TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG–HCM
KHOA TOÁN – TIN HỌC

Báo cáo Seminar

ĐỀ TÀI: ÁP DỤNG CÁC THUẬT TOÁN PHÂN CỤM TRONG BÀI TOÁN PHÂN KHÚC KHÁCH HÀNG

Thực hiện yêu cầu cấp bằng Cử nhân khoa Toán – Tin học

TÁC GIẢ

Sinh viên Trần Huỳnh Nghĩa – MSSV: 20110251

Sinh viên Đỗ Tấn Phát - MSSV: 20110270

GIẢNG VIÊN HƯỚNG DẪN

TS. Trần Anh Tuấn

LỜI CẢM ƠN

Chúng tôi rất vui và cảm ơn thầy Trần Anh Tuấn đã đồng hành hỗ trợ nhóm chúng tôi trong nghiên cứu về vấn đề ứng dụng các thuật toán phân cụm vào bài toán phân khúc khách hàng. Sự sâu sắc và kiến thức chuyên sâu của thầy đã đóng vai trò quan trọng trong việc làm sáng tỏ những khía cạnh phức tạp của dự án nghiên cứu của chúng tôi. Chúng tôi rất biết ơn vì sự tận tâm và thời gian quý báu mà thầy dành cho chúng tôi, giúp chúng tôi tiến bộ và phát triển trong nghiên cứu. Thầy đã tạo ra những cơ hội học hỏi quý báu và truyền cảm hứng cho chúng tôi trong quá trình nghiên cứu. Sự hướng dẫn và hỗ trợ từ thầy không chỉ giúp chúng tôi hiểu rõ hơn về vấn đề mà còn giúp chúng tôi phát triển kỹ năng nghiên cứu và đi sâu vào các khía cạnh quan trọng của đề tài. Qua sự đóng góp của thầy, chúng tôi đã có những bước tiến vững chắc hơn và tin tưởng rằng nghiên cứu của chúng tôi sẽ mang lại giá trị thực tiễn trong lĩnh vực này. Chúng tôi ơn cảm tạ thầy vì sự đóng góp to lớn và sự hỗ trợ nhiệt tình của mình.

Các nội dung trong báo cáo

KẾT LUẬN

TÀI LIỆU THAM KHẢO

Các hình trong báo cáo

Các bảng dùng trong báo cáo

Tóm tắt

Trong thị trường cạnh tranh ngày nay, các doanh nghiệp không ngừng tìm cách để đạt được lợi thế cạnh tranh. Một cách như vậy là hiểu rõ hơn về khách hàng và điều chỉnh các chiến lược tiếp thị để đáp ứng nhu cầu của họ. Phân khúc khách hàng là một công cụ đắc lực có thể giúp doanh nghiệp đạt được mục tiêu này. Phân khúc khách hàng bao gồm việc chia khách hàng thành các nhóm dựa trên các đặc điểm chung như nhân khẩu học, hành vi mua hàng và sở thích. Sau đó, các nhóm này có thể được nhắm mục tiêu bằng các thông điệp và ưu đãi tiếp thị phù hợp, giúp cải thiện mức độ tương tác và lòng trung thành của khách hàng.

CHƯƠNG 1: GIỚI THIỆU

1.1 Một số thông tin cơ bản

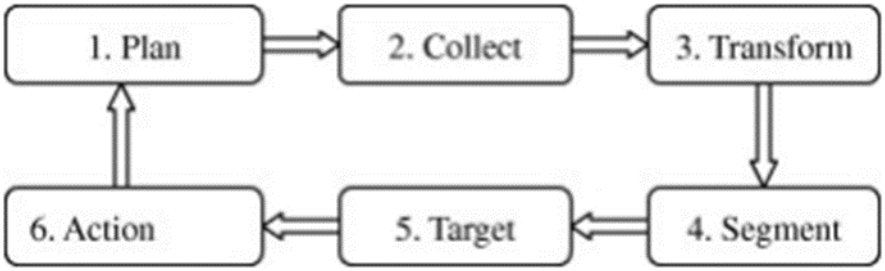
Trong thị trường cạnh tranh ngày nay, các doanh nghiệp không ngừng tìm cách để đạt được lợi thế cạnh tranh. Một cách như vậy là hiểu rõ hơn về khách hàng và điều chỉnh các chiến lược tiếp thị để đáp ứng nhu cầu của họ. Phân khúc khách hàng là một công cụ đắc lực có thể giúp doanh nghiệp đạt được mục tiêu này.

Phân khúc khách hàng bao gồm việc chia khách hàng thành các nhóm dựa trên các đặc điểm chung như nhân khẩu học, hành vi mua hàng và sở thích. Sau đó, các nhóm này có thể được nhắm mục tiêu bằng các thông điệp và ưu đãi tiếp thị phù hợp, giúp cải thiện mức độ tương tác và lòng trung thành của khách hàng.

Cluster analysis là một kỹ thuật phổ biến để phân khúc khách hàng và nó liên quan đến việc phân nhóm khách hàng dựa trên sự giống nhau của họ về các thuộc tính nhất định. Có một số thuật toán phân cụm có thể được sử dụng cho mục đích này, bao gồm K-means, DBSCAN và GMM. Mỗi thuật toán đều có điểm mạnh và điểm yếu riêng và việc lựa chọn thuật toán phụ thuộc vào bản chất của dữ liệu và câu hỏi nghiên cứu.

1.2 Phân cụm khách hàng (Customer Segmentation)

Phân khúc khách hàng là một khía cạnh quan trọng của quản lý quan hệ khách hàng (CRM) và kinh doanh thông minh. Bằng cách chia khách hàng thành các nhóm dựa trên các đặc điểm chung như hành vi, độ tuổi, giới tính, trình độ học vấn, vị trí và tình trạng kinh tế xã hội, doanh nghiệp có thể hiểu rõ hơn về hành vi của khách hàng, điều chỉnh thông điệp tiếp thị cho các phân khúc cụ thể cũng như cải thiện mức độ tương tác và giữ chân khách hàng. Nhân vật tiếp thị, nhân cách hóa một phân khúc khách hàng, thường được tạo từ dữ liệu phân khúc khách hàng và phải phù hợp chặt chẽ với các danh mục khách hàng để có kết quả hiệu quả. Tuy nhiên, mặc dù phân khúc khách hàng có thể là một công cụ mạnh mẽ nhưng nó cũng có những hạn chế, chẳng hạn như thách thức trong việc duy trì phân khúc khi hành vi của khách hàng thay đổi theo thời gian. Bằng cách hiểu và triển khai các loại kỹ thuật phân khúc khách hàng khác nhau, doanh nghiệp có thể cải thiện các chiến dịch tiếp thị của mình, nhắm mục tiêu vào các nhóm khách hàng cụ thể và tăng doanh số bán hàng. Các bước sử dụng để phân khúc người tiêu dùng mục tiêu được mô tả như sau:



Hình 1.1 Các bước phân khúc người tiêu dùng

Hiện nay có nhiều cách để phân chia phân khúc khách hàng, dưới đây là một số cách chia phổ biến và có hiệu quả cao mà các doanh nghiệp thường áp dụng. Dựa vào các dữ liệu mà doanh nghiệp thu thập để chia khách hàng theo các phân khúc khác nhau. Một số được giới thiệu như sau:

ĐẶC ĐIỂM PHÂN KHÚC	NHÂN TỐ
Nhân khẩu học (Demographic)	Tuổi
	Giới tính
	Nghề nghiệp
	Trình độ học vấn
	Dân tộc
	Quốc tịch
	Thế hệ
Thông tin liên quan đến vị trí địa lý (Geographic)	Vị trí địa lý
	Đặc điểm dân cư
	Đặc điểm kinh tế
	Điều kiện địa hình và khí hậu
	Đặc điểm văn hóa và dân tộc
Hành vi khách hàng	Tần suất mua hàng

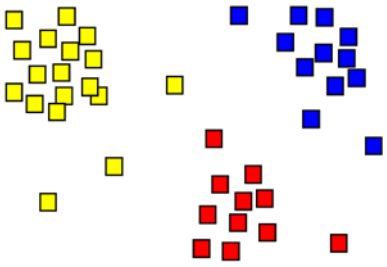
(Behavioral)	Giá trị mua hàng
	Kênh mua
	Sản phẩm mua
	Thời gian mua hàng
Sở thích và lối sống *(Psychographic)*	Giá trị và niềm tin
	Sở thích và lối sống
	Nhận thức và cái nhìn về thương hiệu
	Phong cách mua hàng
	Nhóm cộng đồng mạng xã hội

Bảng 1. Các tiêu chí phân khúc khách hàng

Vậy như đã thấy, sự kết hợp các thuộc tính từ các khía cạnh phân khúc khác nhau là điều cần thiết để phát triển một mô hình phân khúc khách hàng hiệu quả. Nhìn chung, mục đích cuối cùng của phân khúc khách hàng nhằm tối đa hóa giá trị khách hàng cho công ty, tối ưu hóa chiến lược tiếp thị và cải thiện trải nghiệm của khách hàng.

1.3 Thuật toán phân cụm

Phân cụm (Clustering) thuộc loại học không giám sát (Unsupervised learning) là một dữ liệu là bài toán gom nhóm các đối tượng dữ liệu vào thánh từng cụm (cluster) sao cho các đối tượng trong cùng một cụm có sự tương đồng theo một tiêu chí nào đó. Đó là một kiểu phân tích dữ liệu khám phá bao gồm việc nhóm các tập dữ liệu thành một số cụm cụ thể với các đặc điểm có thể so sánh giữa các điểm dữ liệu bên trong mỗi cụm, trực quan hóa hình ảnh khái niệm này như sau:



Hình 1.2 Các điểm dữ liệu sau khi được phân cụm rõ ràng

Các cụm được tạo thành từ các điểm dữ liệu được nhóm lại với nhau theo cách sao cho khoảng cách giữa chúng được giữ ở mức tối thiểu. Nói cách khác, cụm là khu vực có mật độ cao các điểm dữ liệu tương tự.

Phân cụm, còn được gọi là phân tích cụm, là một kỹ thuật phân tích dữ liệu quan trọng và là một lĩnh vực trong khai thác dữ liệu. Nó liên quan đến việc nhóm tập dữ liệu thành các cụm, trong đó các điểm dữ liệu có thuộc tính tương tự được đặt trong cùng một nhóm hoặc cụm. Mục đích là xác định kết nối giữa các điểm dữ liệu dựa trên chất lượng dữ liệu thô của chúng.

Tuy nhiên, thách thức chính là xác định số lượng cụm thích hợp có liên quan và mang lại nhiều thông tin để phân tích. Đây là một quá trình có hệ thống và lặp đi lặp lại bao gồm việc phân tích khối lượng lớn dữ liệu thô để tìm ra các mẫu và điểm tương đồng.

Dữ liệu không có tổ chức sẽ được sàng lọc để có được thông tin chi tiết có ý nghĩa và sau đó các điểm dữ liệu được gán cho các cụm. Để đạt được kết quả tốt trong một lĩnh vực thị trường, một kỹ thuật phân cụm cụ thể kết hợp với việc kiểm tra chi tiết các thuộc tính của cụm dựa trên kết quả phân cụm có thể là lý tưởng.

Tóm lại, các thuật toán phân cụm xác định các nhóm đồng nhất bên trong và đa dạng bên ngoài. Khách hàng có những hành vi, yêu cầu, nhu cầu và đặc điểm riêng biệt và mục đích chính của việc phân nhóm là xác định các loại khách hàng khác nhau và chia cơ sở khách hàng thành các cụm có hồ sơ tương tự để có thể thực hiện tiếp thị mục tiêu hiệu quả hơn.

Phân cụm trong customer segmentation là quá trình chia nhóm khách hàng thành các nhóm con dựa trên các đặc điểm chung như hành vi mua hàng, độ tuổi, giới tính, v.v. Phân cụm giúp doanh nghiệp hiểu rõ hơn về nhu cầu và ưu tiên của từng nhóm khách hàng, từ đó tối ưu hóa chiến lược marketing và dịch vụ để phù hợp với từng nhóm mục tiêu. Các phương pháp phổ biến để phân cụm khách hàng bao gồm phân cụm dựa trên đặc điểm demo, hành vi mua hàng, tâm lý, v.v

1.4 Mô tả bài toán

1.4.1 Chuẩn bị dữ liệu

Phân cụm khách hàng hiệu quả là một nhiệm vụ quan trọng đối với các doanh nghiệp, nhưng việc lựa chọn các biến số phù hợp, xác định số lượng biến cần xem xét và xác định kích thước mẫu thí nghiệm phù hợp có thể gây khó khăn.

Việc chọn lựa sai các đặc trưng, bao gồm các đặc trưng không liên quan hoặc dư thừa, có thể dẫn đến các giải pháp phân cụm không chính xác hoặc không có ý nghĩa, trong khi kích thước mẫu thí nghiệm không phù hợp có thể dẫn đến kết quả thiên vị hoặc không hoàn chỉnh.

Do đó, có nhu cầu xác định các phương pháp tốt nhất và hướng dẫn cho việc lựa chọn các biến số, xác định số lượng biến cần xem xét và xác định kích thước mẫu thí nghiệm phù hợp để đảm bảo phân cụm khách hàng chính xác và ý nghĩa.

1.4.2. Làm sạch và tiền xử lí dữ liệu

Việc làm sạch và tiền xử lý dữ liệu là những bước quan trọng trong việc chuẩn bị dữ liệu cho phân tích. Tuy nhiên, nhiều tổ chức đối diện với những thách thức trong các lĩnh vực này, bao gồm xử lý dữ liệu thiếu, định dạng không đồng nhất, các mục nhập trùng lặp và các điểm ngoại lệ.

