế**

**4.1 Biểu đồ hiển thị toàn bộ thống số của chuyển đi trong tuần**

**Chart, line chart

Description automatically generated**

**4.2 Biểu đồ hiển thị Distance, MaxSpeed, AvgSpeed, AvgMovingSpeed**

**Chart, line chart

Description automatically generated**

**4.3 Biểu đồ hiển thị Distance, MaxSpeed, AvgSpeed, AvgMovingSpeed**

**Chart, line chart

Description automatically generated**

**4.4 Biểu đồ hiển thị TotalTime, MovingTime**

**Chart, line chart

Description automatically generated**

**5. Thống kê các dữ liệu đa biến tính được trên tập data.**

**5.1 Với tính trung bình các mẫu trên data.**

Distance 50.981512

MaxSpeed 127.591707

AvgSpeed 74.477561

AvgMovingSpeed 81.975610

TotalTime 41.904390

MovingTime 37.871707

**5.2 Với độ lệch chuẩn**

Distance 1.317979

MaxSpeed 4.118368

AvgSpeed 11.381953

AvgMovingSpeed 10.086852

TotalTime 6.832750

MovingTime 4.823265

**5.3 Với số liệu lớn nhất**

Distance 60.32

MaxSpeed 140.90

AvgSpeed 107.70

AvgMovingSpeed 112.10

TotalTime 82.30

MovingTime 62.40

**5.4 Với số liệu nhỏ nhất**

Distance 48.32

MaxSpeed 112.20

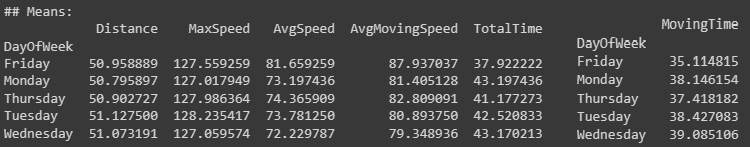
AvgSpeed 38.10

AvgMovingSpeed 50.30

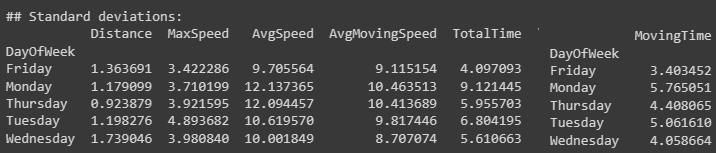
TotalTime 28.20

**6. Trung bình mẫu, độ lệch chuẩn và số lượng mẫu của mỗi nhóm data**

**6.1 Trung bình mẫu**



**6.2. Độ lệch chuẩn**



**6.3 Số lượng mẫu trong tuần**

Table

Description automatically generated with low confidence

**7. Within\_groups và between\_groups với một biến cụ thể: Biến Distance**

- Within\_groups: 

**-** between\_groups: 

**8. Sự khả tách của từng biến**

**A screenshot of a computer

Description automatically generated with medium confidence**

**9. Within\_groups và between\_groups với 2 biến cụ thể: Biến AvgSpeed và AvgMovingSpeed**

- Within\_groups: 

**-** between\_groups: 

**10. Sự tương quan cho dữ liệu đa biến**

**10.1 Tính sự tương quan**

- Giá trị p-value: 0.00023974233393727256

- Giá trị cor: 0.2538685434138951

**10.2 Ma trận tương quan**

**Text

Description automatically generated**

**10.3 Headmap ma trận tương quan**

Square

Description automatically generated

**11. Biểu đồ Hinton để trực quan hoá ma trận trọng số.**

Chart

Description automatically generated

**12. hệ số tương quan tuyến tính cho từng cặp biến theo thứ tự của hệ số tương quan.**

**Text

Description automatically generated**

**CHƯƠNG 3: PCA (principal component analysis)**

**1. Tổng quan về PCA**

- PCA là kĩ thuật giảm chiều dữ liệu từ n chiều sang dữ liệu m chiều (m<n) mà vẫn giữ được nhiều thông tin nhất có thể.

