K-Means

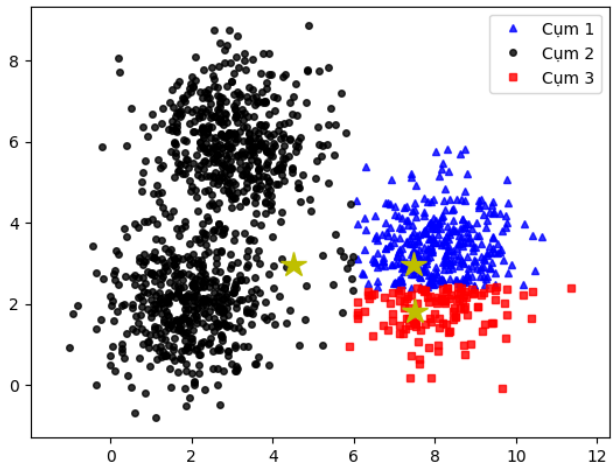
Hồ Nguyễn Phú

December 2024

1 Khái quát về K-Means

1.1 Giới thiệu về K-Means:

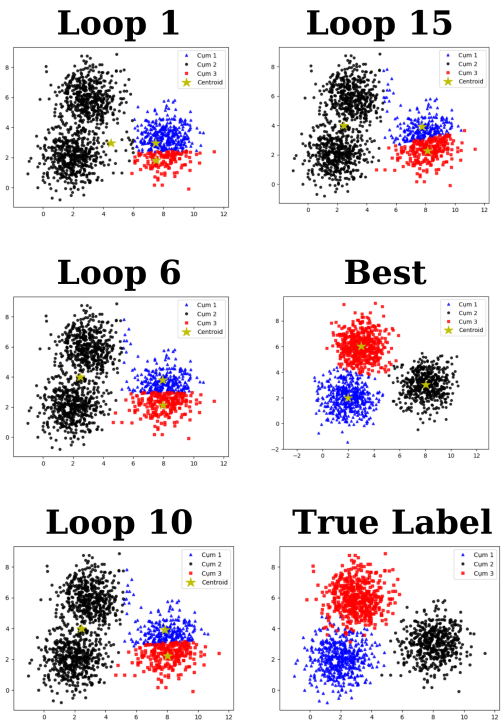
- **K-Means là thuật toán học máy không giám sát.**
- **K-means có ứng dụng rộng rãi trong học máy và các bài toán phân vùng ảnh, xử lý ngôn ngữ tự nhiên.**
- **Thuật toán K-means dựa trên nguyên tắc:**
  - Tìm ra k tâm điểm (centroid) để đại diện cho k cụm dữ liệu.
  - Mỗi điểm dữ liệu sẽ thuộc về cụm có tâm điểm gần nhất với nó. Điều này sẽ đảm bảo các điểm dữ liệu cùng một cụm sẽ có các đặc trưng giống nhau.



Hình 1: Thuật toán K-means đang hoạt động - lần lặp đầu tiên (dữ liệu mẫu từ Bài 4: K-means clustering của Machine Learning Cơ bản).

1.2 Đặc điểm chính:

- **Trong thực tế, ta thường chạy thuật toán K-means nhiều lần và chọn kết quả tốt nhất vì:**
  - Trên lý thuyết, thuật toán K-means có thể có độ phức tạp cao và không gian tìm kiếm lớn (NP-khó). Độ phức tạp tỷ lệ thuận với số chiều của mỗi điểm dữ liệu (ví dụ: mỗi điểm dữ liệu X có thể bao gồm n chiều) và với số k cụm ta cần phân loại.
  - K-means là một thuật toán tham lam ("quyết định sớm và thay đổi hướng đi của giải thuật; không bao giờ xét lại các quyết định cũ"), do đó nó dễ đưa ra những kết quả tối ưu cục bộ thay vì toàn cục.