Ngoài ra, việc lựa chọn các kỹ thuật tiền xử lý phù hợp có thể gặp khó khăn, vì các phương pháp khác nhau có thể ảnh hưởng khác nhau đến chất lượng và độ chính xác của phân tích kết quả. Kết quả là, cần có những chiến lược làm sạch và tiền xử lý dữ liệu hiệu quả có thể đảm bảo tính toàn vẹn của dữ liệu và tối ưu hóa dữ liệu cho phân tích.

1.4.3. Sử dụng các thuật toán phân cụm thích hợp

Phân cụm là một kỹ thuật quan trọng cho phân đoạn khách hàng và phân tích thị trường, nhưng việc lựa chọn thuật toán phân cụm phù hợp là một nhiệm vụ phức tạp. Các nhà tiếp thị và các nhà phân tích dữ liệu phải lựa chọn từ nhiều thuật toán phân cụm khác nhau, mỗi thuật toán có những điểm mạnh và yếu riêng, và đánh giá tính phù hợp của chúng đối với tập dữ liệu cụ thể và vấn đề kinh doanh cụ thể.

Việc lựa chọn các thuật toán không phù hợp có thể dẫn đến các giải pháp phân cụm không chính xác hoặc không hoàn chỉnh, khiến cho các nỗ lực tiếp thị trở nên không hiệu quả.

Do đó, việc xác định thuật toán phân cụm phù hợp nhất có thể cung cấp cái nhìn chính xác và ý nghĩa cho quá trình ra quyết định tiếp thị là rất quan trọng để phân đoạn khách hàng thành công và triển khai các chiến dịch tiếp thị có mục tiêu.

1.4.4. Tính toán số cụm tối ưu

Xác định số cụm tối ưu là một bước quan trọng trong phân tích phân cụm. Tuy nhiên, đây là một nhiệm vụ đầy thách thức vì nó đòi hỏi phải chọn đúng phương pháp để tính toán số lượng cụm và diễn giải kết quả. Việc lựa chọn phương pháp sai có thể dẫn đến kết luận không chính xác, có thể gây ra hậu quả đáng kể.

Như vậy, cần có một phương pháp có hệ thống và đáng tin cậy để tính toán số cụm tối ưu có tính đến các đặc điểm của tập dữ liệu và thuật toán phân cụm được sử dụng. Vấn đề là phát triển một phương pháp có thể xác định chính xác số cụm tối ưu cho một tập dữ liệu và thuật toán phân cụm nhất định, xem xét cỡ mẫu, số lượng biến và các yếu tố liên quan khác.

1.4.5. Kiểm tra tính xác thực cho việc phân cụm

Xác thực hiệu quả các kết quả phân cụm là điều cần thiết để đảm bảo rằng các cụm thu được từ dữ liệu có ý nghĩa và hữu ích. Tuy nhiên, việc xác định thử nghiệm xác nhận tối ưu cho một vấn đề phân cụm nhất định là một điều khó khăn vì có nhiều biện pháp và phương pháp xác thực khác nhau, và mỗi phương pháp đều có điểm mạnh và điểm yếu riêng.

Hơn nữa, hiệu quả của thử nghiệm xác nhận đã chọn có thể phụ thuộc vào loại thuật toán phân cụm được sử dụng, đặc điểm của dữ liệu và kết quả mong muốn của phân tích phân cụm. Do đó, việc xác định thử nghiệm xác nhận thích hợp để sử dụng trong từng vấn đề phân cụm là rất quan trọng để đảm bảo độ tin cậy và tính hợp lệ của kết quả.

1.4.6. Giải thích và mô tả các cụm

Mặc dù việc phân tích phân cụm rất hữu ích trong việc xác định các nhóm khách hàng hoặc các đối tượng khác biệt, nhưng việc diễn giải và mô tả các nhóm phân cụm kết quả vẫn là một thách thức đáng kể. Điều này bởi vì sự phức tạp và thường là tính chủ quan trong việc diễn giải các giải pháp phân cụm, điều này liên quan đến việc hiểu được các mẫu ngầm và mối quan hệ giữa các biến số và các nhóm phân cụm.

Mặt khác, các thuật toán phân cụm và các phương pháp xác minh khác nhau có thể tạo ra các giải pháp phân cụm khác nhau, làm cho việc lựa chọn giải pháp tốt nhất trở nên khó khăn. Kết quả là, có nhu cầu cho các phương pháp mạnh mẽ và chuẩn hóa hơn cho việc diễn giải và mô tả các phân cụm có thể giúp các nhà nghiên cứu và người thực hành có thông tin sâu hơn về các đặc tính và hành vi của các phân cụm và đưa ra các quyết định có căn cứ hơn dựa trên kết quả phân cụm.

1.5 Thiết kế bài báo cáo

Báo cáo này được chia thành sáu chương, mỗi chương đề cập đến các khía cạnh khác nhau của dự án:

Chương 1 cung cấp cái nhìn tổng quan toàn diện về dự án. Nó bao gồm thông tin cơ bản, trình bày vấn đề, động lực của dự án, phạm vi, mục tiêu, đóng góp của dự án, những điểm nổi bật về thành tựu của dự án và cách tổ chức tổng thể của báo cáo. Chương này làm tiền đề cho các chương tiếp theo, xác định bối cảnh và mục đích nghiên cứu

Trong **Chương 2** Chương 2 tập trung vào mô tả và phân tích ba phương pháp phân cụm quan trọng trong lĩnh vực học máy và khám phá dữ liệu, đó là thuật toán K-Means, Mô hình phân cụm Gaussian (GMM) và thuật toán DBSCAN. Thuật toán K-Means là một phương pháp phân cụm không giám sát, nó nhằm tới việc gom nhóm các dữ liệu thành các cụm dựa trên sự tương tự của chúng. Thuật toán này hoạt động dựa trên việc xác định các điểm trung tâm của các cụm và sau đó gán mỗi điểm dữ liệu vào một cụm tương ứng với điểm trung tâm gần nhất. Mô hình phân cụm Gaussian (GMM) là một phương pháp phân cụm thống kê sử dụng phân phối Gaussian để mô hình hóa các cụm dữ liệu. Nó cho phép mô hình hóa các cụm có hình dạng và kích thước linh hoạt thông qua các tham số tự nhiên. Thuật toán DBSCAN (Density-Based Spatial Clustering of Applications with Noise) sử dụng mật độ để xác định các cụm trong không gian dữ liệu. Phương pháp này phân loại các điểm dữ liệu dựa trên mật độ của chúng và có thể xác định được các cụm có hình dạng và kích thước không đều. Ba phương pháp này cung cấp các phương tiện quan trọng để phân tích và phân cụm dữ liệu trong nhiều ứng dụng thực tế.

Trong **Chương 3** thiết kế mô hình tổng thể và các phương pháp được sử dụng trong dự án sẽ được thảo luận. Chương này đưa ra giải thích chi tiết về phương pháp đã chọn, cung cấp cho người đọc sự hiểu biết rõ ràng về các phương pháp và kỹ thuật được sử dụng.

Chương 4, cung cấp kết quả của việc phân cụm từ đó đưa ra các chiến lược nhằm giải quyết các vấn đề đối với từng cụm cho doanh thu các kỳ sắp tới của của hàng.

Chương 5 sẽ thảo luận về các phương pháp học sâu : Deep clustering network (DCN) , Deep Adaptive Clustering (DAC) , Deep Embedded Clustering (DEC) , Information Maximizing Self-Augmented Training (IMSAT) , Variational Deep Embedding (VaDE)

CHƯƠNG 2: TỔNG QUAN LÝ THUYẾT

2.1 Thuật toán K – means

K-means clustering nhắm tới việc phân chia n quan sát thành k cụm trong đó mỗi quan sát thuộc về cụm có trung bình gần nhất, đóng vai trò là mẫu của cụm

Giả sử ta có tập D = { x_1 , … , x_n } gồm N quan sát , D ϵ $R^{d \times N}$ .Ta muốn với D như vậy , ta có thể phân thành K < N cụm . Giờ đây , ta đặt μ_k làm center , $\mu_k \epsilon R^{d \times 1}$, k = 1, … Với mỗi x_i thì y_i [y_{i1}, … , y_{ik}] mô tả cụm k và dữ liệu x_i được gán cho . Nếu x_i được gán cho cụm k thì y_{ik} = 1 , y_{ij} = 0 . Ta hình dung dễ hiểu , nếu ta có [1,0,…,0] thì x_i ta đang xét đang thuộc vào cluster 1 ;[0,1,…,0] thì x_i ta đang xét thuộc vào cluster 2. Ràng buộc mà y_i đưa ra ta biểu diễn : $y_{ik} \epsilon \{0.1\}$, $\sum_{i=1}^{k} y_{ik} = 1$

Xét hàm mục tiêu : J = $\sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} || x_i - \mu_k ||^2$.

Nhiệm vụ của chúng ta là tiến hành tìm y_{ik} và μ_k sao cho hàm J cực tiểu . chúng ta sẽ làm việc này thông qua quá trình lặp để tìm các tham số tối ưu cho y_{ik} và μ_k . Ta hãy cho μ_k 1 giá trị ban đầu , từ đó hạy giảm thiểu J qua y_{ik} .Sang step 2 , cố định y_{ik} ở step 1. Ta giảm thiểu J bằng μ_k . Ta làm đi làm lại quá trình trên cho đến khi hội tụ .

Ta coi lại thuật toán :

+ Step #01 :

- Ta cho rằng đã có sẵn các μ_k , thì ta giải quyết bài toán là tìm y_i sao cho :

$\mathbf{y_i} = \underset{y_i}{\boldsymbol{argmin}} \sum_{i=1}^{k} \boldsymbol{y_{ik}} || \mathbf{x_i} - \boldsymbol{\mu_k} ||^2$ (mà do $\sum_{i=1}^{k} y_{ik}$ =1) nên ta giải quyết

$$\mathbf{k} = \underset{y_i}{\boldsymbol{argmin}} || \mathbf{x_i} - \boldsymbol{\mu_k} ||^2$$

+ Step #02 :

- Với các y_i đã tìm được như trên , ta đã có được các cụm . Giờ đây , ta sẽ tìm các μ_k mới tức là ta đi giải quyết

$$\boldsymbol{\mu_k} = \underset{\mu_k}{\boldsymbol{argmin}} \sum_{i=1}^{N} \boldsymbol{y_{ik}} || \mathbf{x_i} - \boldsymbol{\mu_k} ||^2$$

- Đặt $f(\mu_k)$ là hàm argmin thì khi ta giải quyết

$$\frac{df}{d\mu_k}(\boldsymbol{\mu_k}) = 2 \sum_{i=1}^{N} \boldsymbol{y_{ik}} [\boldsymbol{\mu_k} - \mathbf{x_i}] = \mathbf{0}$$

Rõ ràng khi giải phương trình trên ta thu được $\mu_k = \frac{\sum y_{ik}\, x_i}{\sum y_{ik}}$ => trung bình cộng trong cluster. Ta cùng xem lại tóm tắt thuật toán K – means:

Đầu vào: Dữ liệu quan sát và số lượng cluster cần tìm.

Đầu ra: Các center và label vector cho từng điểm dữ liệu.

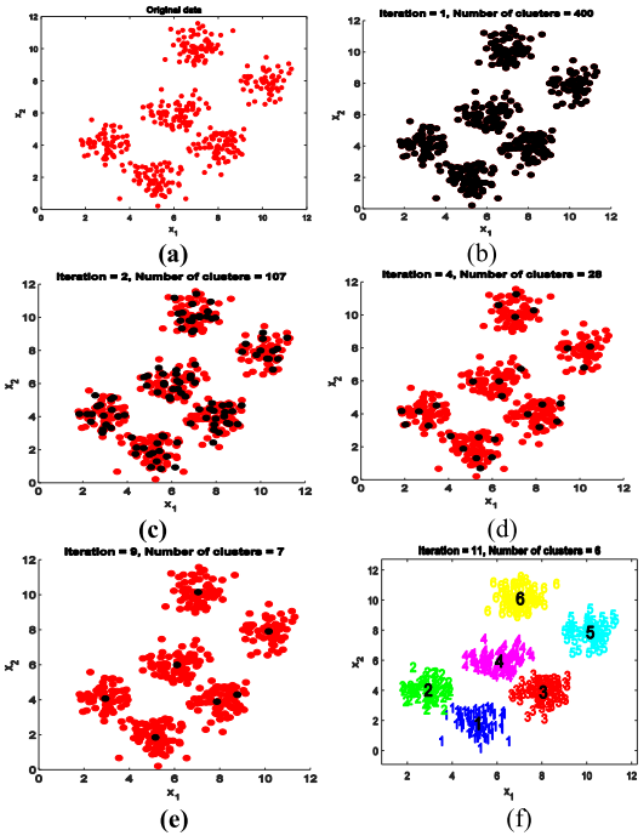
1. Chọn K điểm bất kỳ làm các center ban đầu.

2. Phân mỗi điểm dữ liệu vào cluster có center gần nó nhất.

3. Nếu việc gán dữ liệu vào từng cluster ở bước 2 không thay đổi so với vòng lặp trước nó thì ta dừng thuật toán.

4. Cập nhật center cho từng cluster bằng cách lấy trung bình cộng của tất các các điểm dữ liệu đã được gán vào cluster đó sau bước 2.