**2. Các bước tính của PCA**

Bước 1: Tính giá trị trung bình:

Bước 2: Chuẩn hóa:

Bước 3: Tính ma trận hiệp phương sai:

Bước 4: Tính các trị riêng λi và vector riêng vi:

Bước 5: Chọn K vector riêng ứng với K trị riêng lớn nhất để xây dựng ma trận UK có

các cột tạo thành một hệ trực giao. K vector này, còn được gọi là các thành phần chính, tạo thành một không gian con gần với phân bố của dữ liệu ban đầu đã chuẩn hoá.

Bước 6: Chiếu dữ liệu ban đầu đã chuẩn hoá X xuống không gian con tìm được. Dữ liệu mới chính là toạ độ của các điểm dữ liệu trên không gian mới:

**3. Thực thi**

**3.1 Tiêu chuẩn hoá các biến**

Text

Description automatically generated

**3.2 độ lệch chuẩn của từng thành phần chính**

**Text

Description automatically generated  
3.3 Tổng phương sai**

****

**4. Biểu đồ với các thành phần chính của data**

**Chart, line chart

Description automatically generated**

**5. Biểu đồ phân tán của các thành phần chính**

**Chart, scatter chart

Description automatically generated**

**CHƯƠNG 4: LDA (Linear Discriminant Analysis)**

**1. Tổng quan về LDA**

- LDA là một thuật toán học có giám sát, giảm chiều dữ liệu.

- Mục đích LDA là tìm sự khác nhau giữa các thành phần trong 1 class (within-class) là nhỏ và sự khác nhau giữa các classes là lớn.

- Khác với PCA, LDA tìm phép chiếu sao cho tối đa hóa sự khác biệt giữa các lớp để có thể phân lớp hiệu quả.**2. Các bước tính LDA**

Bước 1: Tính ma trận phân tán giữa các nhóm :

A picture containing text, watch

Description automatically generated

µi là giá trị trung bình của từng lớp.

µ là giá trị trung bình của tất cả dữ liệu.Bước 2: Tính ma trận phân tán tích lũy ứng với từng nhóm

A picture containing text, watch

Description automatically generated

Bước 3 : Xây dựng hàm tiêu chí tách lớp



Bước 4 : Dự đoán nhãn của mẫu dữ liệu nhập (so sánh vector trung bình của từng nhóm – gần vector trung bình nhất).

**3. Thực thi với code**

**3.1 Làm đẹp data bằng tính hệ số phân biệt tuyến tính**

****

**3.2 Chuẩn hoá dữ liệu theo nhóm**

**A picture containing text, black, scoreboard, screenshot

Description automatically generated**

**3.3 Tính sự khả tách đạt được bởi từng hàm phân biệt**

**Graphical user interface

Description automatically generated**

**3.4 Tính khoảng cách dưới dạng tỷ lệ của phương sai giữa các nhóm với phương sai bên trong các nhóm:**

****

**3.5 Tính tỷ lệ dấu vết cho từng phân biệt tuyến tính**

**Text

Description automatically generated with medium confidence**

**4. Biểu đồ xếp chống của các giá trị LDA với hàm phân biệt đầu tiên : LD1**

**Chart, histogram

Description automatically generated**

**4. Biểu đồ xếp chống của các giá trị LDA với hàm phân biệt thứ 2 : LD2**

**Chart, histogram

Description automatically generated**

**5. Biểu đồ phân tán của LDA**

**Chart, scatter chart

Description automatically generated**

**CHƯƠNG 5: Bonus : Trực quan hoá dữ liệu với ICA và FA**

**1. ICA (Independent Component Analysis)**

**1.1 Tương quan về ICA**

**- Independent Component Analysis (phân tích thành phần độc lập)** là một phương pháp thống kê được xây dựng để tách rời tín hiệu nhiều chiều thành các thành phần tín hiệu độc lập ẩn sâu bên dưới dữ liệu. Kỹ thuật này đòi hỏi phải đặt ra giả thuyết tồn tại các nguồn tín hiệu bên dưới nongaussianity và độc lập thống kê từng đôi một. Thuật toán ICA có nhiều ứng dụng rộng rãi trong nhiều bài toán khác nhau như xử lý tín hiệu, kinh tế học, sinh tin học,…