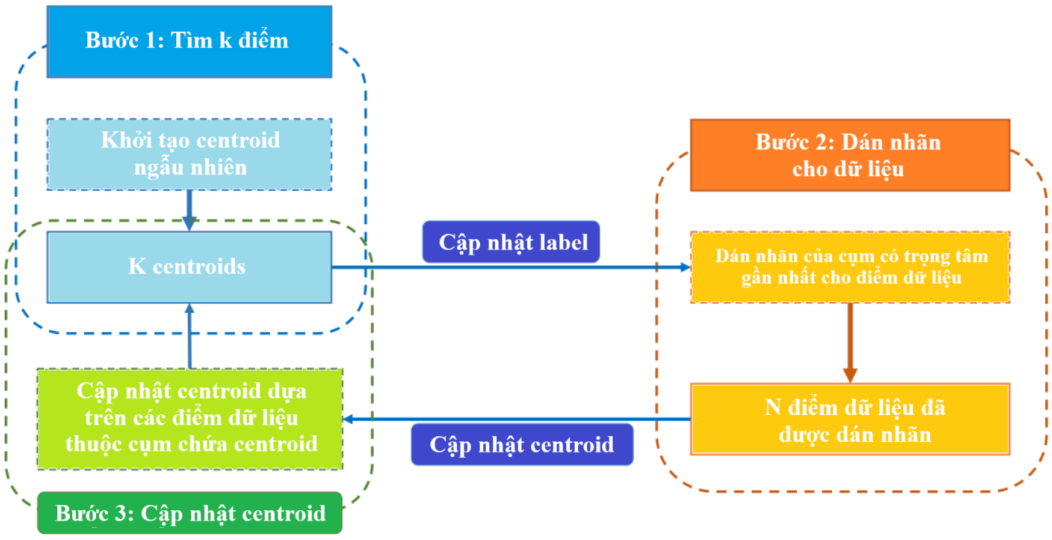


Hình 2: Kết quả cuối cùng của hình 1 - một ví dụ về K-means bị kẹt lại ở cực trị cục bộ.

- **Khoảng cách trong thuật toán K-means là khoảng cách Euclid, tương tự như thuật toán KNN.**

1.3 Các bước của thuật toán K-means:

- **Bước 1**: Chọn ra số cụm k.
- **Bước 2**: Khởi tạo các tâm điểm (centroid) cho từng cụm.
- **Bước 3**: Gán nhãn cho dữ liệu sử dụng one-hot encoding.
- **Bước 4**: Cập nhật các tâm điểm ở mỗi cụm với tâm điểm mới là trung bình cộng của tất cả các điểm trong cụm đó.
- **Bước 5**: Kiểm tra tính hội tụ (nếu hàm mất mát không thay đổi sau một vòng lặp thì ngừng thuật toán).



Hình 3: Các bước thực hiện thuật toán K-means - dịch từ trang 2, slide bài giảng K-means, AIO2024.

2 Lý do nên sử dụng K-mean:

- **Những lợi thế chung của học máy không giám sát:**
  - Tìm ra biểu diễn "tốt"nhất của dữ liệu: vừa giữ lại phần lớn nhưng cũng vừa làm đơn giản hóa thông tin.
  - Tìm ra những đặc trưng ẩn và trích xuất thông tin từ dữ liệu không dán nhãn.
  - Ví dụ: chia dữ liệu thành các cụm, giảm độ nhiễu của data,...
- **Những lợi thế của K-means:**
  - Đơn giản và linh hoạt trong ứng dụng, hội tụ nhanh trong các trường hợp thực tế.
  - Dễ dàng giải thích kết quả mô hình (trong lĩnh vực data).
  - Có khả năng kết hợp với các mô hình khác, đem lại hiệu quả cao trong lĩnh vực thị giác máy tính và xử lý ngôn ngữ tự nhiên.

3 Các bước thực hiện K-means:

3.1 Các thư viện cần sử dụng:

- **Numpy (Numerical Python):**
  - Cho phép thao tác trên vector và ma trận (ndarray).
  - Giúp xử lý dữ liệu dạng bảng / có nhiều chiều.
- **Matplotlib:**
  - Giúp trực quan hóa dữ liệu.
  - Kết hợp tốt với ndarray của Numpy.
- **Scipy:**
  - Sử dụng hàm cdist để tính khoảng cách Euclid giữa hai điểm.