5. Quay lại bước 2.



Hình 2.1. Ví dụ cho K – means

(a) Tập dữ liệu gốc;

(b)-(e) Các quy trình của K -means sau 1, 2, 4 và 9;

(f) Hội tụ

2.2 Thuật toán GMM

Ước lượng MLE cho *phân phối Gaussian đa chiều*

Trước khi giải thích GMM ta đi qua khai niệm ước lượng hợp lý tối đa cho các tham số của *phân phối Gassian đa chiều*

Giả sử chúng ta có một bộ dữ liệu gồm các quan sát độc lập và xác định là $D=\{x_1,x_2\ldots,x_N\}$. Trong đó mỗi một $x_i \in R^d$ là một véc tơ quan sát trong không gian d chiều được lấy mẫu từ *phân phối Gaussian đa chiều*. Chúng ta cần ước lượng phân phối của tham số thông qua *ước lượng hợp lý tối đa MLE*.

N quan sát được giả định là độc lập. Do đó hàm hợp lý của phân phối của N quan sát sẽ bằng tích của xác suất trên từng quan sát:

$$
\begin{aligned}
l(\mu, \mathbf{\Sigma}|\mathcal{D}) &= \log \prod_{i=1}^{m} f_{\mathbf{x}_i}(\mathbf{x}_i|\mu, \mathbf{\Sigma}) \\
&= \log \prod_{i=1}^{N} \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i-\mu)^\intercal \mathbf{\Sigma}^{-1}(\mathbf{x}_i-\mu)\right) \\
&= \sum_{i=1}^{N} \left(-\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{x}_i-\mu)^\intercal \mathbf{\Sigma}^{-1}(\mathbf{x}_i-\mu)\right) \\
&= -\frac{N}{2}\log|\mathbf{\Sigma}| - \sum_{i=1}^{N} \frac{1}{2}(\mathbf{x}_i-\mu)^\intercal \mathbf{\Sigma}^{-1}(\mathbf{x}_i-\mu) - \underbrace{\frac{Nd}{2}\log(2\pi)}_{C} \\
&= -\frac{N}{2}\log|\mathbf{\Sigma}| - \sum_{i=1}^{N} \frac{1}{2}(\mathbf{x}_i-\mu)^\intercal \mathbf{\Sigma}^{-1}(\mathbf{x}_i-\mu) + C
\end{aligned}
$$

Lấy đạo hàm bậc nhất của *hàm hợp lý* theo μ và $\sum$.

Đạo hàm theo μ:

Để tính toán đạo hàm bậc nhất chúng ta cần áp dụng công thức:

$$\frac{\partial \mathbf{w}^\top \mathbf{A}\mathbf{w}}{\partial \mathbf{w}} = 2\mathbf{A}\mathbf{w}$$

Coi $\boldsymbol{\Sigma}^{-1} = \mathbf{A}$ và $\mathbf{x}_i - \mu = \mathbf{w}$, khi đó:

$$\begin{aligned}\frac{\partial l(\mu, \boldsymbol{\Sigma}|\mathcal{D})}{\partial \mu} &= -\sum_{i=1}^{N} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mu) \\ &= \boldsymbol{\Sigma}^{-1}(N\mu - \sum_{i=1}^{N} \mathbf{x}_i) \\ &= 0\end{aligned}$$

Nhân cả hai vế của dòng thứ 2 với $\sum$ về phía ngoài cùng bên trái ta suy ra nghiệm $\hat{\mu}$ chính là:

$$\begin{aligned}N\hat{\mu} - \sum_{i=1}^{N} \mathbf{x}_i = 0 \\ \leftrightarrow \hat{\mu} = \frac{\sum_{i=1}^{N} \mathbf{x}_i}{N}\end{aligned}$$

Tương tự ta cũng có ước lượng MLE nghiệm $\hat{\Sigma}$ chính là:

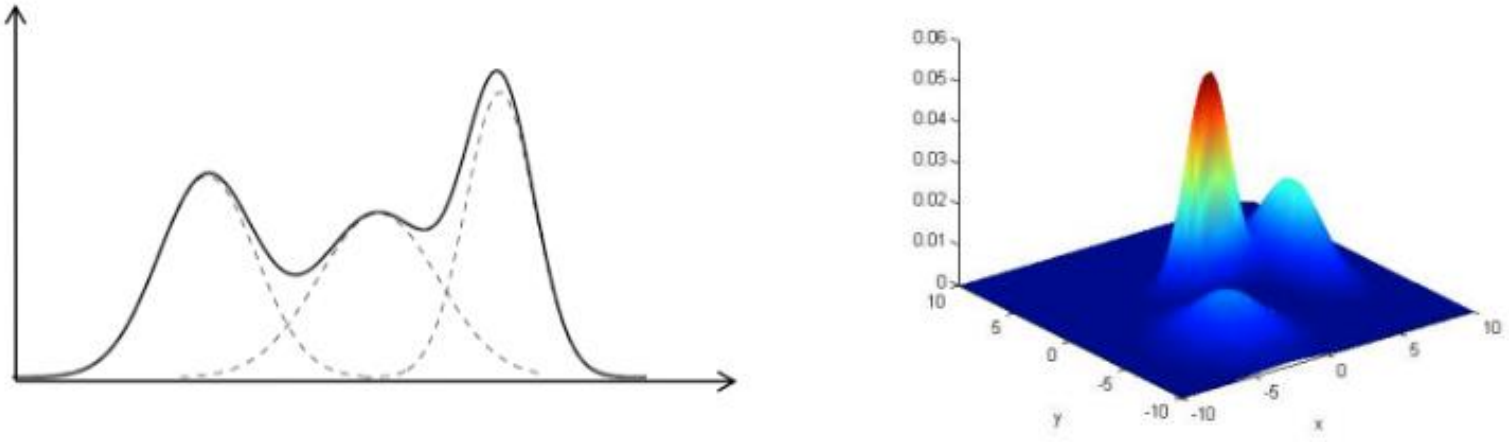
$$\begin{aligned}\frac{N}{2}\hat{\boldsymbol{\Sigma}} - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^\top = 0 \\ \leftrightarrow \hat{\boldsymbol{\Sigma}} = \frac{\sum_{i=1}^{N}(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^\top}{N}\end{aligned}$$

Như vậy ta thu được ước lượng hợp lý tối đa cho các tham số của *phân phối Gassian đa chiều*:

$$\begin{cases}\hat{\mu} &= \frac{\sum_{i=1}^{N} \mathbf{x}_i}{N} = \mathbb{E}(\mathbf{X}) \\ \hat{\boldsymbol{\Sigma}} &= \frac{\sum_{i=1}^{N}(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^\top}{N} = \mathbb{C}\mathrm{ov}(\mathbf{X})\end{cases}$$

GMM

Gaussian Mixture Model (viết tắt GMM) là một mô hình phân cụm thuộc lớp bài toán học không giám sát mà phân phối xác suất của mỗi một cụm được giả định là phân phối Gassian đa chiều. Sở dĩ mô hình được gọi là Mixture là vì xác suất của mỗi điểm dữ liệu không chỉ phụ thuộc vào một phân phối Gaussian duy nhất mà là kết hợp từ nhiều phân phối Gaussian khác nhau từ mỗi cụm.



Hình 2.2 Phân phối Gaussian đa chiều với số cụm k=3 đối với bộ dữ liệu một chiều (bên trái) và hai chiều (bên phải).

Mục tiêu của mô hình GMM là ước lượng tham số phù hợp nhất cho k cụm thông qua phương pháp ước lượng hợp lý tối đa mà chúng ta sẽ thảo luận kĩ hơn ở bên dưới. Một số giả định của mô hình GMM:

- Có k cụm cần phân chia mà mỗi cụm tuân theo phân phối Gaussian đa chiều với tập tham số đặc trưng $\{(\mu_i , \sum_i)\}_{i=1}^{k}$
- Z_k được giả định là một biến ngẫu nhiên nhận giá trị 1 nếu như quan sát x rơi vào cụm thứ k , các trường hợp còn lại nhận giá trị 0.
- Z_k được coi như là một biến ẩn (latent variable hoặc hidden variable) mà ta chưa biết giá trị của nó. Xác suất xảy ra của $P(Z_k =1| x)$ giúp chúng ta xác định tham số phân phối của Gaussian Mixture. Điều này sẽ được thảo luận kĩ hơn bên dưới.

Tập hợp các giá trị Z_k của đối với các cụm sẽ tạo thành một phân phối xác suất sẽ tạo thành một phân phối xác suất $(\pi_1, \pi_2,\ldots, \pi_k)$ trong đó $\pi_k=P(Z_k =1| x)$.

Một xác suất hỗn hợp tại một điểm dữ liệu x sẽ được tính theo công thức Bayes như sau:

$$\begin{aligned} p(\mathbf{x}) &= \sum_{c=1}^{k} p(z_c)p(\mathbf{x}|z_c) \\ &= \sum_{c=1}^{k} p(z_c = 1)p(\mathbf{x}|\mu_c, \mathbf{\Sigma}_c) \\ &= \sum_{c=1}^{k} \pi_c p(\mathbf{x}|\mu_c, \mathbf{\Sigma}_c) \\ &= \sum_{c=1}^{k} \pi_c N(\mathbf{x}|\mu_c, \mathbf{\Sigma}_c) \end{aligned}$$

Thành phần xác suất $P(x | \mu_i , \sum_i)$ được tính từ phân phối Guassian đa chiều và chúng đồng thời là mục tiêu mà chúng ta cần tham số hoá.

Ước lượng hợp lý tối đa

Bài toán đặt ra đó là giả sử chúng ta có một tập dữ liệu $X=\{x_i\}_{i=1}^{N}$ hãy tìm ra ước lượng hợp lý tối đa của các tham số θ sao cho lớp mô hình được giả định là GMM khớp nhất bộ dữ liệu. Như vậy $\theta*$ chính là nghiệm của bài toán:

$$\theta^* = \arg\max_{\theta} p(\mathbf{X}|\theta) = \arg\max_{\theta} \prod_{i=1}^{N} p(\mathbf{x}_i|\theta)$$

Để giải phương trình trên chúng ta có thể dựa trên hai cách tiếp cận:

- Giải trực tiếp phương trình đạo hàm của hàm logarith để theo các hệ số để tìm ra nghiệm tối ưu như đã thực hiện đối với phân phối Gaussian đa biến cho 1 cụm. Tuy nhiên phương pháp này tỏ ra bất khả thi bởi đối với bài toán có nhiều cụm thì hàm mất mát trở nên phức tạp hơn nhiều. Việc giải phương trình đạo hàm dường như là không thể.

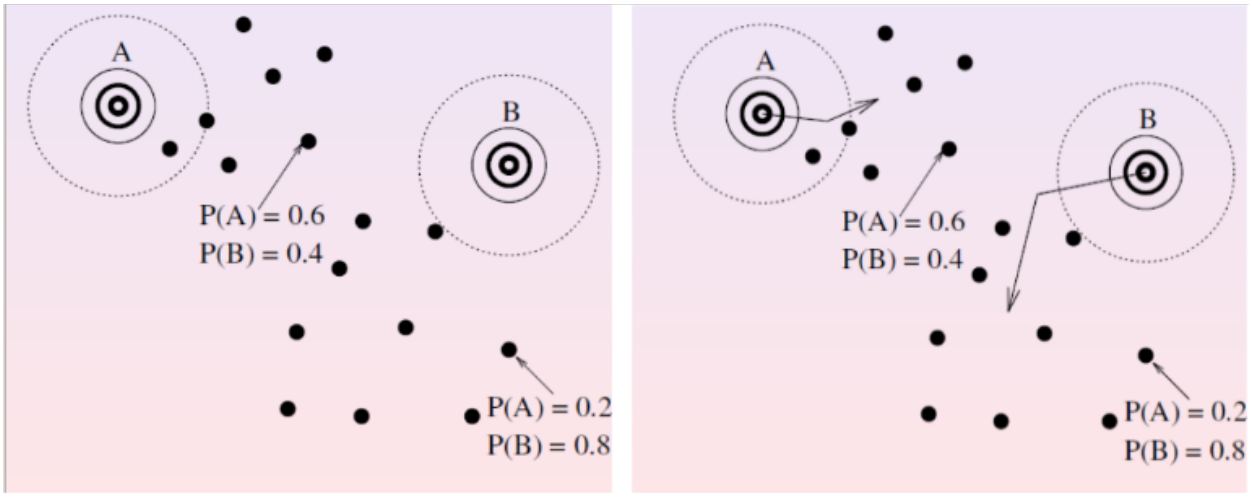
- Sử dụng thuật toán EM (Expectation-Maximization) để cập nhật dần dần nghiệm của θ .

Thuật toán EM là một trong những phương pháp thường được sử dụng để cập nhật nghiệm theo hàm hợp lý. Đây là một phương pháp đơn giản và hiệu quả, phù hợp với các bài toán phức tạp khi mà lời giải trực tiếp từ đạo hàm không dễ dàng tìm kiếm. Bên dưới chúng ta sẽ tiếp tục tìm hiểu phương pháp này:

Trong thuật toán EM chúng ta liên tục thực hiện các vòng lặp mà mỗi vòng lặp bao gồm hai bước huấn luyện chính:

- E-Step: Ước lượng phân phối của biến ẩn z thể hiện phân phối xác suất của các cụm tương ứng với dữ liệu và bộ tham số phân phối.
- M-Step: Tối đa hoá phân phối xác suất đồng thời (join distribution probability) của dữ liệu và biến ẩn.

