Báo Cáo Học Máy Nâng Cao - báo cáo

Phương pháp nghiên cứu khoa học (Đại học Điện lực)

TRƯỜNG ĐẠI HỌC ĐIỆN LỰC

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO CHUYÊN ĐỀ HỌC PHẦN
HỌC MÁY NÂNG CAO

ĐỀ TÀI: ÁP DỤNG PHƯƠNG PHÁP GIẢM CHIỀU PCA DỰ ĐOÁN MOBILE APPSTORE

Sinh viên thực hiện	**: DƯƠNG TUẤN ĐẠT**
	: TRẦN SƠN TÙNG
	: ĐẶNG QUYẾT TIẾN
Giảng viên hướng dẫn	**: ĐÀO NAM ANH**
Ngành	**: CÔNG NGHỆ THÔNG TIN**
Chuyên ngành	**: CÔNG NGHỆ PHẦN MỀM**
Lớp	**: D14CNPM4**
Khóa	**: 2019 - 2024**

Hà Nội, tháng 1 năm 2023

PHIẾU CHẤM ĐIỂM

Sinh viên thực hiện:

Họ và tên sinh viên	Nội dung thực hiện	Chữ ký	Điểm
Dương Tuấn Đạt 19810310101			
Trần Sơn Tùng 19810310127			
Đặng Quyết Tiến 19810310111			

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

MỤC LỤC

CHƯƠNG 1: GIỚI THIỆU VỀ PHƯƠNG PHÁP PHÂN TÍCH THÀNH PHẦN CHÍNH (PCA)....1
1.1. Thuật toán PCA (Principal Component Analysis)....1
1.2. Giảm chiều dữ liệu....2
1.3. Các bước thực hiện thuật toán giảm chiều PCA....3
1.4. Tiêu chí giảm chiều PCA....4
1.5. Ưu, nhược điểm của thuật toán PCA....4
1.5.1. Ưu điểm của thuật toán PCA....4
1.5.2. Nhược điểm của thuật toán PCA....4
1.6. Ứng dụng thuật toán PCA....4
CHƯƠNG 2: CƠ SỞ TOÁN HỌC SỬ DỤNG TRONG PRINCIPAL COMPONENT ANALYSIS – PCA....6
2.1. Độ lệch chuẩn (Standard Deviation)....6
2.2. Kỳ vọng và ma trận hiệp phương sai....6
2.2.1. Dữ liệu một chiều....6
2.2.2. Dữ liệu nhiều chiều....7
3.1. Mô tả bài toán....8
3.1.1. Mô tả bài toán trực quan hóa PCA trong bộ dữ liệu Digits....8
3.2. Môi trường thực nghiệm....8
3.3. Xây dựng bộ dữ liệu....9
3.3.1. Bộ dữ liệu cho bài toán dự đoán giá BĐS trên 1 đơn vị diện tích....9
3.4.1.Kết quả thực nghiệm....10
KẾT LUẬN....15
TÀI LIỆU THAM KHẢO....16

LỜI CẢM ƠN

Nhóm chúng em xin chân thành cảm ơn các thầy, cô giáo trong Khoa Công nghệ thông tin, trường Đại học Điện Lực, đã tạo điều kiện cho em thực hiện đề tài này.

Để có thể hoàn thành báo cáo đề tài **"**Áp dụng phương pháp giảm chiều PCA dự đoán Mobile AppStore**",** nhóm em xin gửi lời cảm ơn chân thành nhất tới thầy Đào Nam Anh đã truyền đạt, giảng dạy cho chúng em những kiến thức, những kinh nghiệm quý báu trong thời gian học tập và rèn luyện, tận tình hướng dẫn chúng em trong quá trình làm báo cáo này.

Nhóm em cũng gửi lời cảm ơn tới bạn bè đã đóng góp những ý kiến quý bàu để nhóm em có thể hoàn thành báo cáo tốt hơn. Tuy nhiên, do thời gian và trình độ có hạn nên báo cáo này chắc chắn không tránh khỏi những thiếu sót, nhóm em rất mong được sự đóng góp ý kiến của các thầy và toàn thể các bạn.

Một lần nữa, em xin chân thành cảm ơn và luôn mong nhận được sự đóng góp của tất cả mọi người.