3.2 Khởi tạo dữ liệu test:

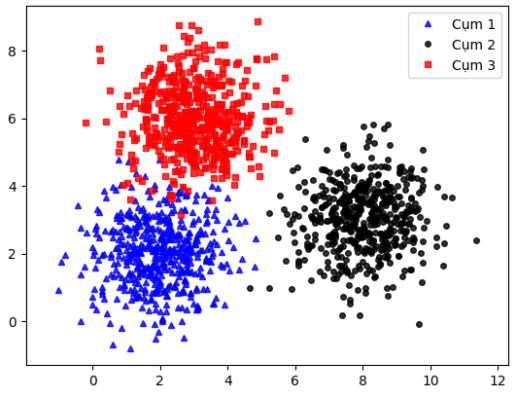
- **Sử dụng khởi tạo của bài viết "Bài 4: K-means Clustering"trên trang web Machine Learning Cơ bản với random seed được đặt cố định (để tái tạo kết quả).**
  - Dữ liệu chia làm 3 cụm X0, X1, X2. Và do mỗi điểm dữ liệu sẽ có 2 chiều nên shape của ndarray X0, X1, X2 đều là (500, 2), tức là một ma trận 500x2.
  - Các điểm dữ liệu được khởi tạo theo phân phối chuẩn 2 chiều, tương ứng với giá trị kỳ vọng means và hiệp phương sai cov.
  - Ta đưa các cụm vào cùng một ma trận với hàm np.concatenate(), tạo thành các điểm cần phân cụm. Do axis = 0 nên np.concatenate() sẽ ghép vào chiều đầu tiên, tạo nên ndarray mới với shape (1500, 2). Nếu ta để axis = 1 thì shape sẽ đổi thành (500, 6).
  - K là số cụm ta muốn chia.
- **Code những phần sau có thể truy cập file.**

```
means = [[2, 2], [8, 3], [3, 6]]
cov = [[1, 0], [0, 1]]
N = 500
X0 = np.random.multivariate_normal(means[0], cov, N)
X1 = np.random.multivariate_normal(means[1], cov, N)
X2 = np.random.multivariate_normal(means[2], cov, N)

X = np.concatenate((X0, X1, X2), axis = 0)
K = 3

original_label = np.asarray([0]*N + [1]*N + [2]*N).T
✓ 0.0s
```

Hình 4: Khởi tạo dữ liệu.



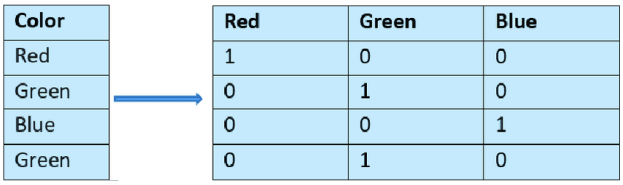
Hình 5: Trực quan hóa dữ liệu với phân cụm gốc.

3.3 Khởi tạo tâm điểm đầu tiên:

- **Có nhiều phương pháp khởi tạo tâm điểm khác nhau:**
  - **Forgy Method**: chọn ngẫu nhiên một trong các điểm dữ liệu => Dễ rơi vào cực tiểu cục bộ.
  - **K-means++**: chọn các tâm điểm càng xa nhau càng tốt => Giảm khả năng rơi vào cực tiểu cục bộ, hội tụ nhanh hơn. Đây là lựa chọn tối ưu
  - **Các lựa chọn khác**: Random Partition, Hierarchical K-means.
- **Để thể hiện rõ tính chất của K-means, phương pháp khởi tạo tâm điểm ta dùng là Forgy Method.**

3.4 Cập nhật nhãn cho các điểm dữ liệu:

- **Ý tưởng**: các điểm sẽ thuộc về cụm có tâm điểm gần nó nhất (dựa trên khoảng cách Euclid).
- **Triển khai**: ta dùng one-hot encoding để tính toán và dán nhãn cho dữ liệu.
  - **One-hot encoding** là phương pháp biểu diễn thưa thớt (sparse representation) đối với dữ liệu.
  - Nhãn là một vector chỉ bao gồm giá trị 1 và 0. Trong đó, nhãn của điểm dữ liệu được đặt là 1, phần còn lại đều là giá trị 0.
  - **One-hot encoding** không chỉ giúp mã hóa nhãn thành một chuỗi số cho máy tính, mà nó còn có ý nghĩa về mặt tính toán khi nhân ma trận / vector: giúp tập trung vào nhãn đang xét (với giá trị là 1) và bỏ qua những nhãn còn lại (với giá trị là 0).
  - **One-hot encoding** thường được kết hợp với hàm argmax() hoặc argmin(). Khi cập nhật nhãn ở K-means, ta dùng các phép tính ma trận để có được một vector khoảng cách từ điểm đang xét đến từng tâm điểm đang có, sau đó dùng argmin() để chọn ra tâm điểm gần nhất và gán nó bằng 1; các nhãn còn lại thì gán giá trị 0.
  - **One-hot encoding** được dùng làm nhãn cho các mô hình phân loại như Softmax Regression, các mô hình học sâu,...