Cụ thể những bước này sẽ được thể hiện qua hình minh hoạ:



Hình 2.3. E&M step

Hình bên trái là bước E-Step. Tại bước này chúng ta tính toán phân phối xác suất tại từng điểm dữ liệu ứng với mỗi cụm theo bộ tham số phân phối trên từng cụm lúc ban đầu. Chẳng hạn tại một điểm trong hình ở phía trên chúng ta tính ra hai xác suất là P(A)=0.6 và P(B)=0.4 và tại một điểm ở phía dưới tính ra xác suất P(A)=0.2 và P(B)=0.8. Tiếp theo hình bên phải là bước M-Step thể hiện cách cập nhật lại tham số để phù hợp với phân phối của các cụm dữ liệu. Ở đây tham số trung bình của các cụm được cập nhật lại đồng nghĩa với việc dịch chuyển cụm sao cho giá trị hợp lý của phân phối lý thuyết được tối đa hoá và tiến gần tới phân phối thực ở mỗi cụm.

Để cập nhật tham số thì chúng ta xét một hàm auxilary như sau:

$$\begin{aligned} Q(\theta, \theta_t) &= \mathbb{E}_z(\log p(\mathbf{X}, \mathbf{Z}|\theta_t)) \\ &= \sum_z p(z|\mathbf{X}, \theta_t) \log p(\mathbf{X}, \mathbf{Z}|\theta) \\ &= \sum_z p(z|\mathbf{X}, \theta_t) \log [\, p(\mathbf{Z}|\mathbf{X}, \theta) p(\mathbf{X}|\theta)] \\ &= \sum_z p(z|\mathbf{X}, \theta_t) \log p(\mathbf{Z}|\mathbf{X}, \theta) + \underbrace{\left[\sum_z p(z|\mathbf{X}, \theta)\right]}_{1} \log p(\mathbf{X}|\theta) \\ &= \sum_z p(z|\mathbf{X}, \theta_t) \log p(\mathbf{Z}|\mathbf{X}, \theta) + \log p(\mathbf{X}|\theta) \end{aligned}$$

Như vậy $Q(\theta, \theta_t)$. chính là kì vọng của logarith xác suất chung của X và Z trên từng cụm dữ liệu. Giá trị kì vọng này bằng tổng theo trọng số của xác suất tiên nghiệm $P(z \mid X, \theta_t)$. trên từng cụm. Xác suất này có thể tính được dựa trên tham số θ_t. trước đó (θ ở đây là đại diện chung cho cả μ và Σ). Tham số mà chúng ta cần cập nhật sẽ nằm ở log likehood của xác suất chung $\log p(X,Z| \theta)$. Để tính xác suất này chúng ta phân tích chúng theo công thức Bayes giữa $p(Z,X| \theta)$ và $p(X| \theta)$. Cuối cùng chúng ta rút gọn thành tổng giữa logarith hàm hợp lý $\log p(X| \theta)$ và logarith xác suất hậu nghiệm $\log p(Z,X| \theta)$.

Tại sao tối đa hoá hàm hợp lý chúng ta lại thông qua $Q(\theta, \theta_t)$. Đó là bởi khi giá trị $Q(\theta, \theta_t)$. gia tăng thì kéo theo sự gia tăng hàm hợp lý. Như vậy tồn tại một chuỗi vô hạn $\{\theta'_j\}_{j=0}^{\infty}$ sao cho $Q(\theta'_j, \theta_t)$ là một chuỗi tăng và dẫn tới $\{\theta'_j\}_{j=0}^{\infty}$ hội tụ về nghiệm cực đại $\theta*$. Khi đó giá trị hàm hợp lý $\log p(X| \theta')$ cũng là một chuỗi tăng và có nghiệm hội tụ về $\theta*$. Tức là quá trình tìm nghiệm của hàm hợp lý có thể tìm được thông qua hàm $Q(\theta, \theta_t)$.

Tiếp theo ta sẽ chứng minh rằng sự gia tăng của $Q(\theta, \theta_t)$ kéo theo sự gia tăng của hàm hợp lý. Thật vậy:

$$\begin{aligned} Q(\theta, \theta_t) - Q(\theta_t, \theta_t) &= \log p(\mathbf{X}|\theta) - \log p(\mathbf{X}|\theta_t) - \sum_z p(z|\mathbf{X}, \theta_t) \log \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{p(\mathbf{Z}|\mathbf{X}, \theta_t)} \\ &= \log p(\mathbf{X}|\theta) - \log p(\mathbf{X}|\theta_t) - \underbrace{\mathrm{KL}(p(\mathbf{Z}|\mathbf{X}, \theta), p(\mathbf{Z}|\mathbf{X}, \theta_t))}_{\geq 0} \\ &\leq \log p(\mathbf{X}|\theta) - \log p(\mathbf{X}|\theta_t) \end{aligned}$$

Dòng thứ 2 được suy ra là bởi $\sum_z p(z|X,\theta_t)\log\frac{p(Z|X,\theta)}{p(Z|X,\theta_t)}$ chính là một độ đo Kullback-Leibler Divergence về khoảng cách giữa hai phân phối. Giá trị này luôn lớn hơn hoặc bằng 0.

Bất đẳng thức trên cho thấy khi $Q(\theta,\theta_t) \geq Q(\theta_t,\theta_t)$ sẽ kéo theo $\log p(X|\theta) \geq \log p(X|,\theta_t)$. Như vậy thay vì tối đa hoá hàm mục tiêu là hàm hợp lý thì chúng ta có thể tối đa hoá hàm $Q(\theta,\theta_t)$.

Khai triển hàm auxilary

Xác suất xảy ra tại một điểm dữ liệu có thể được biểu diễn :

$$p(\mathbf{x}_i, \mathbf{z}|\theta) = \prod_{j=1}^{k}[p(\mathbf{x}_i, z_j|\theta)]^{z_j} = \prod_{j=1}^{k}[p(\mathbf{x}_i|z_j,\theta)p(z_j|\theta)]^{z_j} = \prod_{j=1}^{k}[p(\mathbf{x}_i|z_j,\theta)\pi_j]^{z_j}$$

Như vậy giá trị hàm hợp lý của phân phối xác suất đồng thời có thể được viết như sau:

$$\mathcal{L}(\mathbf{X},\mathbf{Z}|\theta) = p(\mathbf{X},\mathbf{Z}|\theta) = \prod_{i=1}^{N}\prod_{j=1}^{k}[p(\mathbf{x}_i, z_j|\theta)]^{z_j} = \prod_{i=1}^{N}\prod_{j=1}^{k}[p(\mathbf{x}_i|z_j,\theta)\pi_j]^{z_j}$$

Lấy logarith hai vế ta thu được:

$$\log[p(\mathbf{X},\mathbf{Z})] = \sum_{i=1}^{N}\sum_{j=1}^{k} z_j \log p(\mathbf{x}_i|z_j,\theta) + z_j \log \pi_j$$

Như vậy:

$$
\begin{aligned}
&= \mathbb{E}_z \left[\sum_{i=1}^{N} \sum_{j=1}^{k} z_j \log p(\mathbf{x}_i | z_j, \theta) + z_j \log \pi_j | \theta_t \right] \\
&= \sum_{i=1}^{N} \sum_{j=1}^{k} \mathbb{E}_z[z_j | \theta_t] \log p(\mathbf{x}_i | z_j, \theta) + \mathbb{E}_z[z_j | \theta_t] \log \pi_j \\
&= \sum_{i=1}^{N} \sum_{j=1}^{k} p(z_j | \mathbf{x}_i, \theta_t) \left[\log p(\mathbf{x}_i | z_j, \theta) + \log \pi_j \right] \\
&= \sum_{i=1}^{N} \sum_{j=1}^{k} p(z_j | \mathbf{x}_i, \theta_t) \left[\log \frac{\exp\left(-\frac{1}{2} (\mathbf{x}_i - \mu_j)^\intercal \mathbf{\Sigma}_j^{-1} (\mathbf{x}_i - \mu_j) \right)}{(2\pi)^{d/2} |\mathbf{\Sigma}_j|^{1/2}} + \log \pi_j \right] \\
&= \sum_{i=1}^{N} \sum_{j=1}^{k} p(z_j | \mathbf{x}_i, \theta_t) \left[-\frac{1}{2} \log |\mathbf{\Sigma}_j| - \frac{1}{2} (\mathbf{x}_i - \mu_j)^\intercal \mathbf{\Sigma}_j^{-1} (\mathbf{x}_i - \mu_j) + \log \pi_j + C_j \right]
\end{aligned}
$$

Các bước trong GMM

Bước E-Step:

Mục tiêu của bước E-Step là tính xác suất của mỗi điểm dữ liệu dựa vào phân phối Gaussian đa chiều dựa trên tham số θ_t của vòng lặp gần nhất. Xác suất này được tính như sau:

$$
\begin{aligned}
\mathbb{E}_z(z_j | \mathbf{x}_i, \theta_t) &= 1 \times p(z_j = 1 | \mathbf{x}_i, \theta_t) + 0 \times p(z_j = 0 | \mathbf{x}_i, \theta_t) \\
&= p(z_j | \mathbf{x}_i, \theta_t) \\
&= \frac{p(z_j | \theta_t) p(\mathbf{x}_i | z_j, \theta_t)}{p(\mathbf{x}_i | \theta_t)} \\
&= \frac{\pi_j N(\mu_{jt}, \mathbf{\Sigma}_{jt} | \mathbf{x}_i)}{\sum_j \pi_j N(\mu_{jt}, \mathbf{\Sigma}_{jt} | \mathbf{x}_i)}
\end{aligned}
$$

Xác suất π_j chính là xác suất tiên nghiệm (posteriori probability) bằng với tỷ lệ các quan sát thuộc về cụm j ở vòng lặp thứ t. Trong khi $N(\mu_{jt}, Z_{jt} \mid x_i)$ là xác suất của x_i rơi vào cụm thứ j được tính theo phân phối Gaussian đa chiều. Hai xác suất này có thể tính được và sau cùng ta thu được xác suất rơi vào mỗi cụm tại mỗi một quan sát x_i .

Bước M-Step:

Tại bước M-Step chúng ta cần cập nhật lại tham số phân phối theo hàm auxiliary $Q(\theta, \theta_t)$. Cực trị đạt được khi đạo hàm bậc nhất bằng 0:

$$\frac{\partial Q(\theta, \theta_t)}{\partial \theta} = 0$$

Ở đây θ là các tham số $\{\pi_j, \mu_i, \sum_i\}_{j=1}^{k}$. Lần lượt giải phương trình đạo hàm theo μ_i và $\sum_i$ tương tự như đối với ước lượng MLE đã trình bày :

$$\begin{aligned}
\frac{\partial Q(\theta, \theta_t)}{\partial \mu_j} &= \frac{\partial}{\partial \mu_j} \sum_{i=1}^{N} \sum_{j=1}^{k} p(z_j|\mathbf{x}_i, \theta_t) \left[-\frac{1}{2} \log |\mathbf{\Sigma}_j| - \frac{1}{2} (\mathbf{x}_i - \mu_j)^\intercal \mathbf{\Sigma}_j^{-1} (\mathbf{x}_i - \mu_j) + \log \pi_j + C_j \right] \\
&= \frac{\partial}{\partial \mu_j} p(z_j|\mathbf{x}_i, \theta_t) \left[\sum_{i=1}^{N} \mathbf{\Sigma}_j^{-1} (\mu_j - \mathbf{x}_i) \right] \\
&= \frac{\partial}{\partial \mu_j} \mathbf{\Sigma}_j^{-1} \left[\sum_{i=1}^{N} p(z_j|\mathbf{x}_i, \theta_t) (\mu_j - \mathbf{x}_i) \right] \\
&= 0
\end{aligned}$$

Từ đó suy ra:

$$\mu_j^* = \frac{\sum_{i=1}^{N} p(z_j|\mathbf{x}_i, \theta_t) \mathbf{x}_i}{\sum_{i=1}^{N} p(z_j|\mathbf{x}_i, \theta_t)}$$

Trong đó P(z_i | x_i , θ_t) chính là xác suất tương ứng để x_i thuộc về cụm j được tính từ bước E-Step.

