**BÁO CÁO CHƯƠNG 6**

**CLUSTERING IMAGES**

1. K’mean clustering

* Kmean là một thuật toán đơn ginả mà cố gắng đẻ chia dữ liệu đầu vào vào k cụm. nó thực hiện bằng cách lặp đi lặp lại một ước lượng các điểm trung tâm ban đầu như sau:
  + Khởi tạo các điểm trung tâm ui, i=1, ..k, ngẫu nhiên
  + PHân mỗi điểm dữ liệu tới một class ci của cái điểm trung tâm gần nó nhất.
  + Cập nhật các điểm trung tâm là trung bình của tất cả các điểm dữ liệu được phân tới lớp đó.
  + Lặp lại bước 2, 3 tới khi hội tụ
* K-means cố gắng tối ưu tổng phương sai: A picture containing text, watch

  Description automatically generated. Trong đó xj là các vector dữ liệu. Thuật toán trên thường được chạy một vài lần với các điểm trung tâm khác nhau. Sau đó giải pháp với V nhỏ nhất được chọn.
* Nhược điểm lớn nhất của thuật toán này là số lượng clusters cần phải được quyết định trước và một lựa chọn không thoã màn sẽ dẫn đến một kết quả phân cụm tệ.
* Ưu điểm là nó khá đơn giản để triển khai, nó có thể song song hoá và làm việc tốt với phạm vi vấn đề lớn mà không cần điều chỉnh.
* **The Scipy clustering package:**

Chart, scatter chart

Description automatically generated

**Hình 1 Một ví dụ k-means về phân cụm các điểm 2D**

* + **Clustering images**
* **Graphical user interface, text, application

  Description automatically generated**

Hình 2 Kết quả của phân cụm k-means với k=4 và 5 thành phần principal components

Graphical user interface, text, application

Description automatically generated

Hình 3 Kết quả của phân cụm k-means với k=4 và 7 thành phần principal components

* + **Visualizing the images on principal components**

**Chart, scatter chart

Description automatically generated**

**Hình 4 Kết quả của phép chiếu các ảnh trên một cặp các principal components ( the first and second principal components)**

* + **Clustering pixels**

**A picture containing histogram

Description automatically generated**

**Hình 5 Clustering with k=3 and 50\*50 resolution**

A picture containing graphical user interface

Description automatically generated

**Hình 6 Clustering with k=3 and 100\*100 resolution**

1. Hierarchical clustering

* Hierarchical clustering – Phân cụm theo thứ bậc là một thuật toán phân cụm đơn giản nhưng hiệu quả khác. Ý tưởng của nó là xây dựng một cây tương đồng dựa vào khoảng cách theo cặp. Thuật toán bắt đầu với việc nhóm 2 đối tượng gần nhất( dựa vào khoảng cách giữa các vector khoảng cách) và tạo một nốt trung bình trong cây với hai đối tượng như là con. Sau đó cặp gần nhất được tìm thấy giữa các đối tượng còn lại nhưng cũng bao gồm các node trung bình khác. Tạo mỗi node là khoảng cách giữa hai node con. Phân cụm có thể được trích xuất bằng việc đi ngang qua cây này và dừng tại các điểm với khoảng cách nhỏ hơn một ngưỡng mà sau đó xác định được kích thước cụm.
* Tạo ra hai lớp cho các node của câu là ClusterNode và ClusterLeafNode, để tạo ra cây cluster. Hàm hcluster() dùng để tạo cây. Chúng ta sẽ tính khoảng cách giữa hai node và tìm ra khoảng các bé nhất để tạo một cây mới hoàn chỉnh. Sử dụng các thủ tục đệ quy để tìm chiều cao, chiều sâu, id, khoảng cách giữa cây.
* **Clustering images**
  + Xem mỗi kênh màu như các vectors và chọn 8 bín cho mỗi chiều dữ liệu màu. Sau khi làm phẳng sẽ được vector đặc trưng 8 \* 8\* 8 = 512 bins.

**Chart, bar chart, histogram

Description automatically generated**

**Hình 7 Cluster dendrogram**

* **Graphical user interface, application

  Description automatically generated**

**Hình 8 Clusters from the 100 images of sunsets using threshold set to 23% of maximum distance of tree**

Diagram, schematic

Description automatically generated

Hình 9 Dendrogram for the font images

1. Spectral clustering

* Similarity matrix (or affinity matrix or distance matrix) cho n phần tử (ví dụ là các ảnh) là một ma trận n \* n với các điểm tương tự giữa các cặp. Spectral clustering lấy tên từ việc sử dụng phổ của ma trận được xây dựng từ một similarity matrix. Các vector tiêng của ma trận được sử dụng cho việc giảm kích thước sau đó phân cum.
* Lợi ích của phương pháp phân cụm phổ là đầu vào cần chỉ là matrix này và có có thể được xây dựng từ bất kì độ đo tương tự nào. Phương pháp tương tự như k-means và hierarchical clustering tính toán trung bình của vector thuộc tính và điều này giới hạn các thuộc tính vào một vectors. Với phương pháp phổ, không cần có các vector thuộc tính của bất kì loại nào, chỉ cần khoảng cách hoặc là độ tương tự
* Graphical user interface, text, application, email

  Description automatically generated

**Hình 10 Spectal clustering with k=5**