Hình 6: Minh họa One-hot Encoding.

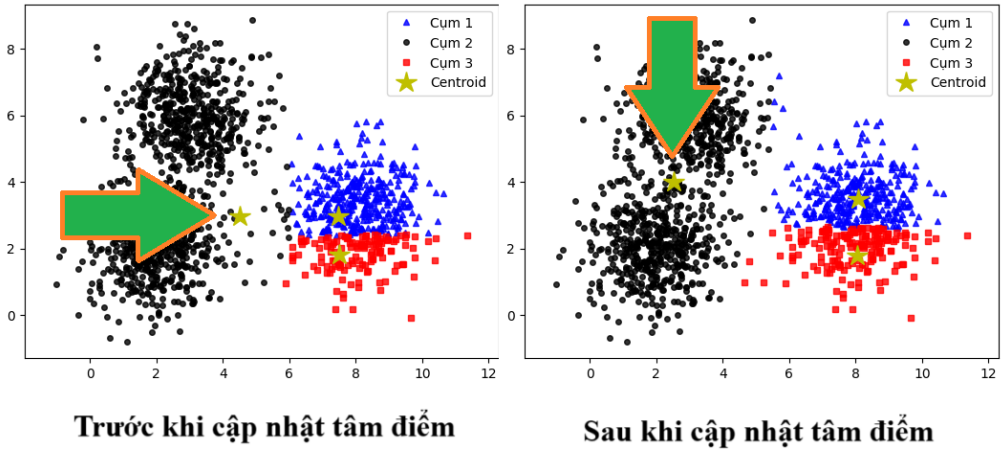
3.5 Cập nhật centroid theo các điểm dữ liệu:

- **Ý tưởng**:
  - Hàm mất mát được sử dụng cho thuật toán K-means là tổng các khoảng cách của từng điểm dữ liệu đến tâm điểm của chúng, biểu diễn như hình dưới đây:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^i - \mu_k\|^2$$

Hình 7: Hàm loss của K-means.

  - Trong đó:
    - Sigma chạy từ i = 1 đến m dùng để quét qua hết các điểm dữ liệu,
    - Sigma chạy từ k = 1 đến K dùng để quét qua hết các tâm điểm và tính khoảng cách.
    - Mỗi w_{ik} là nhãn của điểm dữ liệu biểu diễn dưới dạng vector One-hot encoding. Do vector One-hot encoding nên chỉ giá trị khoảng cách đối với cụm mà điểm dữ liệu thuộc về được xét, các khoảng cách còn lại đều bị bỏ qua (bằng 0). Ta dùng argmax() để trích xuất thông tin từ vector.
  - Để giảm hàm mất mát, ở từng cụm, ta gán tâm điểm mới với điểm trung bình cộng (mean) của tất cả các điểm thuộc cụm đó. Lý do chọn trung điểm là vì đạo hàm của hàm mất mát bằng 0 (hàm số đạt cực tiểu) tại trung bình cộng các điểm trong cụm.



Hình 8: Cập nhật tâm điểm dựa trên trung bình cộng của tất cả các điểm thuộc cụm, rõ nhất ở cụm 2 (Tiếp tục quá trình từ Hình 1).

4 K-means với thị giác máy tính:

Giới thiệu

Thuật toán *K-means* là một phương pháp phân cụm phổ biến được sử dụng rộng rãi trong lĩnh vực *Computer Vision*. Bằng cách phân loại các điểm dữ liệu thành K cụm dựa trên các giá trị tương đồng, K-means giúp giải quyết nhiều bài toán quan trọng trong xử lý và phân tích hình ảnh.

Các ứng dụng chính

1. **Phân đoạn ảnh (Image Segmentation):** K-means có thể được sử dụng để phân đoạn các vùng ảnh thành nhiều phần khác nhau dựa trên màu sắc, cường độ sáng hoặc các thuộc tính khác. Mục tiêu là gom các pixel có đặc điểm tương đồng vào cùng một nhóm, tạo ra các vùng ảnh có ý nghĩa.

