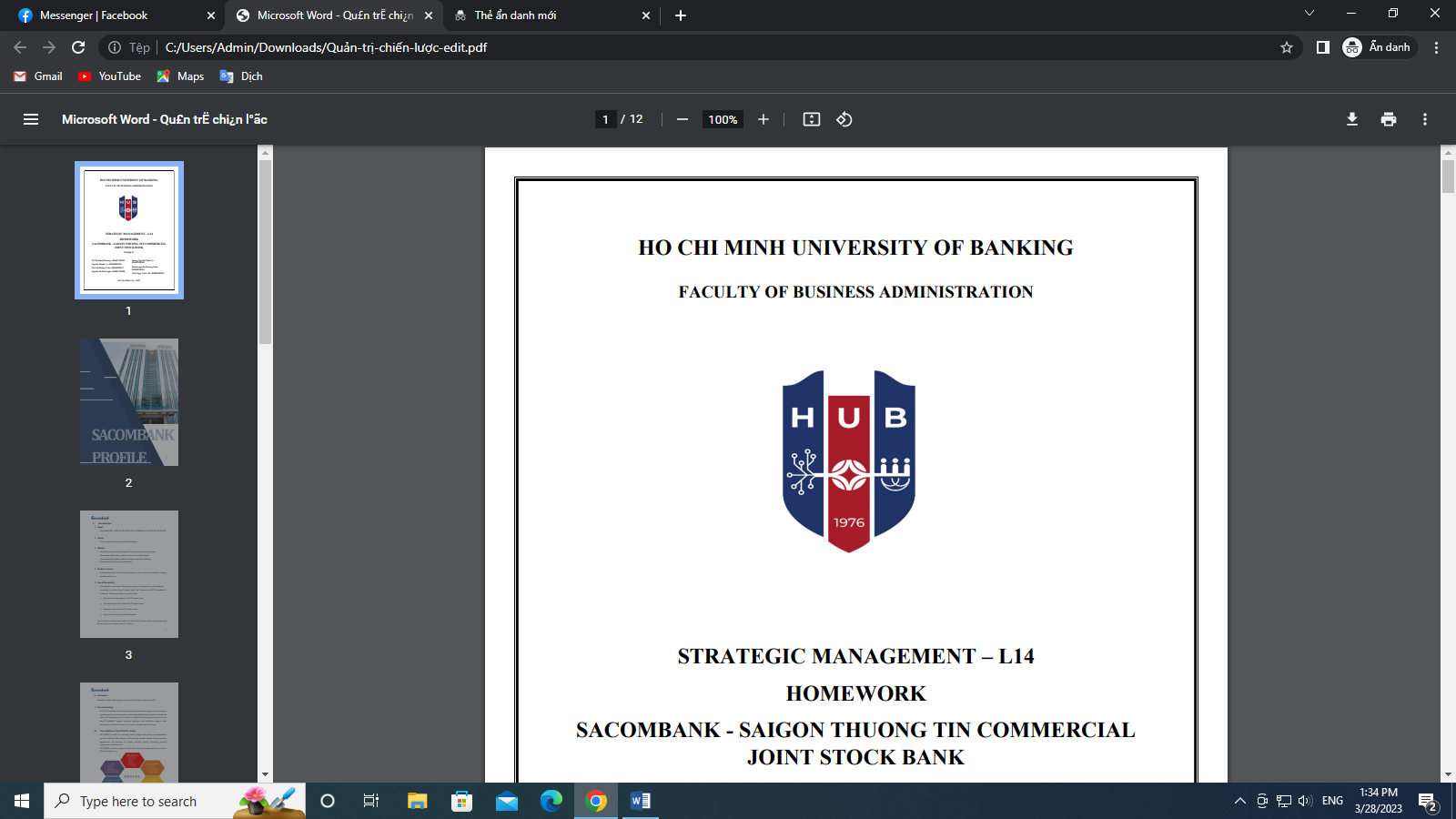
**TRƯỜNG ĐẠI HỌC NGÂN HÀNG THÀNH PHỐ HỒ CHÍ MINH**

**KHOA QUẢN TRỊ KINH DOANH**



**BÀI TẬP NHÓM**

**MÔN: PHÂN TÍCH DỮ LIỆU LỚN**

**MÃ HỌC PHẦN: DAT712\_222\_8\_L14**

**GIẢNG VIÊN HƯỚNG DẪN: TS. NGUYỄN THANH HIÊN**

|  |  |
| --- | --- |
| **Họ tên sinh viên** | **MSSV** |
| Phan Thảo Nhân | 050608200119 |
| Lê Tấn Phát | 050608200559 |

**Thành phố Hồ Chí Minh, tháng 3 năm 2023**

**Câu 1:**

1. Các ma trận và đều có tính đối xứng, vuông góc và thực.

Đầu tiên, xét về tính đối xứng, ta có: nên = và . Từ hai suy luận ở trên, có thể thấy rằng, và . Do đó, có thể kết luận rằng và đều có tính đối xứng.