LỜI MỞ ĐẦU

Lý do chọn đề tài

Ngày nay, với sự phát triển mạnh mẽ của Công nghệ thông tin, các mô hình tự động hóa ngày càng được ứng dụng trong thực tế nhiều hơn. Song song với nó, khai thác dữ liệu để phục vụ trong công cuộc Cách mạng 4.0 là không thể thiếu. Dữ liệu trong thực tế thì vô cùng đa dạng. Muốn sử dụng dữ liệu một cách thông minh và có ích nhất, chúng ta cần quan tâm tới các đặc tính (feature) của dữ liệu. Chúng ta có thể quan sát được trong không gian 2 chiều, 3 chiều, nhưng dữ liệu thì lại có rất nhiều chiều. Làm sao để có thể trực quan hóa dữ liệu lên không gian 2 chiều, 3 chiều? Để trả lời câu này, chúng em xin chọn đề tài: **"**Áp dụng phương pháp giảm chiều PCA dự đoán Mobile AppStore**"** để làm rõ.

Trong khuôn khổ bài tập lớn của nhóm, chúng em xin được trình bày giảm chiều dữ liệu bằng phương pháp phân tích thành phần chính (PCA) ứng dụng trong bộ dữ liệu Digits và dự đoán Mobile AppStore.

Cấu trúc báo cáo bao gồm các chương như sau:

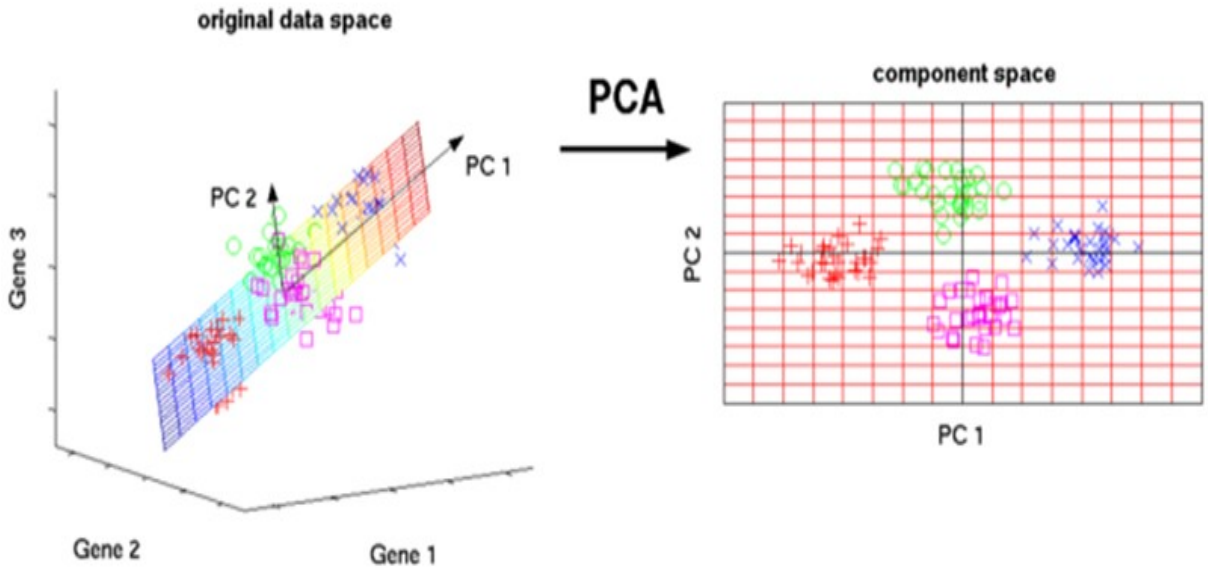
Chương 1: Giới thiệu về Phương pháp phân tích thành phần chính (PCA)

Chương 2: Cơ sở toán học trong PCA

Chương 3: Ứng dụng thuật toán PCA trong bộ dữ liệu Digits và dự đoán Mobile AppStore.

CHƯƠNG 1: GIỚI THIỆU VỀ PHƯƠNG PHÁP PHÂN TÍCH THÀNH PHẦN CHÍNH (PCA)

1.1. Thuật toán PCA (Principal Component Analysis)



Hình 1.1: Hình ảnh đại diện cho phương pháp giảm chiều PCA

Thuật toán phân tích thành phần chính (Principal Components Analysis - PCA) là một thuật toán thống kê sử dụng phép biến đổi trực quan để biến đổi một tập hợp dữ liệu từ một không gian nhiều chiều sang một không gian mới ít chiều hơn (2 hoặc 3 chiều) nhằm tối ưu hóa việc thể hiện sự biến thiên của dữ liệu.

Ý tưởng chính của PCA là ánh xạ các đặc trưng **n** chiều thành **k** chiều. **k** chiều này là một đối tượng trực giao hoàn toán mới, còn được gọi là thành phần chính, là đối tượng k chiều được tái tạo lại trên cơ sở đối tượng n chiều ban đầu.

