Giảm chiều dữ liệu sử dụng PCA

Le Nhat Tung

Contents

1 Giới thiệu về PCA

1.1 Khái niệm

Principal Component Analysis (PCA) là một kỹ thuật giảm chiều được sử dụng rộng rãi trong phân tích dữ liệu và học máy. PCA chuyển đổi một tập dữ liệu có nhiều biến (nhiều chiều) thành một tập dữ liệu với ít biến hơn (ít chiều hơn) nhưng vẫn giữ được thông tin quan trọng nhất.

1.2 Mục tiêu của PCA

PCA có các mục tiêu chính sau:

- **Giảm số lượng biến:** Chuyển đổi dữ liệu sang không gian mới với ít chiều hơn
- **Giữ lại thông tin quan trọng:** Các thành phần chính (principal components) mới sẽ giữ lại phần lớn sự biến thiên của dữ liệu gốc
- **Loại bỏ đa cộng tuyến:** Các thành phần chính không tương quan với nhau
- **Trực quan hóa dữ liệu:** Có thể hiển thị dữ liệu nhiều chiều trên không gian 2D hoặc 3D

1.3 Quy trình thực hiện PCA

1. Chuẩn hóa dữ liệu (nếu cần)
2. Tính ma trận hiệp phương sai (covariance matrix) hoặc ma trận tương quan (correlation matrix)
3. Tính các giá trị riêng (eigenvalues) và vector riêng (eigenvectors) của ma trận
4. Sắp xếp các vector riêng theo thứ tự giảm dần của giá trị riêng tương ứng
5. Chọn k vector riêng đầu tiên để tạo ma trận chiếu
6. Chiếu dữ liệu gốc lên không gian mới k chiều

2 Ứng dụng PCA với bộ dữ liệu mtcars

2.1 Tải thư viện cần thiết

```
library(tidyverse)   # Bộ thư viện chứa nhiều công cụ xử lý dữ liệu như dplyr, tidyr, ggplot2,... để th
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(ggplot2)     # Thư viện tạo đồ thị với cú pháp ngữ pháp đồ họa (grammar of graphics) giúp tạo b
#install.packages(GGally)
library(GGally)      # Mở rộng của ggplot2, cung cấp các hàm để tạo ma trận tương quan, biểu đồ cặp (pa
```

```
## Warning: package 'GGally' was built under R version 4.4.3
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(factoextra)  # Thư viện cho phân tích đa chiều, hỗ trợ phân tích thành phần chính (PCA) và phân
```

```
## Warning: package 'factoextra' was built under R version 4.4.3
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

2.2 Hiểu về bộ dữ liệu mtcars

```
# Xem cấu trúc bộ dữ liệu mtcars
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
# Hiển thị một số dòng đầu tiên
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
```

```
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
# Tóm tắt thống kê
summary(mtcars)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

Bộ dữ liệu mtcars chứa thông tin về 32 mẫu xe với 11 biến mô tả các đặc điểm kỹ thuật:

- mpg: Miles per gallon (hiệu suất tiêu thụ nhiên liệu)
- cyl: Số xi-lanh
- disp: Dung tích xi-lanh
- hp: Mã lực
- drat: Tỷ số truyền động sau
- wt: Trọng lượng (1000 lbs)
- qsec: Thời gian chạy 1/4 dặm
- vs: Kiểu động cơ (0 = chữ V, 1 = thẳng hàng)
- am: Kiểu hộp số (0 = tự động, 1 = số sàn)
- gear: Số lượng số
- carb: Số lượng bộ chế hòa khí