Tiếp theo ta cần tính đạo hàm theo $\sum_j$.

$$\begin{aligned}\frac{\partial Q(\theta,\theta_t)}{\partial \Sigma_j^{-1}} &= \frac{\partial}{\partial \mu_j}\sum_{i=1}^{N}\sum_{j=1}^{k} p(z_j|\mathbf{x}_i,\theta_t)\left[-\frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(\mathbf{x}_i-\mu_j)^\intercal\Sigma_j^{-1}(\mathbf{x}_i-\mu_j) + \log\pi_j + C_j\right] \\ &= \sum_{i=1}^{N} p(z_j|\mathbf{x}_i,\theta_t)\left[\frac{1}{2}\Sigma_j - \frac{1}{2}(\mathbf{x}_i-\mu_j)(\mathbf{x}_i-\mu_j)^\intercal\right] \\ &= 0\end{aligned}$$

Suy ra :

$$\Sigma_j^* = \frac{\sum_{i=1}^{N} p(z_j|\mathbf{x}_i,\theta_t)[(\mathbf{x}_i-\mu_j)(\mathbf{x}_i-\mu_j)^\intercal]}{\sum_{i=1}^{N} p(z_j|\mathbf{x}_i,\theta_t)}$$

Như vậy tham số tối ưu ở mỗi cụm sẽ được cập nhật theo công thức:

$$\mu_j^* = \frac{\sum_{i=1}^{N} p(z_j|\mathbf{x}_i,\theta_t)\mathbf{x}_i}{\sum_{i=1}^{N} p(z_j|\mathbf{x}_i,\theta_t)}$$

$$\Sigma_j^* = \frac{\sum_{i=1}^{N} p(z_j|\mathbf{x}_i,\theta_t)[(\mathbf{x}_i-\mu_j)(\mathbf{x}_i-\mu_j)^\intercal]}{\sum_{i=1}^{N} p(z_j|\mathbf{x}_i,\theta_t)}$$

Để tính π_j chúng ta dựa vào điều kiện ràng buộc $\sum_{j=1}^{k}\pi_j$=1. Khi đó hàm Lagrange tương ứng với Q(θ ,θ_t) là:

$$J(\theta,\theta_t) = Q(\theta,\theta_t) + \lambda(1-\sum_{j=1}^{k}\pi_j)$$

Do đó:

$$\frac{\partial J(\theta, \theta_t)}{\partial \pi_j} = \frac{\partial Q(\theta, \theta_t)}{\partial \pi_j} - \lambda$$
$$= \frac{\sum_{i=1}^{N} p(z_j|\mathbf{x}_i, \theta_t)}{\pi_j} - \lambda = 0$$

Từ đó suy ra:

$$\pi_j = \frac{\sum_{i=1}^{N} p(z_j|\mathbf{x}_i, \theta_t)}{\lambda}$$

Mặt khác ta có $\sum_{j=1}^{k} \pi_j$=1. Do đó:

$$\sum_{j=1}^{k} \pi_j = \frac{\sum_{i=1}^{N} \sum_{j=1}^{k} p(z_j|\mathbf{x}_i, \theta_t)}{\lambda} = \frac{N}{\lambda} = 1$$

Suy ra λ= N và thế vào công thức (1) ta được:

$$\pi_j^* = \frac{\sum_{i=1}^{N} p(z_j|\mathbf{x}_i, \theta_t)}{N}$$

Như vậy chúng ta đã tìm ra được tham số tối ưu của thuật toán GMM sau mỗi vòng lặp. Việc giải trực tiếp bài toán tối ưu hàm hợp lý theo ước lượng MLE là bất khả thi trong điều kiện có nhiều cụm dữ liệu. Chính vì thế thuật toán EM được áp dụng để cập nhật dần dần tham số của mô hình. Thuật toán sẽ dần dần hội tụ sau một hữu hạn bước. Về lý thuyết của thuật toán GMM chúng ta sẽ phải trải qua nhiều tính toán đạo hàm tương đối phức tạp. Tuy nhiên để thực hành thuật toán này lại tương đối dễ dàng trong sklearn.

2.3 Thuật toán DBSCAN

Trước khi tìm hiểu về thuật toán DBSCAN chúng ta xác định một số định nghĩa mà thuật toán này sử dụng.

Định nghĩa 1: Vùng lân cận epsilon (Eps-neighborhood) của một điểm dữ liệu P được định nghĩa là tợp hợp tất cả các điểm dữ liệu nằm trong phạm vi bán kính epsilon (kí hiệu ϵ) xung quanh điểm P. Kí hiệu tập hợp những điểm này là:

$$N_{eps}(P) = \{Q \in \mathcal{D} : d(P,Q) \leq \epsilon\}$$

Trong đó D , là tập hợp tất cả các điểm dữ liệu của tập huấn luyện.

Định nghĩa 2: Khả năng tiếp cận trực tiếp mật độ (directly density-reachable) đề cập tới việc một điểm có thể tiếp cận trực tiếp tới một điểm dữ liệu khác. Cụ thể là một điểm Q được coi là có thể tiếp cận trực tiếp bởi điểm P tương ứng với tham số epsilon và minPts nếu như nó thoả mãn hai điều kiện:

1. Q nằm trong vùng lân cận epsilon của P : $Q \in N_{eps}(P)$
2. Số lượng các điểm dữ liệu nằm trong vùng lân cận epsilon tối thiểu là minPts: $|N_{eps}(Q)| \geq$ minPts

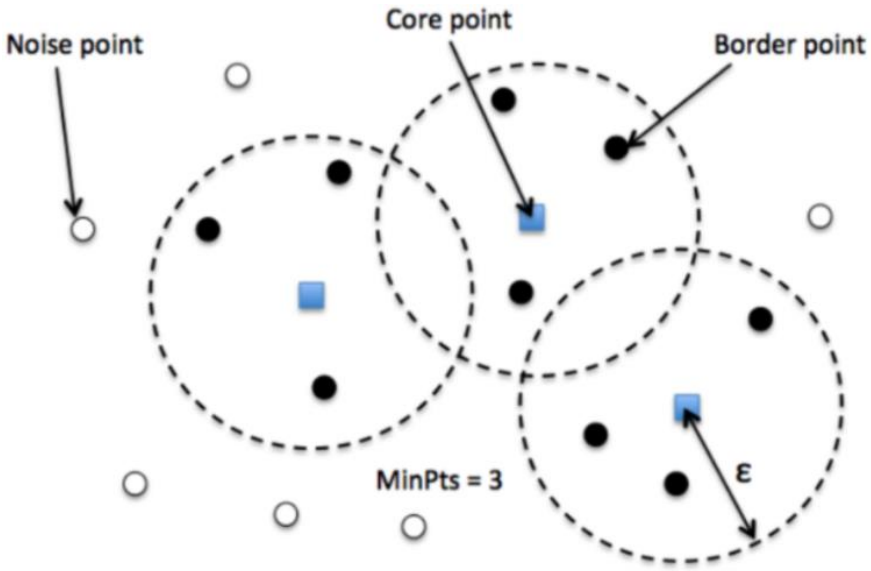
Như vậy một điểm dữ liệu có thể tiếp cận được trực tiếp tới một điểm khác không chỉ dựa vào khoảng cách giữa chúng mà còn phụ thuộc vào mật độ các điểm dữ liệu trong vùng lân cận epsilon phải tối thiểu bằng minPts. Khi đó vùng lân cận được coi là có mật độ cao và sẽ được phân vào các cụm. Trái lại thì vùng lân cận sẽ có mật độ thấp. Trong trường hợp mật độ thấp thì điểm dữ liệu ở trung tâm được coi là không kết nối trực tiếp tới những điểm khác trong vùng lân cận và những điểm này có thể rơi vào biên của cụm hoặc là một điểm dữ liệu nhiễu không thuộc về cụm nào.

Định nghĩa 3: Khả năng tiếp cận mật độ (density-reachable) liên quan đến cách hình thành một chuỗi liên kết điểm trong cụm. Cụ thể là trong một tập hợp chuỗi điểm $\{Pi\}_{i=1}^{n} \subset$ D mà nếu như bất kì một điểm P_i nào cũng đều có thể tiếp cận trực tiếp mật độ (định nghĩa 2) bởi P_{i-1} theo tham số epsilon và minPts thì khi đó ta nói điểm $P= P_n$ có khả năng kết nối mật độ tới điểm $Q=P_1$.

Từ định nghĩa 3 ta suy ra hai điểm P_i và P_j bất kì thuộc chuỗi $\{Pi\}_{i=1}^{n}$ thoả mãn i<j thì P_j đều có khả năng kết nối mật độ tới P_i. Hai điểm bất kì có khả năng kết nối mật độ với nhau thì sẽ thuộc cùng một cụm. Từ đó suy ra các điểm trong chuỗi $\{Pi\}_{i=1}^{n}$ đều được phân về cùng cụm. Khả năng tiếp cận mật độ thể hiện sự mở rộng phạm vi của một cụm dữ liệu dựa trên liên kết theo chuỗi. Xuất phát từ một điểm dữ liệu ta có thể tìm được các điểm có khả năng kết nối mật độ tới nó theo lan truyền chuỗi để xác định cụm.

Phân loại dạng điểm trong DBSCAN

Căn cứ vào vị trí của các điểm dữ liệu so với cụm chúng ta có thể chia chúng thành ba loại: Đối với các điểm nằm sâu bên trong cụm chúng ta xem chúng là điểm lõi. Các điểm biên nằm ở phần ngoài cùng của cụm và điểm nhiễu không thuộc bất kì một cụm nào. Bên dưới là hình vẽ mô phỏng thể hiện ba loại điểm tương ứng nêu trên.



Hình 2.4. Các điểm dữ liệu trong DBSCAN

Trong thuật toán DBSCAN sử dụng hai tham số chính đó là:

minPts: Là một ngưỡng số điểm dữ liệu tối thiểu được nhóm lại với nhau nhằm xác định một vùng lân cận epsilon có mật độ cao. Số lượng minPts không bao gồm điểm ở tâm.

epsilon (kí hiệu ϵ): Một giá trị khoảng cách được sử dụng để xác định vùng lân cận epsilon của bất kỳ điểm dữ liệu nào.

Hai tham số trên sẽ được sử dụng để xác định vùng lân cận epsilon và khả năng tiếp cận giữa các điểm dữ liệu lẫn nhau. Từ đó giúp kết nối chuỗi dữ liệu vào chung một cụm.

Hai tham số trên giúp xác định ba loại điểm:

điểm lõi (core): Đây là một điểm có ít nhất minPts điểm trong vùng lân cận epsilon của chính nó.

điểm biên (border): Đây là một điểm có ít nhất một điểm lõi nằm ở vùng lân cận epsilon nhưng mật độ không đủ minPts điểm.

điểm nhiễu (noise): Đây là điểm không phải là điểm lõi hay điểm biên.

Đối với một cặp điểm (P,Q) bất kì sẽ có ba khả năng:

Cả P và Q đều có khả năng kết nối mật độ được với nhau. Khi đó P, Q đều thuộc về chung một cụm.

P có khả năng kết nối mật độ được với Q nhưng Q không kết nối mật độ được với P. Khi đó P sẽ là điểm lõi của cụm còn Q là một điểm biên.

P và Q đều không kết nối mật độ được với nhau. Trường hợp này P và Q sẽ rơi vào những cụm khác nhau hoặc một trong hai điểm là điểm nhiễu.

Các bước trong thuật toán DBSCAN

Các bước của thuật toán DBSCAN khá đơn giản. Thuật toán sẽ thực hiện lan truyền để mở rộng dần phạm vi của cụm cho tới khi chạm tới những điểm biên thì thuật toán sẽ chuyển sang một cụm mới và lặp lại tiếp quá trình trên

Qui trình của thuật toán:

Bước 1: Thuật toán lựa chọn một điểm dữ liệu bất kì. Sau đó tiến hành xác định các điểm lõi và điểm biên thông qua vùng lân cận epsilon bằng cách lan truyền theo liên kết chuỗi các điểm thuộc cùng một cụm.

Bước 2: Cụm hoàn toàn được xác định khi không thể mở rộng được thêm. Khi đó lặp lại đệ qui toàn bộ quá trình với điểm khởi tạo trong số các điểm dữ liệu còn lại để xác định một cụm mới.

Xác định tham số

Xác định tham số là một bước quan trọng và ảnh hưởng trực tiếp tới kết quả của các thuật toán. Đối với thuật DBSCAN cũng không ngoại lệ. Chúng ta cần phải xác định chính xác tham số cho thuật toán DBSCAN một cách phù hợp với từng bộ dữ liệu cụ thể, tuỳ theo đặc điểm và tính chất của phân phối của bộ dữ liệu. Hai tham số cần lựa chọn trong DBSCAN đó chính là minPts và epsilon:

minPts: Theo quy tắc chung, minPts tối thiểu có thể được tính theo số chiều D trong tập dữ liệu đó là minPts $\geq$ D+1. Một giá trị minPts=1 không có ý nghĩa, vì khi đó mọi điểm bản thân nó đều là một cụm. Với minPts$\leq$2, kết quả sẽ giống như phân cụm phân cấp (hierarchical clustering) với single linkage với biểu đồ dendrogram được cắt ở độ cao y = epsilon. Do đó, minPts phải được chọn ít nhất là 3. Tuy nhiên, các giá trị lớn hơn thường tốt hơn cho các tập dữ liệu có nhiễu và kết quả phân cụm thường hợp lý hơn. Theo quy tắc

chung thì thường chọn minPts=2×dim. Trong trường hợp dữ liệu có nhiễu hoặc có nhiều quan sát lặp lại thì cần lựa chọn giá trị minPts lớn hơn nữa tương ứng với những bộ dữ liệu lớn.

epsilon: Giá trị ϵ có thể được chọn bằng cách vẽ một biểu đồ k-distance. Đây là biểu đồ thể hiện giá trị khoảng cách trong thuật toán k-Means clustering đến k=minPts−1 điểm láng giềng gần nhất. Ứng với mỗi điểm chúng ta chỉ lựa chọn ra khoảng cách lớn nhất trong k khoảng cách. Những khoảng cách này trên đồ thị được sắp xếp theo thứ tự giảm dần. Các giá trị tốt của ϵ là vị trí mà biểu đồ này cho thấy xuất hiện một điểm khuỷ tay (elbow point): Nếu ϵ được chọn quá nhỏ, một phần lớn dữ liệu sẽ không được phân cụm và được xem là nhiễu; trong khi đối với giá trị ϵ quá cao, các cụm sẽ hợp nhất và phần lớn các điểm sẽ nằm trong cùng một cụm. Nói chung, các giá trị nhỏ của ϵ được ưu tiên hơn và theo quy tắc chung, chỉ một phần nhỏ các điểm nên nằm trong vùng lân cận epsilon.