Công việc của PCA là tìm một cách tuần tự một tập các trục tọa độ mới có liên quan mật thiết đến bản thân dữ liệu. Trong số đó, lựa chọn trục tọa độ mới thứ hai là mặt phẳng trực giao với trục toạn độ đầu tiên để tối đa hóa phương sai và trục thứ ba giống với trục thứ nhất. Bằng phép loại suy, có thể thu được n trục tọa độ như vậy. Với trục tọa độ mới thu được theo cách này, chúng ta thấy rằng hầu hết phương sai được chưa trong k trục tọa độ đầu tiên và phương sai chứa

trong trục tọa độ sai gần như bằng 0. Do đó, chúng ta có thể bỏ qua các trục còn lại và chỉ giữ lại k trục đầu tiên chứa hầu hết các phương sai. Trên thực tế, điều này tương đương với việc chỉ giữ lại các đặc trưng chứa hầu hết phương sai và bỏ qua các kích thược đặc trưng chứa phương sai gần như bằng 0, để đạt được quá trình giảm kích thước cho các đối tượng dữ liệu.

Nói một cách ngắn gọn: Sử dụng ít chỉ số toàn diện hơn để đại diện cho nhiều loại thông tin khác nhau trong mỗi biến, phân tích thành phần chính và phân tích nhân tố thuộc loại thuật toán giảm chiều này.

1.2. Giảm chiều dữ liệu

Giảm chiều dữ liệu là sự biến đổi dữ liệu từ không gian nhiều chiều thành không gian ít chiều để biểu diễn ở dạng chiều thấp đồng thời giữ lại một số thuộc tính có ý nghĩa của dữ liệu gốc, có ý tưởng là gần với chiều nội tại.

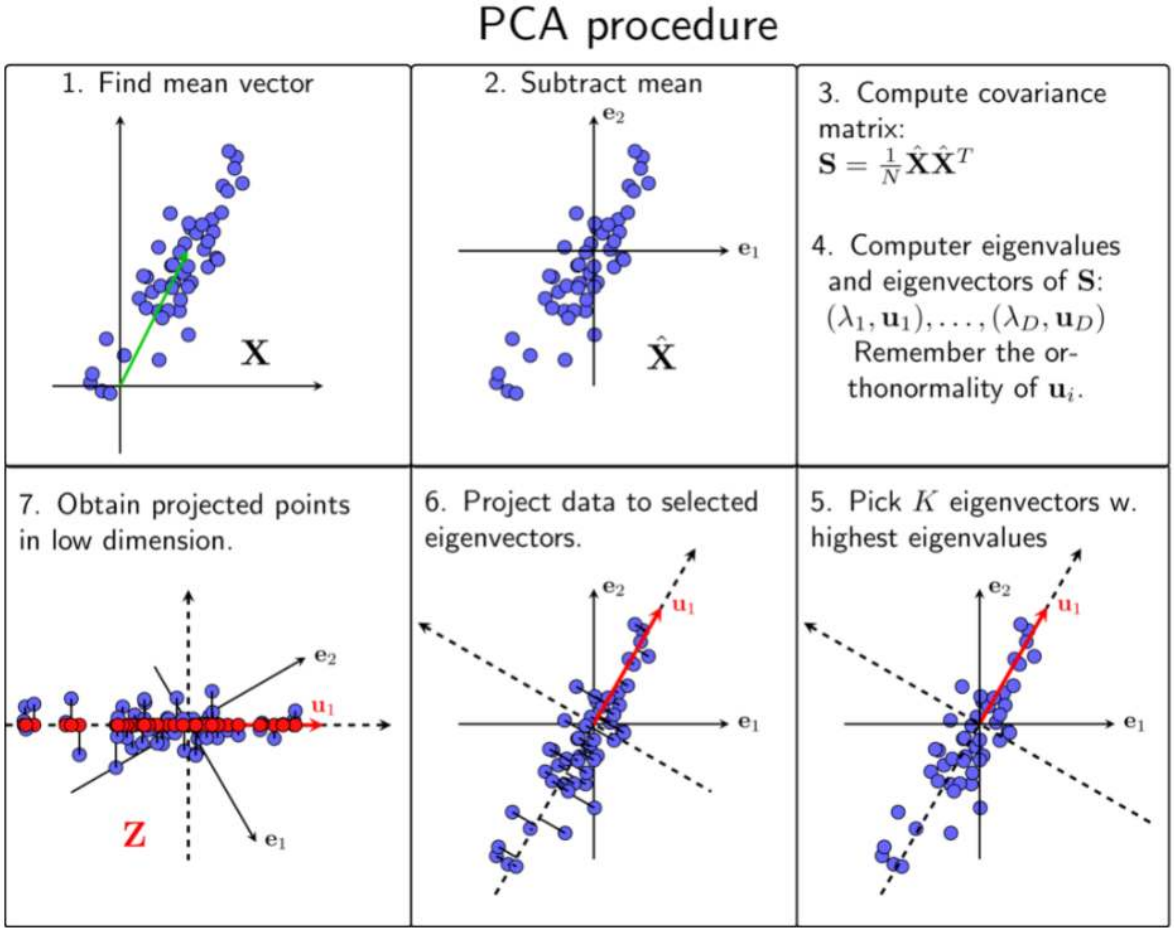
Phân tích dữ liệu trong không gian nhiều chiều có thể khó khăn vì nhiều lý do, dữ liệu thô có tính thưa thớt là một hậu quả của lời nguyền chiều và do đó việc phân tích trở lên khó tính toán, hơn nữa thuật toán có thể mất rất nhiều thời gian để xử lý dữ liệu. Giảm chiều dữ liệu là phổ biến trong các lĩnh vực có số lượng quan sát lớn hoặc số lượng biến lớn chẳng hạn như nhận dạng tiếng nói, tin học thần kinh và tin sinh học.

