MÔN : Lập trình song song ứng dụng

GVHD : Thầy [Phạm Trọng Nghĩa](https://courses.fit.hcmus.edu.vn/user/view.php?id=21&course=1)

Nhóm thực hiện :

Ngô Đăng Quang: 18120526

Nguyễn Đăng Quang: 18120527

Dương Đoàn Bảo Sơn: 18120533

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN—

--

**REASEARCH PROPOSAL**

**Customers Segmentation based on their Credit Card usage behaviour**

Thành phố Hồ chí minh ngày 19/03/2023



**Keywords & Definitions:**

|  |  |
| --- | --- |
| **Customers Segmentation** | **Phân khúc khách hàng**: Là việc doanh nghiệp sắp xếp và phân nhóm khách hàng dựa trên những đặc điểm riêng biệt như nhân khẩu học, hành vi mua hàng, hay tâm lý mua sắm… |
| **Data Preprocessing** | **Tiền xử lý dữ liệu**: Kỹ thuật được sử dụng để chuyển đổi dữ liệu thô thành một tập dữ liệu sạch. |
| **Clustering** | **Gom cụm**: Kỹ thuật phổ biến nhất trong học tập không giám sát, nơi dữ liệu được nhóm dựa trên sự giống nhau của các điểm dữ liệu. |
| **KMeans** | **Phân cụm k-means**: Một phương pháp lượng tử hóa vector dùng để phân các điểm dữ liệu cho trước vào các cụm khác nhau |
| **PCA (Principal component analysis)** | **Phân tích thành phần chính**: Một phương pháp giảm kích thước thường được sử dụng để giảm kích thước của các tập dữ liệu lớn. |

# **Giới thiệu:**

- Dựa vào các hành vi của khách hàng khi sử dụng thẻ tín dụng để phân khúc khách hàng nhằm mục đích xác định các chiến lược tiếp thị. Sử dụng lập trình song song để cải thiện thời gian chạy của thuật toán.

# **Một số kiến thức:**

- Sử dụng các tính toán song song trong các thuật toán về khoa học dữ liệu, trong bài là các bước data preprocessing, Kmeans, PCA.

## **a) MultiprocessPandas**

- Giúp mở rộng chức năng của Pandas để dễ dàng chạy các hoạt động trên nhiều lõi (core), tức là song song hóa các hoạt động.

- Phiên bản hiện tại của gói cung cấp khả năng song song hóa các phương thức ***apply()*** trên DataFrames, Series và DataFrameGroupBy.

- Module **applyparallel** sẽ thêm phương thức ***apply\_parallel()*** vào DataFrames, Series và DataFrameGroupBy, cho phép bạn chạy hoạt động trên nhiều lõi. Với:

+ static\_data: Dữ liệu bên ngoài được yêu cầu bởi hàm đã truyền, mặc định là None)

+ num\_processes: Mặc định là số lõi khả dụng tối đa trên CPU của bạn.

+ axis: Chỉ dành cho DataFrames, mặc định là 0 đối với hàng. Với cột, ta đặt axis = 1.

### **b) PCA**

- Trong các bài toán thực tế, các feature vectors có thể có số chiều rất lớn, tới vài nghìn. Ngoài ra, số điểm dữ liệu cũng thường rất lớn. Nếu thực hiện lưu trữ và tính toán trực tiếp trên các tập dữ liệu lớn thì sẽ gặp khó khăn cả về việc lưu trữ và tốc độ tính toán. Do đó, giảm số chiều dữ liệu là một bước quan trọng trong nhiều bài toán. Đây cũng được coi là một phương pháp nén dữ liệu.

**Phân tích thành phần chính (PCA)** là một phương pháp đơn giản nhất trong các thuật toán Giảm chiều dữ liệu (Dimensionality Reduction) dựa trên một mô hình tuyến tính.

- Trở lại với project hiện tại, chúng ta sẽ sử dụng PCA để chuyển đổi dữ liệu thành 2 chiều để trực quan hóa.

## **c) Data Preprocessing**

**Tiền xử lý dữ liệu**

Làm sạch dữ liệu

Mô tả thống kê dữ liệu

Xử lý các outliers

Xử lý các giá trị thiếu

## **d) Modeling**

Định nghĩa các cluster

Nhận xét

**Mô hình hóa**

Kmeans clustering

Trực quan hóa các cluster

Sử dụng PCA để chuyển đổi dữ liệu thành 2 chiều để trực quan hóa

# **Thách thức và học hỏi:**

- Về Data Science, dữ liệu ngày càng nhiều và liên tục nên thời gian xử lý dữ liệu hay một bài toán luôn là vấn đề quan trọng, thử thách cũng như hy vọng học được khi thực hiện thuật toán lần này là khi thực hiện song song hóa trong một bài toán cũ đã biết nhiều với cách thực hiện tuần tự của nó. Việc thời gian chạy và độ chính xác của kết quả là một thách thức.

- Các nhận xét, đánh giá xung quanh việc trực quan dữ liệu đều chỉ mang tính khách quan.

- Nhóm còn chưa hiểu rõ các khía cạnh của các vấn đề về song song khi thực hiện trên python.

- Hy vọng khi thực hiện đồ án này nhóm có thể học hỏi được cách thiết kế và tổ chức một bài toán song song trên python.

# **Tài liệu:**

* Dữ liệu là tập các hành vi sử dụng Credit card của khách hang gồm 18 hành vi

Link data: [Credit Card Dataset for Clustering | Kaggle](https://www.kaggle.com/datasets/arjunbhasin2013/ccdata)

* Tham khảo cách làm xử lý và làm tuần tự của một bài toán trên Kaggle : [Clustering & Visualization of Clusters using PCA | Kaggle](https://www.kaggle.com/code/sabanasimbutt/clustering-visualization-of-clusters-using-pca/notebook)
* Một số tài liệu tham khảo cho Kmeans và PCA: [Beginner’s Guide To K-Means Clustering (analyticsindiamag.com)](https://analyticsindiamag.com/beginners-guide-to-k-means-clustering/), [Phép phân tích thành phần chính – Wikipedia tiếng Việt](https://vi.wikipedia.org/wiki/Ph%C3%A9p_ph%C3%A2n_t%C3%ADch_th%C3%A0nh_ph%E1%BA%A7n_ch%C3%ADnh)

# **Kế hoạch:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Weekly schedule:** 〈what each member has done every week〉   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Week 01  (14/3-21/3) | Week 02  (22/3-28/3) | Week 03  (29/3-4/4) | Week 4  (5/4-11/4) | | 18120533 | Tìm hiểu bài toán | Cài đặt tuần tự | Cài đặt tuần tự | Hoàn thành tuần tự | | 18120526 | Giống 18120533 | Giống 18120533 | Giống 18120533 | Giống 18120533 | | 18120527 | Giống 18120533 | Giống 18120533 | Tìm hiểu cài đặt song song | Tìm hiểu cài đặt song song |  |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Week 05  (12/4-18/4) | Week 06  (19/4-25/4) | Week 07  (26/4-2/5) | Week 8  (3/5-9/5) | | 18120533 | Tìm hiểu cài đặt song song | Cài đặt song song | Kiểm thử | Tổng hợp | | 18120526 | Tìm hiểu cài đặt song song cải tiến | Cải tiến | Giống 18120533 | Tổng hợp | | 18120527 | Cài đặt song song | Giống 18120533 | Giống 18120533 | Làm report | |