Hàm khoảng cách: Việc lựa chọn hàm khoảng cách có mối liên hệ chặt chẽ với lựa chọn ϵ và tạo ra ảnh hưởng lớn tới kết quả. Điểm quan trọng trước tiên đó là chúng ta cần xác định một thước đo hợp lý về độ khác biệt (disimilarity) cho tập dữ liệu trước khi có thể chọn tham số ϵ. Khoảng cách được sử dụng phổ biến nhất là euclidean distance.

CHƯƠNG 3: PHƯƠNG PHÁP LUẬN

3.1 Tiền xử lí dữ liệu

Tập dữ liệu được sử dụng trong nghiên cứu này là tập dữ liệu "Mall Data" được lấy từ trang Kaggle (**https://www.kaggle.com/code/anoshessam/k-means-clustering/input**).

Bộ dữ liệu Mall Customer bao gồm các thông tin chi tiết về khách hàng tại các trung tâm mua sắm, với cột CustomerID biểu diễn các mã số duy nhất, cột Gender ghi nhận giới tính của từng khách hàng, cột Age thể hiện độ tuổi, cột Annual Income (k$) thể hiện thu nhập hàng năm của khách hàng và cuối cùng, cột Spending Score (1-100) biểu diễn điểm số chi tiêu của họ.

Bộ dữ liệu này cung cấp cơ sở dữ liệu chắc chắn cho việc phân tích đặc tính mua sắm và tiêu dùng của khách hàng tại các trung tâm mua sắm, từ đó tạo ra các cơ hội hiểu sâu hơn về cách họ tương tác với môi trường mua sắm và đề xuất các chiến lược kinh doanh hướng tới sự tối ưu hóa trải nghiệm của khách hàng.

Bằnng cách phân tích các phân khúc khách hàng khác nhau, trung tâm thương mại có thể xác định phân khúc nào có nhiều khả năng mua một sản phẩm cụ thể hơn, điều này cho phép họ tiếp thị sản phẩm một cách hiệu quả đến phân khúc đó thay vì tốn tiền tiếp thị tới tất cả khách hàng trong cơ sở dữ liệu.