Tóm lại, giảm chiều là một phương pháp xử lý trước dữ liệu tính năng nhiều chiều. Giảm chiều là giữ lại các tính năng quan trọng nhất của dữ liệu, loại bỏ nhiễu và các tính năng không quan trọng, để đạt được mục đích cải thiện tốc độ xử lỹ dữ liệu.

Trong thực tế, sản xuất và ứng dụng, việc giảm chiều trong một phạm vi tổn thất thông tin nhất định có thể giúp chúng ta tiết kiệm rất nhiều thời gian và chi phí. Giảm chiều cũng đã trở thành một phương pháp tiền xử lý dữ liệu được sử dụng rất rộng rãi.

1.3. Các bước thực hiện thuật toán giảm chiều PCA

- Bước 1: Tính vector kỳ vọng của toàn bộ dữ liệu
- Bước 2: Trừ mỗi điểm dữ liệu đi vector kỳ vọng của toàn bộ dữ liệu
- Bước 3: Tính ma trận hiệp phương sai
- Bước 4: Tính các trị riêng và vector riêng của norm bằng một ma trận này, sắp xếp chúng theo thứ tự giảm dần của trị riêng.
- Bước 5: Chọn K vector riêng ứng với K trị riêng lớn nhất để xây dựng ma trân U_k 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 hóa.
- Bước 6: Chiếu dữ liệu ban đầu đã chuẩn hóa xuống không gian con tìm được.
- Bước 7: Dữ liệu mới chính là tọa độ của các điểm dữ liệu trên không gian mới.



Hình 1.1: Các bước thực hiện PCA

1.4. Tiêu chí giảm chiều PCA

- Tái tạo gần nhất: Đối với tất cả các điểm trong tập mẫu, tổng sai số giữa điểm được tái tạo và điểm ban đầu là nhỏ nhất.
- Khả năng phân tách tối đa: Hình chiếu của mẫu trong không gian chiều thấp càng tách biệt càng tốt.

1.5. Ưu, nhược điểm của thuật toán PCA

1.5.1. Ưu điểm của thuật toán PCA

- Loại bỏ các đặc trưng tương quan (giảm các đặc trưng)
- Làm cho tập dữ liệu dễ sử dụng hơn.
- Cải thiện hiệu suất thuật toán.
- Giảm quá khớp (overfitting).
- Cải thiện trực quan hóa dữ liệu (dễ trực quan hóa khi có ít chiều)

1.5.2. Nhược điểm của thuật toán PCA

- Nếu người sử dụng đã có kiến thức nhất định về đối tượng quan sát và nắm vững một số đặc điểm của dữ liệu nhưng không thể can thiệp vào quá trình xử lý thông qua tham số hóa và các phương pháp khác thì có thể không đạt được hiệu quả mong đợi và hiệu quả không cao;
- Phân rã Eigenvalue có một số hạn chế, ví dụ, ma trận được biến đổi phải là ma trận vuông;
- Trong trường hợp phân bố không theo Gaussian, các thành phần chính thu được bằng phương pháp PCA có thể không tối ưu.
- Các biến độc lập trở nên khó hiểu hơn.
- Chuẩn hóa dữ liệu trước khi sử dụng PCA.
- Mất thông tin.