Tiếp theo, xét về tính vuông góc của hai ma trận, bởi vì ma trận M là ma trận cấp p x q (p dòng, q cột) và ma trận là ma trận cấp q x p (q dòng, p cột). Khi thì đó là ma trận cấp p x p (p dòng và p cột). Tương tự, là một ma trận cấp q x q (kích thước ma trận là q dòng và q cột). Chính vì vậy, cả hai ma trận và đều có tính vuông góc với nhau.

Cuối cùng, xét về tính thực của ma trận, bởi vì ma trận M là một ma trận thực, nên sản phẩm của ma trận M tất nhiên sẽ chuyển vị thành số thực.

1. Gọi là vector riêng của ma trận và là trị riêng tương ứng, khi đó ta có:

Sau đó, nhân cho cả hai vế của phương trình (1), ta được . Vậy, .

Đặt , từ phương trình (2), ta có: . Từ phương trình (3) có thể thấy rằng, u là vector riêng của với trị riêng tương ứng là . Do đó, và có cùng trị riêng, nhưng lại có các vector riêng khác nhau, trong đó vector riêng của là nhưng vector riêng của lại là .

1. Dựa theo lý thuyết đề bài đã cung cấp, ta có công thức của phép phân rã trị riêng như sau:

Trong đó: chứa các trị riêng của dọc theo trục chính

là ma trận trực giao chứa các vector riêng của cũng chính là mỗi cột của

Từ câu 1a) ở trên đã được trình bày, ta có thể nhận định được ma trận là một ma trận có tính đối xứng, vuông góc và thực. Điều đó có thể được hiểu rằng là:

Trong đó: chứa các trị riêng của dọc theo trục chính

là ma trận trực giao chứa các vector riêng của cũng chính là mỗi cột của

1. Trong phân tích SVD, vì U và V ma trận trực chuẩn nên ta có: và còn là ma trận đường chéo nên . Do đó:
2. Tính SVD của M:

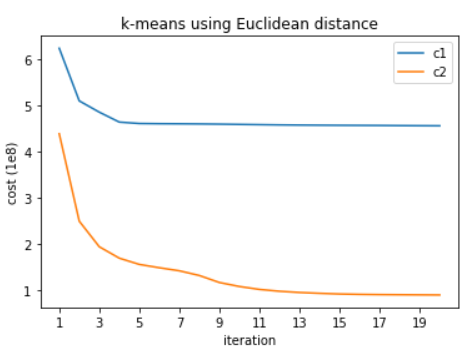
Tính Evals và Evecs:

Cột đầu tiên của V gấp -1 lần vector riêng thứ nhất và cột thứ của V bằng vector riêng thứ hai.

Theo câu (d), ta có: hay nên là các trị riêng của .Vì vậy, các giá trị suy biến của chính là căn bậc hai của các trị riêng của .

**Câu 2:**

1. Đồ thị được vẽ bằng phương pháp Euclidean distance như sau:

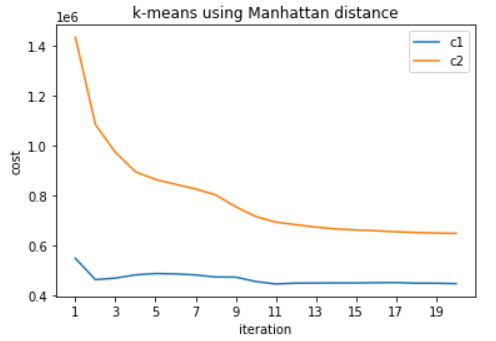


Phần trăm giữa hai tập dữ liệu được tính bằng Euclidean distance tính ra kết quả như sau:

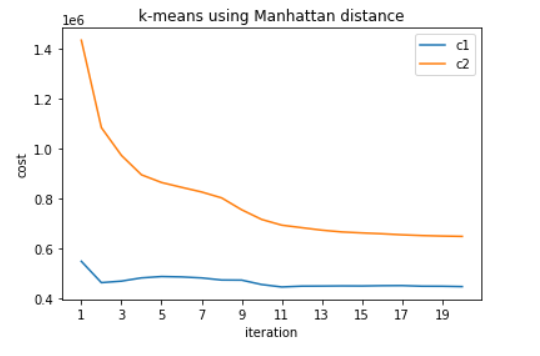


Có thể thấy rằng c2 tốt hơn c1 vì các cụm ban đầu cách xa nhau và ít trùng lặp. Do đó, các cụm sẽ ít phân chia hơn và điều đó dẫn đến kết quả cuối cùng sẽ tốt hơn. Thuật toán K-means giúp giảm khoảng cách Euclidean và xác định tâm cụm giữa các dữ liệu với nhau, ở đây sau 10 vòng lặp của k-means, cost của c2 ít hơn c1 nên nhận định nó tốt hơn.

1. Đồ thị được vẽ bằng phương pháp Manhattan distance như sau:



Phần trăm giữa hai tập dữ liệu được tính bằng Manhattan distance tính ra kết quả như sau:



Tương tự như câu a, có thể thấy rằng c2 tốt hơn c1 vì các cụm ban đầu cách xa nhau và ít trùng lặp. Do đó, các cụm sẽ ít phân chia hơn và điều đó dẫn đến kết quả cuối cùng sẽ tốt hơn. Sau 10 vòng lặp của k-means, cost của c2 ít hơn c1 nên nhận định nó tốt hơn.

Câu 3:

1. Ta có:

(1)

(2)

(3)

(4)

(5)

Nên, biểu thức của có dạng:

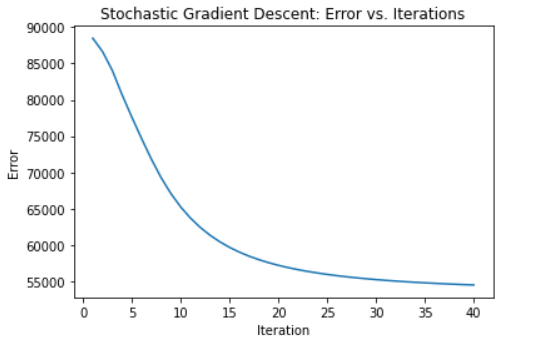
(6)

Kết hợp (2). (3), (4) và (5) ta sẽ có biểu thức của và là:

(7)

(8)

1. Với , ta có sau 40 vòng lặp lại khi đem dữ liệu đi huấn luyện.



Câu 4:

1. Ta có: và .

Mà khi bình phương lên đều cũng bằng 0 hoặc 1. Do đó, có thể kết luận rằng: , điều đó có nghĩa là số lượng vật phẩm mà người sử dụng i thích được tính bằng cấp bậc của người dùng i.

Ta có: vì hoặc 0

Khi có nghĩa là cả người dùng i và người dùng j đều thích vật phẩm k. Vậy, có thể kết luận rằng mang ý nghĩa số lượng vật phẩm k đều được cả hai người dùng i và j đều thích.

1. Gọi là hàng vector thứ i trong có nghĩa là rating của vật phẩm i.

Mà =>

Cosine similarity của vật phẩm i và vật phẩm j là:

The item similarity matrix là:

(1)

Mà (vì có tính đối xứng)

1. =>

Tương tự, gọi là hàng của vector thứ i trong R nghĩa là rating của người dùng i, ta có được the user similarity matrix là:

(2)

Mà (vì có tính đối xứng)

1. =>
2. User-user collaborative filtering:

* Ma trận khuyến nghị là:

1. Item-item collaborative filtering

* Ma trận khuyến nghị là:

1. Tên của 5 truyền hình tivi có điểm chung gần nhất cho Alex đối với lọc cộng tác giữa người dùng-người dùng:

* FOX 28 News at 10pm với điểm số 908.48
* Family Guy với điểm số 861.18
* 2009 NCAA Basketball Tournament với điểm số 827.60
* NBC 4 at Eleven với điểm số 784.78
* Two and a Half Men với điểm số 757.60

Tên của 5 truyền hình tivi có điểm chung gần nhất cho Alex đối với lọc cộng tác giữa phim-phim:

* FOX 28 News at 10pm với điểm số 31.37
* Family Guy với điểm số 30
* NBC 4 at Eleven với điểm số 29.4
* 209 NCAA Basketball Tournament với điểm số 29.23
* Access Hollywood với điểm số 28.97