ĐẠI HỌC QUỐC GIA
THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

MTH00057 - Toán ứng dụng và thống kê cho Công nghệ thông tin

BÁO CÁO ĐỒ ÁN 1

K Mean Compression Color

Họ tên	**MSSV**
Nguyễn Lê Hồ Anh Khoa	23127211

Giảng viên hướng dẫn

Nguyễn Văn Quang Huy
Trần Hà Sơn
Nguyễn Đình Thúc
Nguyễn Ngọc Toàn

Ngày 20 tháng 6 năm 2025

Mục lục

1 Thông tin sinh viên

Họ và tên: Nguyễn Lê Hồ Anh Khoa. MSSV: 23127211. Lớp: 23CLC09

2 Đánh giá

2.1 Bảng tự đánh giá các yêu cầu đã hoàn thành

Bảng 1: Bảng tự đánh giá đồ án

STT	**Yêu cầu**	**Mức độ hoàn thành**
1	Đọc ảnh.	100%
2	Hiển thị ảnh.	100%
3	Lưu ảnh.	100%
4	Chuyển đổi ảnh từ kích thước 2D (height, width, channels) sang 1D (height × width, channels)	100%
5	Gom nhóm màu sử dụng K-Means.	100%
6	Tạo ảnh mới từ các màu trung tâm (từ K-Means).	100%
7	Cho phép nhập vào tên tập tin ảnh mỗi lần chương trình thực thi.	100%
8	Cho phép lưu ảnh với tối thiểu 2 định dạng là pdf và png.	100%
	Tổng cộng	**100%**

2.2 Đánh giá tổng thể mức độ hoàn thành của bài nộp

Bài nộp đã hoàn thành đầy đủ các yêu cầu đề ra trong bài tập. Tất cả các yêu cầu đều đã được cài đặt và kiểm thử thành công. Tổng thể, bài nộp đã hoàn thành 100% các yêu cầu đề ra.

3 Ý tưởng thực hiện

3.1 Thuật toán K-Means clustering.

3.2 Mục tiêu của K-Means

K-means clustering là một trong những thuật toán cơ bản nhất trong học không giám sát (Unsupervised learning). Thuật toán này dùng để phân dữ dữ liệu đầu vào thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau [1]

3.3 Ý tưởng cơ bản

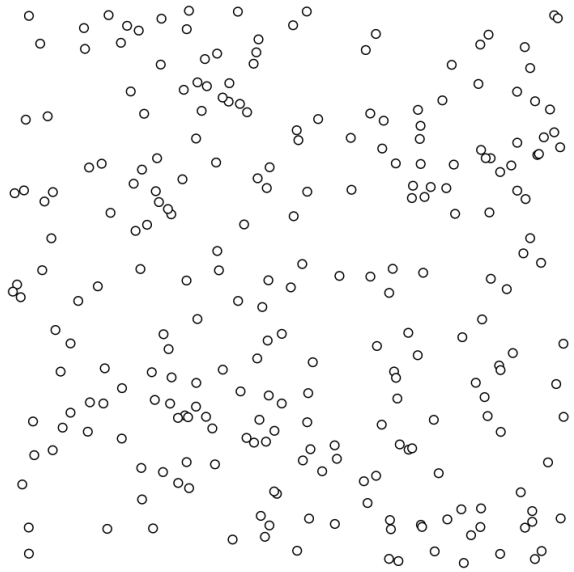
Thuật toán gồm 2 bước chính được lặp đi lặp lại cho đến khi đạt được kết quả tối ưu:

1. Gán mỗi điểm dữ liệu vào cụm gần nhất với nó.
2. Tính toán lại tâm của các cụm dựa trên các điểm dữ liệu đã được gán.
3. Lặp lại bước 1 và 2 cho đến khi không có sự thay đổi nào trong việc gán cụm hoặc đạt đến một số lần lặp tối đa.