1.6. Ứng dụng thuật toán PCA

- Khám phá và trực quan hóa các tập dữ liệu nhiều chiều.
- Nén dữ liệu.
- Tiền xử lý dữ liệu.
- Phân tích và xử lý hình ảnh, giọng nói và giao tiếp.

- Giảm kích thước (quan trọng nhất), loại bỏ dư thừa dữ liệu và nhiễu.
- PCA trong nhận dạng ảnh nư nhận dạng khuôn mặt, …
- ứng dụng PCA trong phân tích mô tả định lượng
- Nếu ta có thể giảm chiều về 2 hoặc 3 chiều ta có thể dùng các loại đồ thị để hiểu thêm về dữ liệu mà ta đang có giúp dễ trực quan hơn.
- Xử lý vấn đề tương quan giữa các biến trong dữ liệu ban đầu bằng cách sử dụng biến mới trong không gian mà phương pháp PCA tìm được để mô tả dữ liệu.

CHƯƠNG 2: CƠ SỞ TOÁN HỌC SỬ DỤNG TRONG PRINCIPAL COMPONENT ANALYSIS – PCA

2.1. Độ lệch chuẩn (Standard Deviation)

- **Ý nghĩa:** đó tính biến động của giá trị mang tính thống kê. Nó cho thấy sự chênh lệch về giá trị của từng thời điểm đánh giá so với giá trị trung bình.
- **Biểu diễn toán học:**

$$\sigma = s = E\{X(t) - m_x(t)\}$$

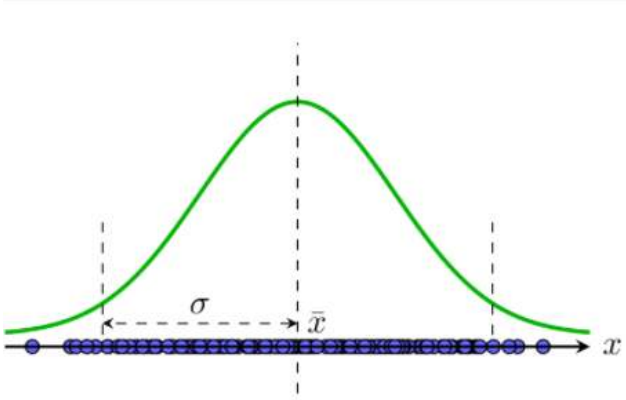
2.2. Kỳ vọng và ma trận hiệp phương sai

2.2.1. Dữ liệu một chiều

- Cho N giá trị từ x_1 đến x_N thì kỳ vọng và phương sai của bộ dữ liệu này được định nghĩa là:

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{1}{N}\mathbf{X1}\,\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2$$

- Với 1 thuộc R^N là vector cột chứa toàn bộ phần từ 1, Kỳ vọng đơn giản là trung bình công của toàn bộ các giá trị.
- Phương sai là trung bình cộng của bình phương khoảng cách từ mỗi điểm tới kỳ vọng, phương sai càng nhỏ thi các điểm dữ liệu càng gần với kỳ vọng, tức là các điểm dữ liệu càng giống nhau, phương sai càng lớn thì ta nói dữ liệu càng có tính phân tán.