Bộ dữ liệu gồm 5 cột và 200 dòng được Python miêu tả:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual Income (k$)      200 non-null    int64
 4   Spending Score (1-100)  200 non-null    int64
dtypes: int64(4), object(1)
    memory usage: 7.9+ KB
```

(i) CustomerID: Mã số khách hàng, kiểu dữ liệu là int64 và không có dữ liệu thiếu (non-null count = 200).
(ii) Gender: Giới tính của khách hàng, có kiểu dữ liệu là object và số lượng dữ liệu không rỗng cũng bằng 200.
(iii) Age: Độ tuổi của khách hàng, kiểu dữ liệu là int64 và có dữ liệu không thiếu.

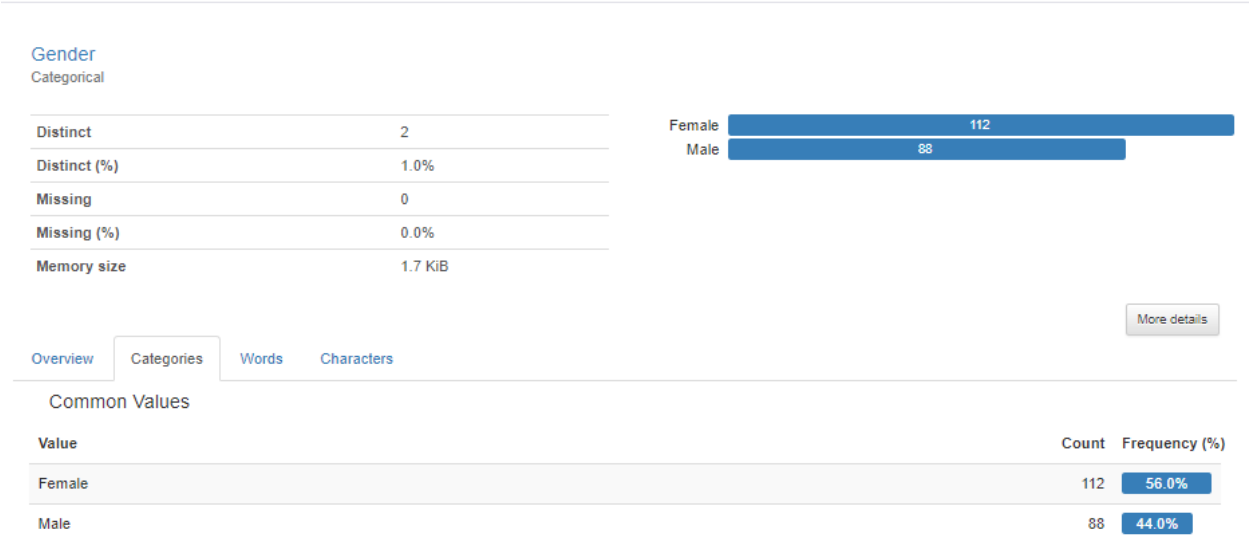
(iv) Annual Income (k$): Thu nhập hàng năm của khách hàng (nghìn đô), kiểu dữ liệu là int64 và không có dữ liệu nào thiếu.
(v) Spending Score (1-100): Điểm chi tiêu (từ 1-100) của khách hàng, kiểu dữ liệu là int64 và không có dữ liệu nào thiếu.

Việc trực quan hóa dữ liệu cho chúng ta thấy vài điều thú vị. Ta thử xem xét việc loại bỏ cột CustomerID bởi nó không ảnh hưởng đến nghiên cứu. Đầu tiên ta xem xét thử các phân bố của các cột, và nói vài thứ liên quan đến phân bố này.

Về phân bố Gender, ta có thể thấy rằng số lượng nữ mua hàng ở trung tâm này nhiều hơn so với số lượng của khách hàng nam. ầu tiên, điều này có thể gợi ý rằng sản phẩm hoặc dịch vụ của trung tâm mua sắm được đánh giá cao hoặc phổ biến đối với đối tượng khách hàng nữ. Điều này khiến trung tâm có thể tập trung vào việc cực kỳ hiệu quả trong việc phân loại, quảng bá và cung cấp sản phẩm và dịch vụ phù hợp với mục tiêu là khách hàng nữ.

Ngoài ra, việc có lượng khách hàng nữ nhiều hơn cũng có thể ảnh hưởng đến cách quảng cáo và marketing của trung tâm, đồng thời giúp xác định các xu hướng mua hàng cụ thể và tạo điều kiện thuận lợi cho việc cải thiện trải nghiệm mua sắm dành cho khách hàng.

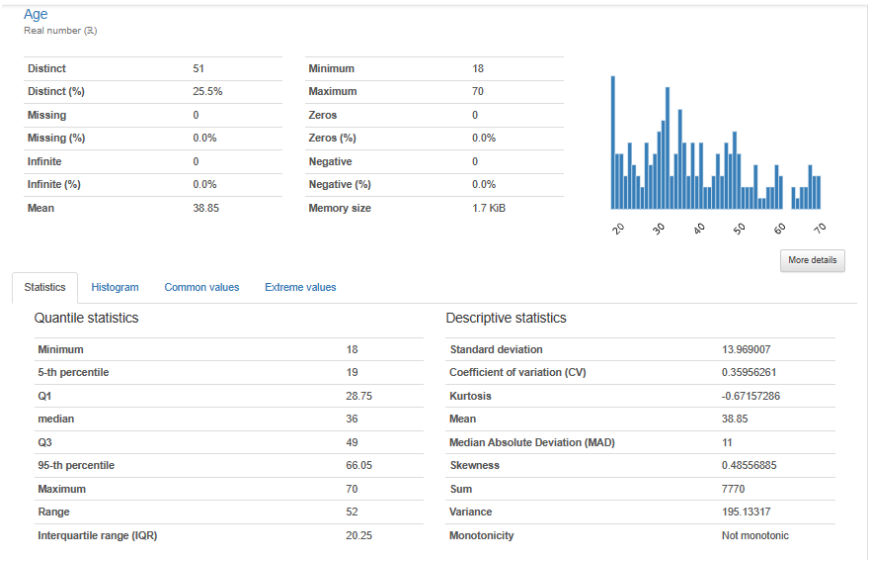
Tuy nhiên, để hiểu rõ hơn ý nghĩa cụ thể của việc có lượng khách hàng nữ nhiều hơn nam, cần phân tích cụ thể thêm về hành vi mua sắm, sở thích và nhu cầu của khách hàng nữ và nam trong ngữ cảnh cụ thể của trung tâm mua sắm đó.



Hình 3.1 Thống kê cho cột Gender

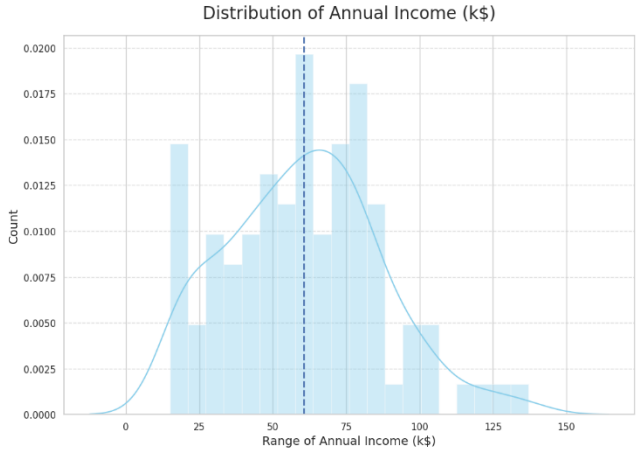
Về thuộc tính Age của dữ liệu, ta thấy rằng dữ liệu này cho thấy độ tuổi của khách hàng tại trung tâm thương mại và cung cấp một cái nhìn tổng quan về phân bố độ tuổi trong cơ sở khách hàng này. Với một

kích thước mẫu 51 và giá trị trung bình là 38.85, chúng ta có thể nhận thấy mức độ trung niên đại diện (độ tuổi khoảng 35-55 tuổi) trong cơ sở khách hàng này có thể cao hơn so với các nhóm độ tuổi khác. Điều này có thể ảnh hưởng đến việc xác định những loại sản phẩm và dịch vụ nào nên được tập trung phát triển hoặc quảng bá, cũng như cách tiếp cận khách hàng và chiến lược marketing. Ngoài ra, với độ tuổi trẻ nhất là 18 và độ tuổi cao nhất là 70, trung tâm thương mại có thể cần xem xét việc cung cấp các sản phẩm và dịch vụ phù hợp với đa dạng độ tuổi của khách hàng, từ người trẻ tuổi đến người cao tuổi trong cộng đồng. Điều này có thể tạo cơ hội để mở rộng sự đa dạng của các sản phẩm và dịch vụ để thu hút và phục vụ khách hàng đa dạng tuổi tác



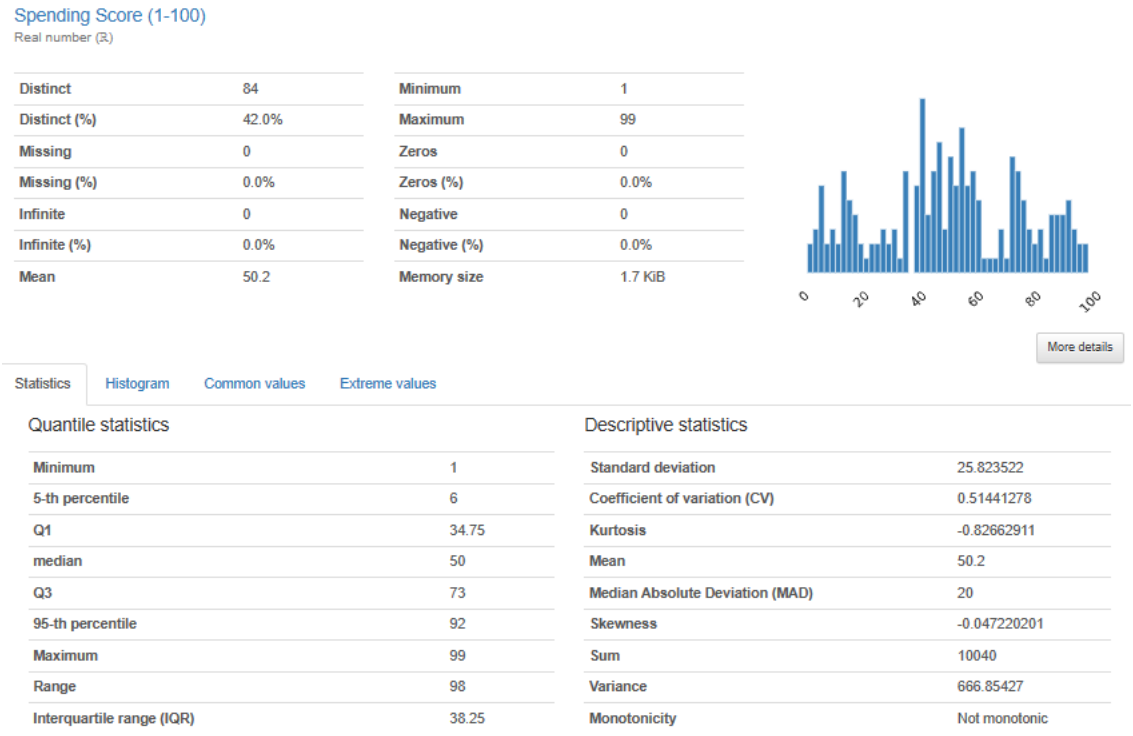
Hình 3.2 Thống kê cho cột Age

Về thuộc tính Anual Income, như đã thấy, ung cấp một cái nhìn sâu sắc về tình hình tài chính và tiêu dùng. Với mức thu nhập trung bình là 60.56k$ và mức thu nhập cao nhất là 137k$, có thể suy ra rằng trung tâm thương mại này đang phục vụ một phân khúc người tiêu dùng có thu nhập tương đối cao. Sự đa dạng của các mức thu nhập từ 15k$ đến 137k$ cũng có thể gợi ý đến sự phong phú và đa dạng của các sản phẩm và dịch vụ mà trung tâm thương mại đang cung cấp. Có thể xem xét việc tăng cường dịch vụ hoặc sản phẩm dành cho đối tượng khách hàng có thu nhập cao hơn để tối ưu hóa doanh thu. Bên cạnh đó, việc không có giá trị thiếu, giá trị vô cùng, hoặc giá trị âm cũng cho thấy tính toàn vẹn và đáng tin cậy của dữ liệu thu thập, tạo điều kiện thuận lợi cho quá trình phân tích kinh tế và quyết định chiến lược kinh doanh.



Hình 3.3 Phân bố của dữ liệu Anual Income

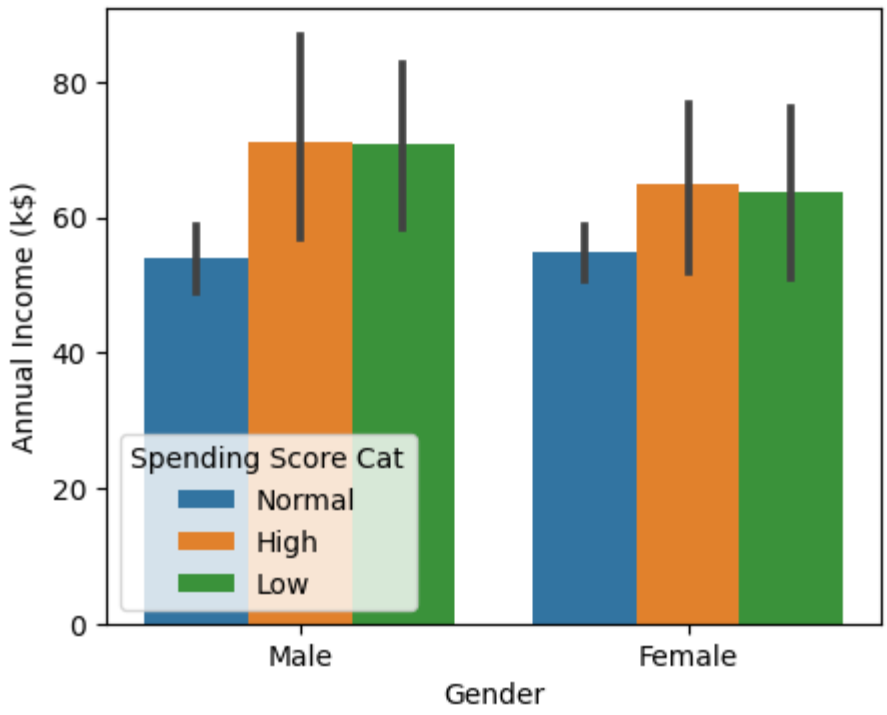
Về thuộc tính Spending Score, Phân tích thống kê mô tả cho thấy rằng thu nhập hàng năm của khách hàng có sự biến động lớn, từ mức thu nhập thấp nhất là 1k$ đến mức thu nhập cao nhất là 99k$. Sự đa dạng này cho thấy trung tâm thương mại phục vụ một phạm vi rộng khách hàng có thu nhập khác nhau Giá trị trung bình (mean) của thu nhập là 50.2k$, với độ lệch chuẩn (standard deviation) là khoảng 25.82k$. Điều này có thể ngụ ý rằng có sự biến động lớn trong thu nhập của khách hàng. Sự biến động này có thể tạo ra cơ hội cho trung tâm thương mại mở rộng hoặc tối ưu hóa dịch vụ và sản phẩm để phù hợp với nhu cầu của các phân khúc khách hàng có thu nhập khác nhau. Khoảng giữa 25% và 75% của dữ liệu thu nhập tập trung trong khoảng từ 34.75k$ đến 73k$, với phạm vi Interquartile (IQR) là 38.25k$. Điều này cho thấy rằng có một phần đáng kể số khách hàng có thu nhập trung bình đến cao. Sự lệch nhỏ về phía trái của độ nhọn (skewness) và chỉ số Kurtosis âm cho thấy phân phối của thu nhập có vẻ đối xứng và nhọn hơn so với phân phối chuẩn. Điều này có thể gợi ýến việc các khách hàng có thu nhập tập trung trong một khoảng hẹp, tuy nhiên, cũng có sự đa dạng đáng kể về thu nhập giữa các khách hàng. Tổng quát, dữ liệu này cung cấp thông tin quan trọng về tiềm năng tiêu dùng và mức độ đa dạng của khách hàng về thu nhập tại trung tâm thương mại. Việc hiểu rõ về tình hình kinh tế của khách hàng có thể giúp trung tâm thương mại tối ưu hóa chiến lược tiếp thị và sản phẩm, từ đó tăng cường doanh số và tạo lợi nhuận.



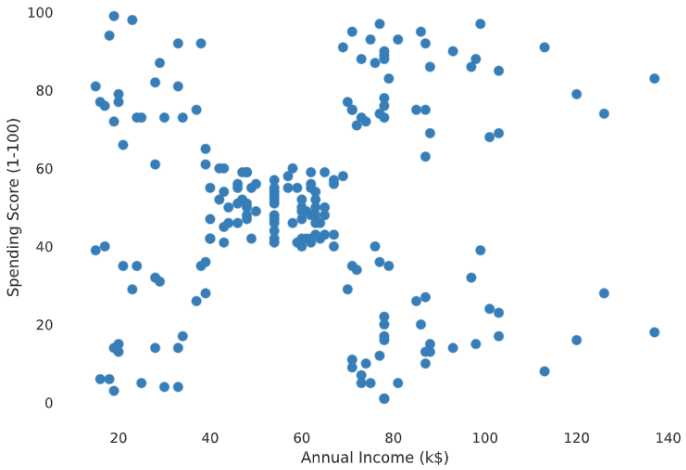
Hình 3.4 Thống kê cho Spending Core

Hình 3.5 Phân bố của dữ liệu Spending Core

Ngoài các phân tích trên chúng tôi cũng đối chiếu các mối quan hệ giữa các thuộc tính với nhau ở đây, ta thấy rằng chúng ta thấy có vài điểm khác biệt trong thu nhập hàng năm giữa các khách hàng nam và nữ, cũng như mối liên quan giữa Annual Income và Spending Score,....



Hình 3.6 Biểu đồ hiển thị mối quan hệ giữa giới tính (Gender), thu nhập hàng năm (Annual Income) và phân loại điểm chi tiêu (Spending Score Cat) từ dữ liệu được lưu trữ trong dataframe



Hình 3.7. Biểu đồ thể hiện sự tương quan giữa Spending Score (1-100) và Annual Icome (k$)

3.2 Áp dụng phương pháp K-Means

Việc phân vùng tập dữ liệu thành một số cụm được xác định trước đạt được bằng cách sử dụng thuật toán học không giám sát phổ biến được gọi là phân cụm K-mean. Mục đích của thuật toán là giảm thiểu tổng khoảng cách bình phương giữa các điểm dữ liệu và tâm cụm được chỉ định của chúng, còn được gọi là centroid. Các bước cơ bản của thuật toán k-means bao gồm chọn số cụm, khởi tạo ngẫu nhiên các centroid, gán điểm dữ liệu cho centroid gần nhất, tính toán lại centroid làm giá trị trung bình của tất cả các điểm dữ liệu được gán cho mỗi cụm và lặp lại cho đến khi hội tụ.

Phân cụm K-mean là một công cụ hữu ích để phân tích dữ liệu khám phá và nhận dạng mẫu, đồng thời nó có thể được áp dụng cho nhiều ứng dụng như phân khúc thị trường, xử lý hình ảnh và phát hiện sự bất thường. Tuy nhiên, điều quan trọng là phải lựa chọn cẩn thận số lượng cụm và đánh giá chất lượng của các cụm kết quả để tránh kết quả dưới mức tối ưu.

Và sau đây là kết quả của việc dùng thuật toán K – Means:

Hình 3.8 Các cụm dữ liệu sau khi dùng K – Means

3.3 Áp dụng thuật toán GMM

GMM là một mô hình xác suất biểu thị sự phân bố dữ liệu dưới dạng hỗn hợp của nhiều phân bố Gaussian. Trong GMM, mỗi cụm được mô hình hóa bằng phân phối Gaussian với ma trận trung bình và hiệp phương sai. Hàm mật độ xác suất của GMM được định nghĩa là tổng của các phân bố Gauss riêng lẻ được tính theo xác suất tương ứng của chúng.

Phân cụm GMM liên quan đến việc tìm các tham số tối ưu cho phân bố Gaussian phù hợp nhất với dữ liệu. Điều này thường được thực hiện bằng cách sử dụng thuật toán Tối đa hóa kỳ vọng (EM), đây là một thuật toán lặp luân phiên giữa việc tính toán xác suất dự kiến của từng điểm dữ liệu thuộc mỗi cụm ("E - step") và cập nhật các tham số của phân bố Gaussian dựa trên về các xác suất này ("M - step"). Thuật toán EM tiếp tục cho đến khi hội tụ, thường được định nghĩa là một thay đổi nhỏ trong hàm khả năng hoặc các tham số.

Một ưu điểm của GMM so với K-means là nó cho phép tạo ra các cụm có hình dạng và kích thước khác nhau, trong khi K-means giả định rằng các cụm có dạng hình cầu và có kích thước bằng nhau. Ngoài ra, GMM có thể xử lý các cụm chồng chéo và có thể gán cho mỗi điểm dữ liệu một xác suất thuộc về từng cụm, thay vì gán cứng như trong K-means.