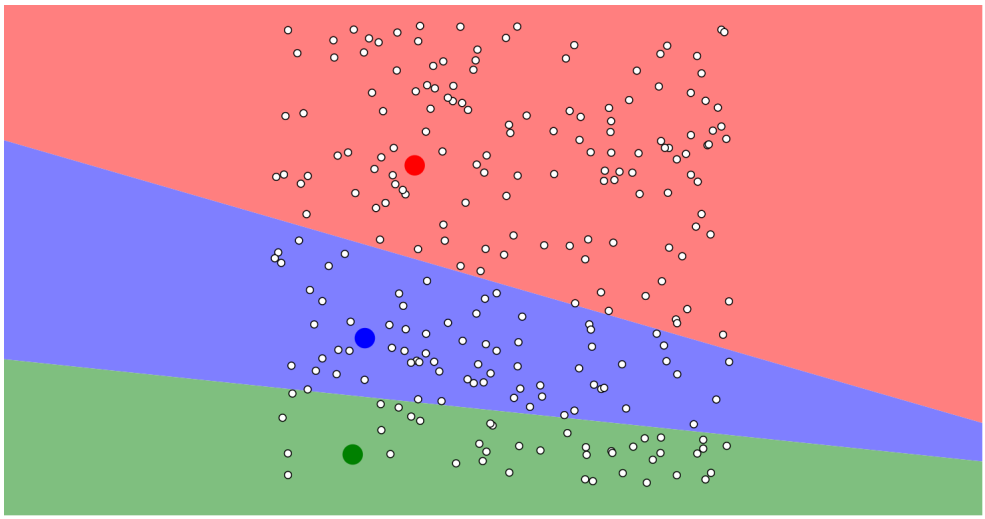
3.4 Mô tả thuật toán

Ta có thể mô tả thuật toán K-Means clustering qua các bước sau:

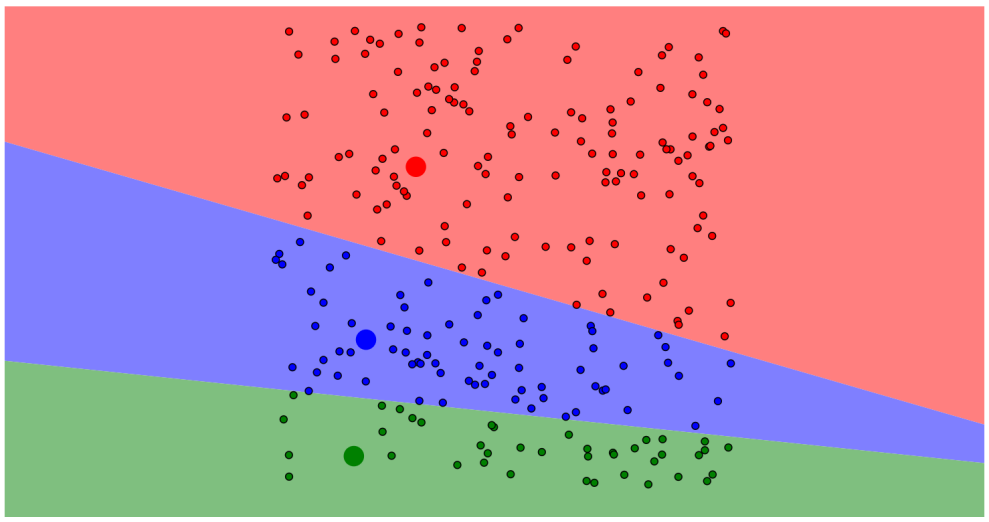
1. Ta có tập dữ liệu ban đầu như sau:



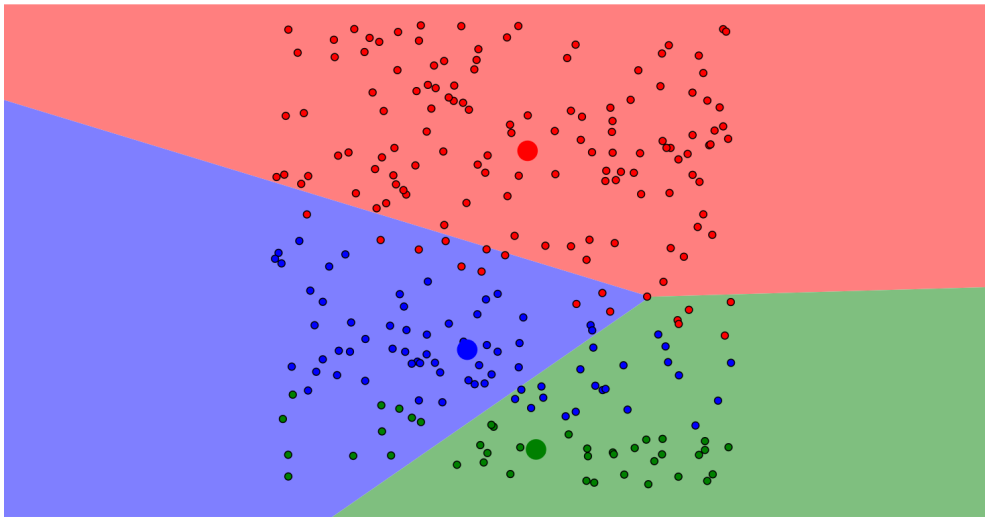
2. Chọn ngẫu nhiên K điểm làm tâm của các cụm phân loại dữ liệu vào từng nhóm. Giả sử ta chọn $K = 3$:



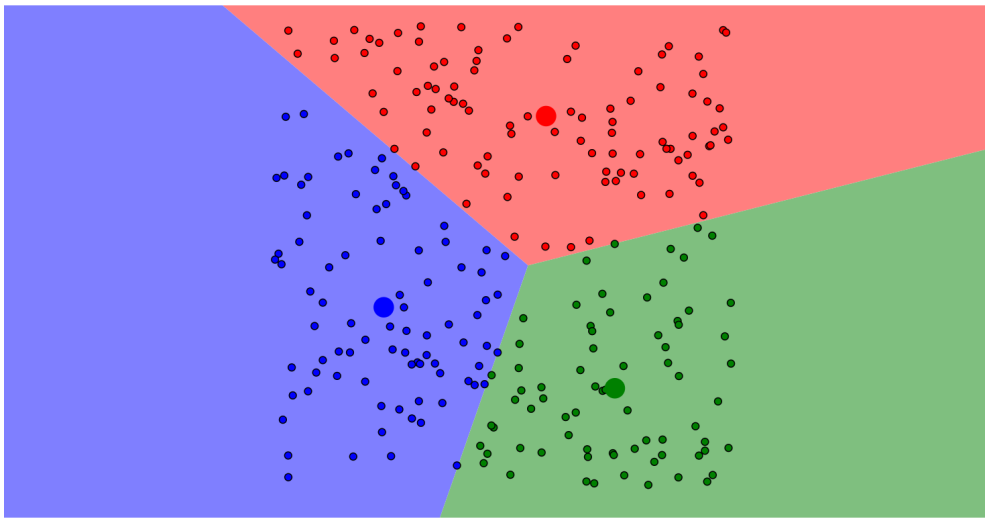
3. Gán mỗi điểm dữ liệu vào cụm gần nhất với nó. Khoảng cách giữa các điểm dữ liệu và tâm cụm được tính bằng khoảng cách Euclid:



4. Tính toán lại tâm của các cụm dựa trên các điểm dữ liệu đã được gán. Tâm cụm mới sẽ là trung bình của tất cả các điểm dữ liệu trong cụm đó:



5. Lặp lại bước 3 và 4 cho đến khi không có sự thay đổi nào trong việc gán cụm hoặc đạt đến một số lần lặp tối đa. Sau một vài lần lặp, ta có thể thu được kết quả cuối cùng như sau:



6. Kết quả cuối cùng là các cụm dữ liệu được phân loại rõ ràng, với mỗi điểm dữ liệu thuộc về một cụm duy nhất

3.5 Ứng dụng thuật toán K-Means clustering trong giảm màu ảnh.

Dựa vào ý tưởng cơ bản của thuật toán K-Means clustering bên trên, ta có thể áp dụng nó để giảm số lượng màu sắc có trong ảnh mà vẫn giữa được các đặc trưng của ảnh ban đầu.

1. Đầu tiên, ta sẽ đọc ảnh và chuyển đổi nó từ định dạng 2D (height, width, channels) sang định dạng 1D (height $\times$ width, channels) để dễ dàng xử lý.

2. Tiếp theo, chọn số lượng màu sắc K mà ta muốn giữ lại trong ảnh.

3. Tạo ra K điểm centroids từ hình ảnh. Có 2 phương pháp để tạo ra các điểm này:

   - **random:**Chọn ngẫu nhiên K điểm centroids trong đoạn từ [0, 255] cho mỗi kênh màu (R, G, B).

   - `in_pixels`: Chọn ngẫu nhiên K điểm centroids từ các màu sắc có trong ảnh.

4. Sử dụng thuật toán K-Means clustering để phân loại các điểm dữ liệu (màu sắc) vào các cụm dựa trên khoảng cách Euclid đến các điểm centroids.

5. Tính toán lại các điểm centroids dựa trên các điểm dữ liệu đã được phân loại.

6. Lặp lại bước 4 và 5 cho đến khi không có sự thay đổi nào trong việc gán cụm hoặc đạt đến một số lần lặp tối đa.

7. Cuối cùng, tạo ra một ảnh mới từ các điểm centroids đã được tính toán. Mỗi điểm dữ liệu trong ảnh ban đầu sẽ được thay thế bằng màu sắc của điểm centroid gần nhất.

4 Mô tả các hàm

5 Kết quả

Tất cả các kết quả dưới đây đều được thực hiện với `max_iter = 100` với kích thước ảnh là 512px x 512px.

5.1 Kết quả với giá trị K = 3

5.2 Kết quả với giá trị K = 5

5.3 Kết quả với giá trị K = 7

5.4 Kết quả với giá trị K = 9

6 Nhận xét

6.1 Chất lượng ảnh đầu ra

6.2 Thời gian thực thi

6.3 Tổng kết

Tài liệu

[1] Tiep Vu Huu, *K-means Clustering*, Jan 1, 2017, https://machinelearningcoban.com/2017/01/01/kmeans