Hình 2.1: Ví dụ về kỳ vọng và phương sai trong không gian một chiều

2.2.2. Dữ liệu nhiều chiều

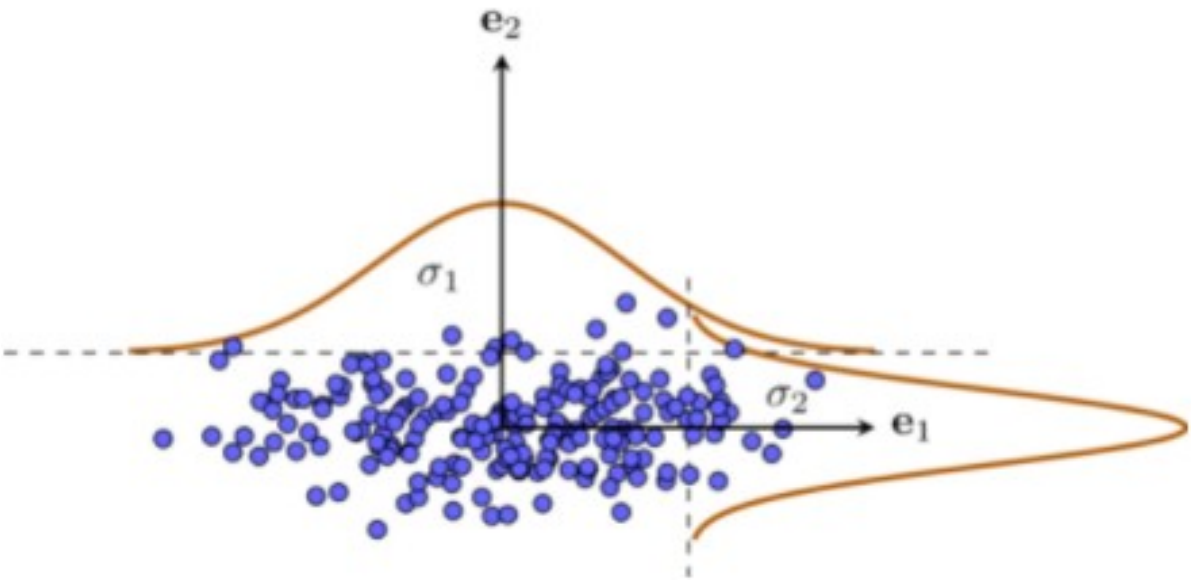
- Cho N điểm dữ liệu được biểu diễn bởi các vector cột x_1 đến x_N khi đó vector kỳ vọng và ma trận hiệp phương sai của toàn bộ dữ liệu được định nghĩa là:

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \quad \mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T = \frac{1}{N}\hat{\mathbf{X}}\hat{\mathbf{X}}^T$$

Trong đó $\hat{\mathbf{X}}$ được tạo bằng cách trừ mỗi cột của $\mathbf{X}$ đi $\bar{\mathbf{x}}$:

$$\hat{\mathbf{x}}_n = \mathbf{x}_n - \bar{\mathbf{x}}$$

- Các công thức này khá tương đồng với với các công thức của dữ liệu một chiều, cho nên có một vài lưu ý như sau:
  - Ma trận hiệp phương sai là ma trận đối xứng hơn thế nữa nó là một ma trận nửa xác định dương.
  - Mọi phần tử trên đường chéo của ma trận hiệp phương sai là các số không âm, chúng cũng chính là phương sai của từng chiều của dữ liệu.
  - Nếu ma trận hiệp phương sai là ma trận đường chéo, ta có dữ liệu hoàn toàn không tương quan giữa các chiều.



Hình 2.2: Dữ liệu trên không gian hai chiều không tương quan

CHƯƠNG 3: ỨNG DỤNG TRỰC QUAN HÓA PCA DỰ ĐOÁN DỰ ĐOÁN MOBILE APPSTORE

3.1. Mô tả bài toán

3.1.1. Mô tả bài toán trực quan hóa PCA trong bộ dữ liệu Digits

Hiện nay, ở hầu hết các quốc gia đều phân chia tài sản quốc gia thành 2 loại: bất động sản và động sản, nhưng còn có sự khác nhau trong khái niệm cụ thể về bất động sản. Tuy nhiên, có một điểm tương đối thống nhất trong khái niệm bất động sản là những tài sản gắn liền với đất đai và không di dời được. Theo qui định tại Điều 181 của Bộ Luật Dân sự nước Cộng hòa xã hội chủ nghĩa Việt Nam năm 2005, bất động sản là các tài sản không di dời được bao gồm:

- Đất đai.
- Nhà ở, công trình xây dựng gắn liền với đất đai, kể cả các tài sản gắn liền với nhà ở, công trình xây dựng đó.
- Các tài sản khác gắn liền với đất đai.
- Các tài sản khác do pháp luật qui định.

Bài toán Dự đoán giá BĐS trên 1 đơn vị diện tích được thực hiện khi có đầy đủ các thông tin liên quan. Sau đó các chuyên gia sẽ dự đoán giá BĐS dựa trên các thông tin đã có

- Input: Thông tin, vị trí
- Ouput: Giá BĐS trên một đơn vị diện tích.

3.2. Môi trường thực nghiệm

Hình 3.1: Ngôn ngữ python

Python là ngôn ngữ lập trình được sử dụng rất phổ biến ngày nay để phát triển nhiều loại ứng dụng phần mềm khác nhau như các chương trình chạy trên desktop, server, lập trình các ứng dụng web... Ngoài ra Python cũng là ngôn ngữ ưa thích trong ngành khoa học về dữ liệu (data science) cũng như là ngôn ngữ phổ biến để xây dựng các chương trình trí tuệ nhân tạo trong đó bao gồm machine learning.