2. **Nén ảnh (Image Compression):** K-means được áp dụng trong nén ảnh bằng cách giảm số lượng màu sắc trong ảnh. Quá trình này thực hiện bằng cách nhóm các pixel có màu tương tự vào một cụm và chỉ giữ lại màu trung bình cho mỗi cụm, giúp giảm dung lượng ảnh mà vẫn duy trì chất lượng hình ảnh.

3. **Phát hiện biên (Edge Detection):** Trong một số trường hợp, K-means có thể được sử dụng để phân cụm các điểm ảnh theo cường độ

hoặc độ tương phản, từ đó phát hiện các đường biên của đối tượng trong ảnh.

4. **Nhận dạng đối tượng (Object Recognition):** Thuật toán K-means cũng được áp dụng để phát hiện và nhận dạng các đối tượng trong ảnh. Bằng cách phân cụm các đặc trưng của ảnh, K-means giúp giảm độ phức tạp của dữ liệu và tăng độ chính xác của các thuật toán nhận dạng sau đó.

Kết luận

K-means là một công cụ mạnh mẽ và linh hoạt trong *Computer Vision*, từ việc phân đoạn ảnh đến nén ảnh và phát hiện đối tượng. Với khả năng phân cụm dữ liệu hiệu quả, thuật toán này đóng vai trò quan trọng trong nhiều ứng dụng thực tế.

5 K-means với xử lý ngôn ngữ tự nhiên:

Giới thiệu

Thuật toán *K-means* là một phương pháp phân cụm dựa trên khoảng cách thường được sử dụng trong Xử lý Ngôn ngữ Tự nhiên (NLP). Với khả năng nhóm các đối tượng có đặc tính tương đồng, K-means giúp giải quyết nhiều bài toán trong NLP, đặc biệt là trong việc phân tích dữ liệu văn bản và phát hiện cấu trúc trong ngôn ngữ.

Các ứng dụng chính

1. **Phân cụm tài liệu (Document Clustering):** K-means được sử dụng để nhóm các tài liệu hoặc đoạn văn có nội dung tương tự vào cùng một cụm. Điều này hữu ích trong việc tổ chức các kho văn bản lớn, phân loại tài liệu, và phát hiện các chủ đề tiềm ẩn trong dữ liệu văn bản.

2. **Phân loại chủ đề (Topic Modeling):** K-means có thể hỗ trợ phát hiện các chủ đề ẩn trong tập hợp tài liệu bằng cách phân cụm các từ vựng dựa trên tần suất xuất hiện. Phương pháp này giúp khám phá các chủ đề chính mà không cần phải dán nhãn trước cho dữ liệu.

3. **Phân tích cảm xúc (Sentiment Analysis):** Trong bài toán phân tích cảm xúc, K-means có thể được áp dụng để nhóm các câu hoặc đoạn văn bản dựa trên các cảm xúc tương đồng (tích cực, tiêu cực,

trung tính). Điều này giúp tự động phân loại cảm xúc trong các bài đánh giá, bình luận hoặc các bài đăng trên mạng xã hội.

4. **Giảm số chiều dữ liệu (Dimensionality Reduction):** Với việc sử dụng K-means để phân cụm các từ hoặc cụm từ trong không gian từ vựng (word embeddings), thuật toán giúp giảm số chiều dữ liệu, từ đó làm tăng hiệu quả của các mô hình học máy khi xử lý dữ liệu văn bản.

5. **Tóm tắt văn bản (Text Summarization):** Trong bài toán tóm tắt văn bản, K-means được dùng để chọn lọc các câu tiêu biểu từ tài liệu, dựa trên sự tương đồng giữa các câu. Điều này giúp tạo ra bản tóm tắt cô đọng và phản ánh nội dung chính của văn bản.

Kết luận

Thuật toán K-means mang lại nhiều ứng dụng quan trọng trong Xử lý Ngôn ngữ Tự nhiên, từ phân cụm tài liệu đến phân tích cảm xúc và giảm số chiều dữ liệu. Khả năng nhóm các đối tượng có đặc điểm tương đồng giúp K-means trở thành một công cụ mạnh mẽ trong việc khai thác và phân tích dữ liệu văn bản.