**1.2 Biểu đồ**

Chart, scatter chart

Description automatically generated

**2. FA (Factor Analysis)**

**2.1 Tương quan về FA**

- Phân tích nhân tố (FA) là một phương pháp thống kê được sử dụng để xác định các mẫu giữa các biến quan sát và giải thích cấu trúc cơ bản của các biến đó. Nó nhằm mục đích giảm số lượng các biến bằng cách nhóm chúng thành các yếu tố cơ bản.

- FA giả định rằng các biến quan sát bị ảnh hưởng bởi một hoặc nhiều yếu tố cơ bản và các yếu tố này độc lập với nhau.

- FA được sử dụng cho mục đích khám phá hoặc xác nhận và thường được sử dụng trong các lĩnh vực như tâm lý học, xã hội học và nghiên cứu tiếp thị để xác định các cấu trúc tiềm ẩn giải thích dữ liệu quan sát được

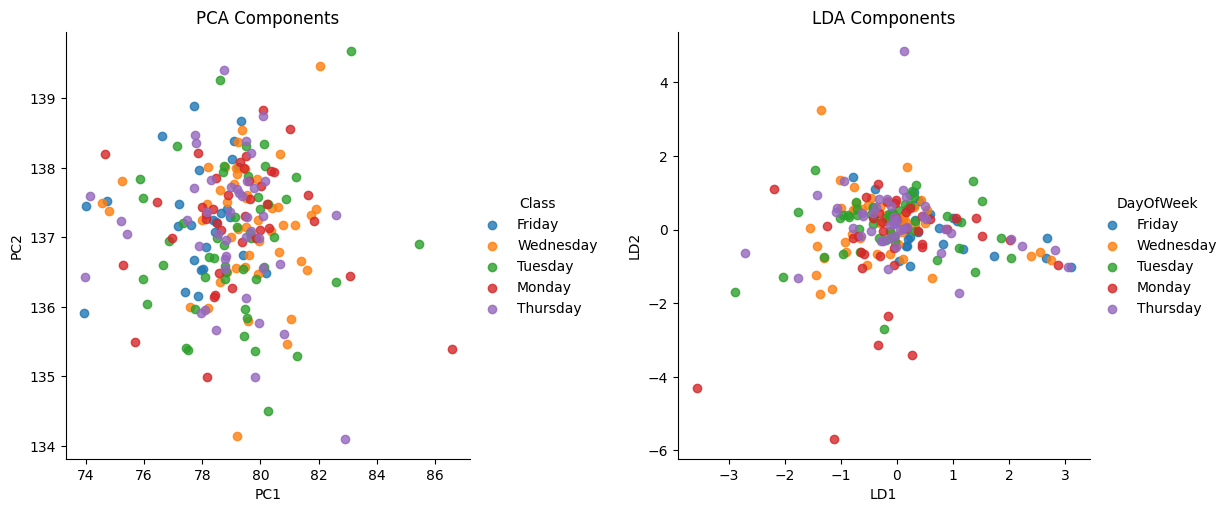
**2.2 Biểu đồ**

**Chart, scatter chart

Description automatically generated**

**CHƯƠNG 6: So sánh PCA và LDA**

**1. Biểu đồ của PCA và LDA**

****

**2. So sánh**

**-** Dựa vào biểu đồ có thể thấy được, dữ liệu ở PCA cho ra nhiều hơn so với LDA nhưng ở PCA rải rác khắp nơi khó nhận định được đâu là thông tin cần thiết để triển khai so với LDA. LDA thì hiển thị những thông tin cần thiết nhất để có thể trực quan được về dữ liệu hơn.

- Lí do vì sao LDA làm được điều đó là vì LDA còn xét trên phương sai còn PCA thì không. LDA được giám sát trong khi PCA không được giám sát và PCA bỏ qua các nhãn lớp.