Python là ngôn ngữ dễ học: Ngôn ngữ Python có cú pháp đơn giản, rõ ràng, sử dụng một số lượng không nhiều các từ khoá, do đó Python được đánh giá là một ngôn ngữ lập trình thân thiện với người mới học.

Python là ngôn ngữ dễ hiểu: Mã lệnh (source code hay đơn giản là code) viết bằng ngôn ngữ Python dễ đọc và dễ hiểu. Ngay cả trường hợp bạn chưa biết gì về Python bạn cũng có thể suy đoán được ý nghĩa của từng dòng lệnh trong source code.

Python có tương thích cao (highly portable): Chương trình phần mềm viết bằng ngôn ngữ Python có thể được chạy trên nhiều nền tảng hệ điều hành khác nhau bao gồm Windows, Mac OSX và Linux.

3.3. Xây dựng bộ dữ liệu

3.3.1. Bộ dữ liệu cho bài toán dự đoán giá BĐS trên 1 đơn vị diện tích

- Tập dữ liệu gồm thông tin của 414 BĐS với các thông tin khác nhau.
- Đặt Y là giá BĐS trên 1 đơn vị diện tích.
- Bộ dữ liệu gồm 13 thuộc tính
    - App
    - Category
    - Rating
    - Reviews
    - Size
    - Install
    - Type

- Price
- Content Rating
- Genres
- Last Update
- Current Ver
- Android Ver

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up
5	Paper flowers instructions	ART_AND_DESIGN	4.4	167	5.6M	50,000+	Free	0	Everyone	Art & Design	March 26, 2017	1.0	2.3 and up
6	Smoke Effect Photo Maker - Smoke Editor	ART_AND_DESIGN	3.8	178	19M	50,000+	Free	0	Everyone	Art & Design	April 26, 2018	1.1	4.0.3 and up
7	Infinite Painter	ART_AND_DESIGN	4.1	36815	29M	1,000,000+	Free	0	Everyone	Art & Design	June 14, 2018	6.1.61.1	4.2 and up
8	Garden Coloring Book	ART_AND_DESIGN	4.4	13791	33M	1,000,000+	Free	0	Everyone	Art & Design	September 20, 2017	2.9.2	3.0 and up
9	Kids Paint Free - Drawing Fun	ART_AND_DESIGN	4.7	121	3.1M	10,000+	Free	0	Everyone	Art & Design;Creativity	July 3, 2018	2.8	4.0.3 and up

Hình 3.2: Bộ dữ liệu dự đoán Mobile AppStore

3.4.1 Kết quả thực nghiệm

- In 5 mẫu đầu tiên của tập dữ liệu và số lượng nhãn của các lớp

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

- Chuyển đổi dữ liệu Size thành dữ liệu số:

```
0        19456.0
1        14336.0
2         8908.8
3        25600.0
4         2867.2
          ...
10836    54272.0
10837     3686.4
10838     9728.0
10839        NaN
10840    19456.0
Name: Size, Length: 10841, dtype: float64
```

- Chuyển đổi dữ liệu Install thành dữ liệu số:

```
0           10000.0
1          500000.0
2         5000000.0
3        50000000.0
4          100000.0
              ...
10836        5000.0
10837         100.0
10838        1000.0
10839        1000.0
10840    10000000.0
Name: Installs, Length: 10840, dtype: float64
```

- Chuyển đổi dữ liệu Review thành dữ liệu số:

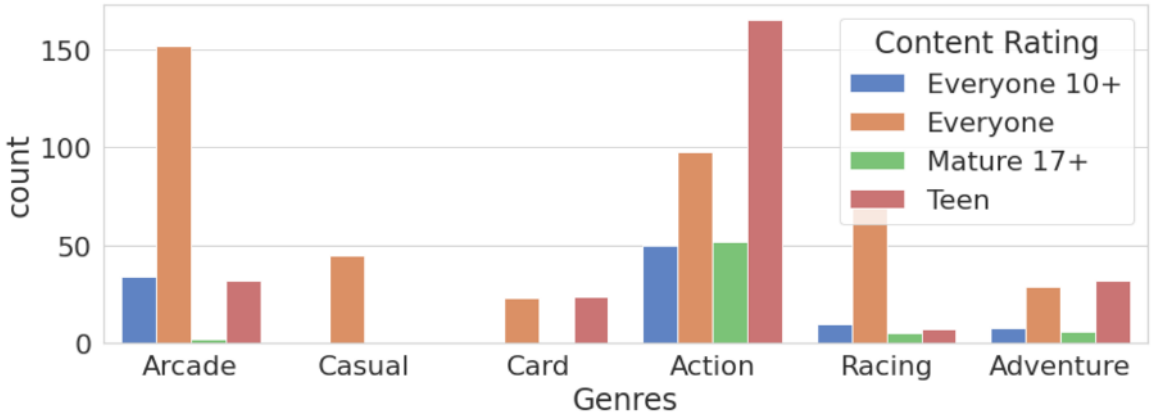
```
0           159
1           967
2         87510
3        215644
4           967
            ...
10836        38
10837         4
10838         3
10839       114
10840    398307
Name: Reviews, Length: 10840, dtype: int64
```

- Chuyển đổi dữ liệu Price thành dữ liệu số:

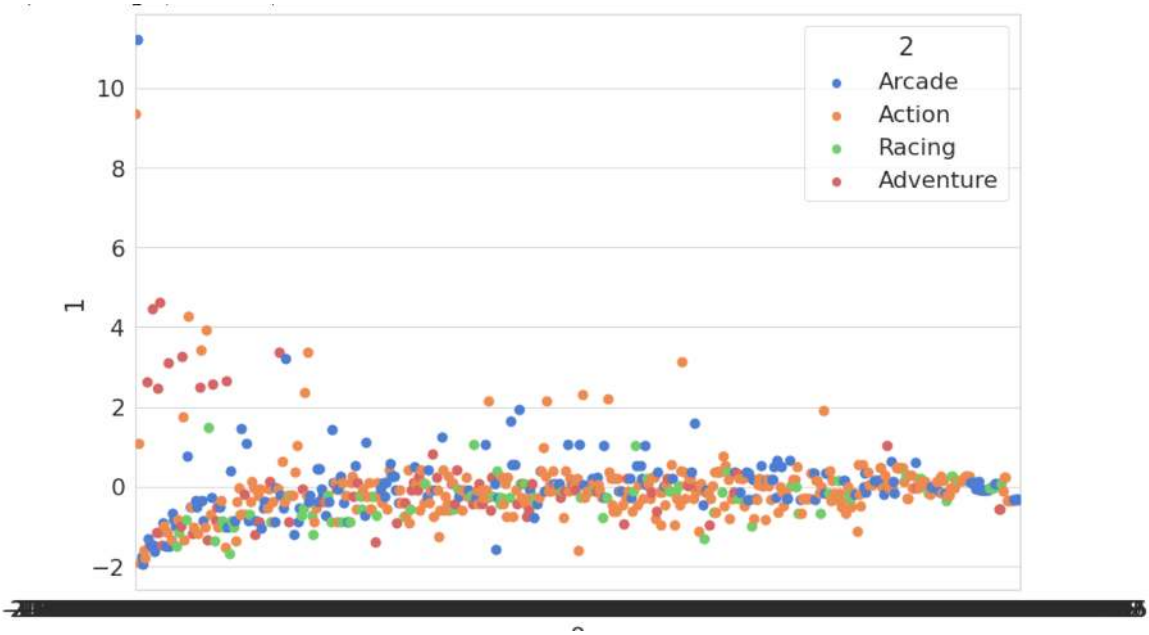
```
0        0.0
1        0.0
2        0.0
3        0.0
4        0.0
        ...
10836    0.0
10837    0.0
10838    0.0
10839    0.0
10840    0.0
Name: Price, Length: 10840, dtype: float64
```

- Điểm tương giữa các thuộc tính Reviews,Rating,Size,Install,Price

- Xem xét chỉ số thú vị - tỷ lệ "Lượt đánh giá và số lượng cài đặt"



- Đồ thị sau khi sử dụng phương pháp PCA



KẾT LUẬN

Đối với dữ liệu nhiều chiều, phương pháp sử dụng thuật toán phân tích thành phần chính PCA cho kết khả quan, có ý nghĩa khoa học và giá trị thực tiễn. Tuy nhiên trong gia đoạn thử nghiệm nên các kết quả giảm chiều chưa được như mong đợi. Điều này do việc trích chọn đặc trưng cũng như việc lựa chọn các tham số phù hợp cho bài toán.

Trong thời gian tới, chúng em sẽ tiếp tục nâng cấp và hoàn thiện nhằm nâng cao tỉ lệ chính xác để giải quyết bài toán một cách nhanh gọn, tiết kiệm chi phí tối đa và dữ liệu được sử dụng một cách có ích.

TÀI LIỆU THAM KHẢO

[1] https://www.easy-tensorflow.com/tf-tutorials/linear-models/linear-classifier

[2] https://machinelearningcoban.com/2017/01/08/knn/

[3] https://github.com/thandongtb/tf_tutorial/blob/master/classification/mnist_softmax