Để thực hiện phân cụm GMM, số cụm (k) và phương thức khởi tạo cho các tham số trước tiên phải được tìm và xác định. Sau đó, điều chỉnh mô hình GMM cho dữ liệu bằng thuật toán EM và lấy các phép gán cụm cũng như các tham số cho từng phân bố Gaussian. Sau đó, sử dụng các phép gán cụm này để gắn nhãn các điểm dữ liệu mới và phân tích các đặc điểm của từng cụm.

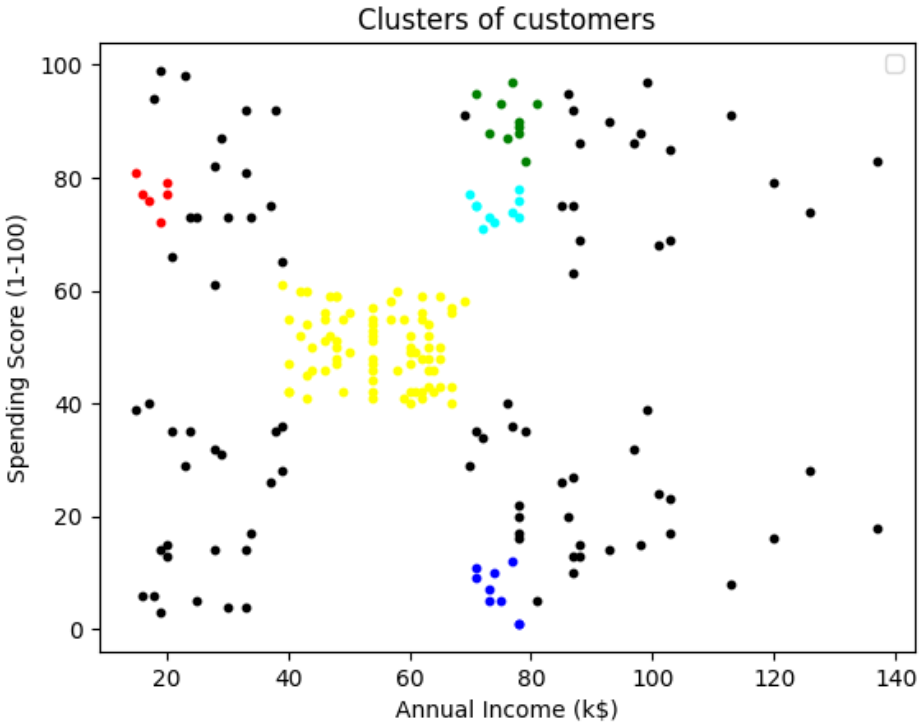
Và đây là kết quả của việc dùng GMM lên dữ liệu của chúng ta:

Hình 3.9 Các cụm dữ liệu sau khi dùng GMM

3.4 Áp dụng thuật toán DBSCAN

DBSCAN ((Density-Based Spatial Clustering of Applications with Noise) đã được sử dụng để xác định các cụm trong tập dữ liệu. DBSCAN là một thuật toán phân cụm dựa trên mật độ, nhóm các điểm dữ liệu gần nhau lại theo một thước đo khoảng cách được chỉ định. Các siêu tham số của thuật toán DBSCAN, cụ thể là số lượng mẫu tối thiểu (min_samples) và khoảng cách tối đa giữa hai điểm (eps), đã được điều chỉnh để thu được các giá trị tối ưu.

Và đây là kết quả của việc dùng DBSCAN lên dữ liệu của chúng ta:



Hình 3.10 Các cụm dữ liệu sau khi dùng DBSCAN

CHƯƠNG 4: PHƯƠNG PHÁP PHÁT TRIỂN HỌC SÂU TRONG PHÂN CỤM

4.1 Đôi nét về deep learning trong phân cụm

Mục tiêu chính của việc gom cụm là phân chia dữ liệu thành các nhóm các điểm dữ liệu tương tự. Việc phân chia tốt các điểm dữ liệu thành các cụm là cực kỳ quan trọng đối với nhiều ứng dụng trong phân tích dữ liệu và trực quan hóa dữ liệu. Tuy nhiên, hiệu suất của các phương pháp gom cụm hiện tại phụ thuộc rất nhiều vào dữ liệu đầu vào. Các bộ dữ liệu khác nhau thường đòi hỏi các độ đo tương đồng và kỹ thuật phân chia khác nhau. Do đó, việc giảm chiều dữ liệu và học biểu diễn đã được sử dụng rộng rãi cùng với việc gom cụm để ánh xạ dữ liệu đầu vào vào một không gian đặc trưng mà phân chia dễ dàng hơn. Bằng cách sử dụng mạng neural sâu (DNNs), có thể học các ánh xạ phi tuyến cho phép biến đổi dữ liệu thành các biểu diễn thân thiện với việc gom cụm hơn mà không cần trích xuất/chọn lọc đặc trưng thủ công.

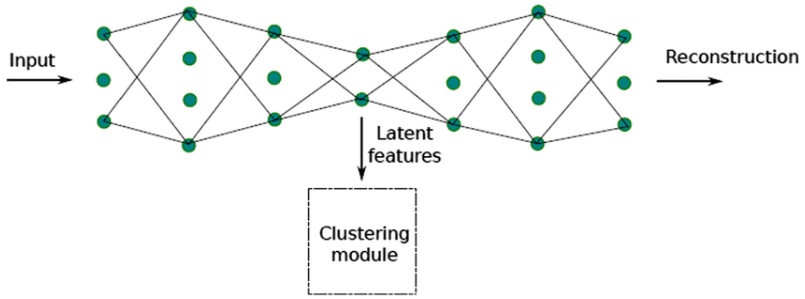
Các phương pháp gom cụ truyền thống như đã nói được sử dụng để xác định các mô hình dữ liệu. Hầu hết thời gian, những phương pháp này phù hợp khi phân tích dữ liệu không cần thiết lập sâu. Tuy nhiên, các kỹ thuật gom truyền thống không phù hợp cho các tập dữ liệu lớn và có chiều cao vì chúng không thể xác định các mô hình phức tạp. Ngược lại, gom sâu sử dụng mạng gom sâu, gom cụ truyền thống linh hoạt và gom sâu nhúng để thực hiện phân tích có chiều sâu trong các tập dữ liệu lớn. Do đó, gom sâu có thể xác định các mô hình phức tạp và ẩn trong dữ liệu hiệu quả hơn. Ngoài ra, hồi quy và phân loại cũng có thể được sử dụng cho tổ chức dữ liệu. Do đó, người ta thường bị nhầm lẫn giữa các kỹ thuật này với gom cụ. Tuy nhiên, so với gom cụ, hồi quy và phân loại là các kỹ thuật học có giám sát. Vì vậy, chúng ta cần cần gắn nhãn dữ liệu để sử dụng hồi quy và phân loại.

4.2 Deep clustering network (DCN)

Các phương pháp học hiện đại phổ biến xử lý việc giảm chiều dữ liệu (DR) và phân cụm một cách riêng lẻ (tức là theo thứ tự), nhưng nghiên cứu gần đây đã chỉ ra rằng tối ưu hóa cùng lúc hai nhiệm vụ này có thể cải thiện đáng kể hiệu suất cả hai. Cơ sở của thể loại sau là các mẫu dữ liệu được thu thập thông qua phép biến đổi tuyến tính của biểu diễn ẩn mà có thể dễ dàng phân cụm; nhưng trong thực tế, phép biến đổi từ không gian ẩn sang dữ liệu có thể phức tạp hơn. Trong công việc này, chúng tôi giả định rằng phép biến đổi này là một hàm không xác định và có thể phi tuyến. Để phục hồi biểu diễn ẩn "thân thiện với phân cụm" và phân cụm dữ liệu tốt hơn, chúng tôi đề xuất một phương pháp kết hợp DR và phân cụm K-means trong đó DR được thực hiện thông qua việc học một mạng neural sâu (DNN). Động lực của chúng tôi là giữ lại những ưu điểm của việc tối ưu hóa cùng lúc hai nhiệm vụ, đồng thời khai thác khả năng của mạng neural sâu để xấp xỉ bất kỳ hàm phi tuyến nào. Như vậy, phương pháp đề xuất có thể hoạt động tốt cho một loạt các mô hình

sinh dữ liệu. Để đạt được điều này, chúng tôi cẩn thận thiết kế cấu trúc DNN và tiêu chuẩn tối ưu hóa chung đi kèm, đồng thời đề xuất một thuật toán hiệu quả và có thể mở rộng để xử lý vấn đề tối ưu hóa được đề ra. Các thí nghiệm với các bộ dữ liệu thực tế khác nhau được sử dụng để thể hiện tính hiệu quả của phương pháp đề xuất.

Trong deep clustering networks (mạng gom cụ sâu), autoencoder được sử dụng để học cách biểu diễn dữ liệu một cách hiệu quả để thuật toán K-means có thể áp dụng được. Autoencoder là một loại thuật toán học máy không giám sát, nó giúp huấn luyện các mạng neural một cách hiệu quả và bỏ qua nhiễu trong dữ liệu. Trong trường hợp này, mạng gom cụ sâu tiền huấn luyện autoencoder và sử dụng nó để cực đại hóa việc tái tạo dữ liệu và mất mát K-means khi thay đổi phân cụm



Hình 4.1 Mạng phân cụm sâu được đề xuất (DCN).

4.3 Deep Adaptive Clustering (DAC)

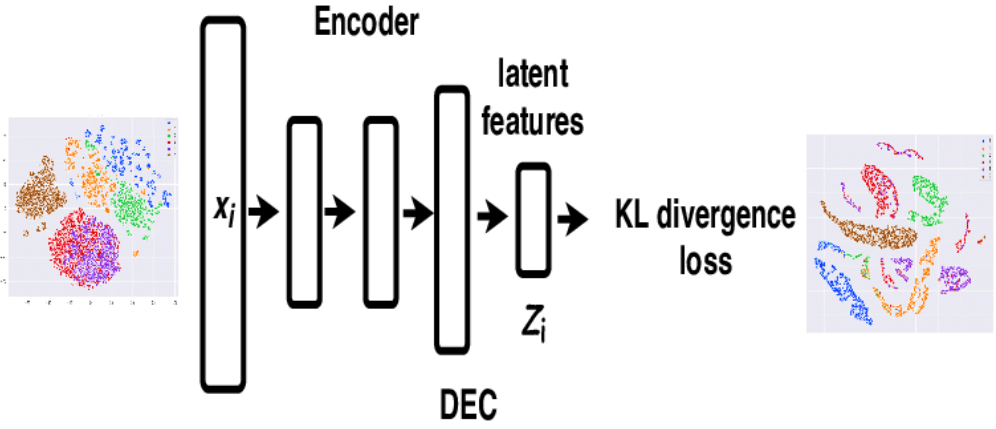
Deep adaptive clustering (gom cụ thích nghi sâu) là một phương pháp gom cụ sâu phổ biến khác, sử dụng một cấu trúc phân loại nhị phân theo cặp để tạo ra các cụ. Nó lấy hai điểm dữ liệu làm đầu vào và quyết định liệu chúng có thuộc cùng một cụ hay không. Dưới đó là một mạng neural với kích hoạt Softmax, nhận các điểm dữ liệu và tạo ra xác suất của một điểm dữ liệu thuộc về một cụ cụ thể. Nếu tích vô hướng của hai đầu ra từ cặp dữ liệu bằng 0, chúng thuộc về cùng một cụ. Nếu tích vô hướng là 1, cặp điểm dữ liệu thuộc về các cụ khác nhau.

Phương pháp này mang tính hiệu quả và linh hoạt, đặc biệt trong việc xử lý dữ liệu không được gán nhãn. Điều này cho phép mô hình tự động xác định cấu trúc của dữ liệu một cách linh hoạt, mà không cần sự can thiệp của con người để định nghĩa trước số lượng và đặc điểm của các cụ. Sự kết hợp giữa mạng neural với kích hoạt Softmax và cấu trúc phân loại nhị phân theo cặp tạo ra một cách tiếp cận mạnh mẽ cho bài toán gom cụ sâu, cho phép mô hình học được các biểu diễn dữ liệu và tự động tạo ra các cụ phù hợp.

4.4 Deep Embedded Clustering (DEC)

Deep Embedded Clustering (Phương pháp gom cụ sâu nhúng sâu) được coi là tiêu chuẩn để so sánh hiệu suất của các phương pháp gom cụ sâu trong học sâu. Nó sử dụng mạng neural sâu để học biểu diễn đặc trưng và phân công cụ một cách liên tục. Ngoài ra, nó tối ưu hóa các mục tiêu gom cụ bằng cách ánh xạ không gian dữ liệu vào không gian

đặc trưng chiều thấp hơn. Trong quá trình tiền huấn luyện DEC, các tham số mã hóa và giải mã được khởi tạo trong vài epochs với mất mát tái cấu trúc. Sau đó, mạng mã hóa được loại bỏ, và mạng giải mã được điều chỉnh tinh chỉnh bằng cách tối ưu hóa sự sai khác KL giữa phân công cụ mềm và phân phối phụ trợ. Tổng công đoạn, DEC là một quá trình tự huấn luyện gom cụ liên tục điều chỉnh biểu diễn dữ liệu trong khi thực hiện phân công cụ.



Hình 4.2 Mô hình cho DEC

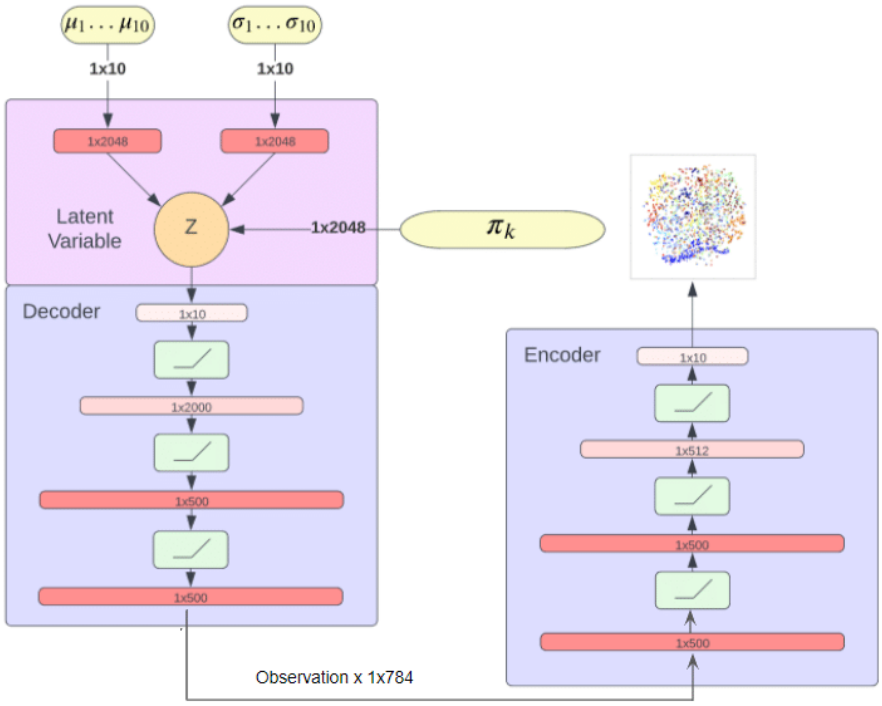
4.5 Information Maximizing Self-Augmented Training (IMSAT)

Tại các cụm được tạo ra bằng cách cân bằng số lượng điểm dữ liệu trong các cụm. Nó chỉ xem xét việc dùng regularization penalty trên các tham số mô hình để đảm bảo việc gán cụm. Information Maximizing Self-Augmented Training (huấn luyện tự tăng cường tối đa hóa thông tin), hay IMSAT, sửa đổi việc tối đa hóa thông tin bằng cách kết hợp nó với các kỹ thuật huấn luyện tự tăng cường mà phạt sự khác biệt về biểu diễn dữ liệu giữa các điểm dữ liệu gốc và điểm dữ liệu được tăng cường. Trong IMSAT, bạn có thể tuân theo biến đổi ngẫu nhiên hoặc tạo gian đoạn ảo để mở rộng dữ liệu. Nếu bạn chọn huấn luyện biến đổi ngẫu nhiên, một sự lệch ngẫu nhiên sẽ được thêm vào đầu vào từ một phân phối nhiễu đã định. Mặt khác, khi chọn huấn luyện gian đoạn ảo, sự lệch sẽ được gán sao cho mô hình không thể gán nó vào cùng một cụm. Ngoài việc phân cụm, phương pháp này cũng được sử dụng trong hash learning.

4.6 Variational Deep Embedding (VaDE)

Variational Deep Embedding (VaDE) là một quá trình tạo dữ liệu 4 bước sử dụng mạng nơ-ron sâu và mô hình Gaussian Mixture. Nó tổng quát hóa mô hình Variational autoencoder (VAE) bằng cách thay thế gaussian đơn bằng một mô hình Gaussian Mixture. Cuối cùng, nó làm cho VaDE phù hợp hơn cho việc gom cụm hơn là VAE bằng cách giảm thiểu tổn thất tái tạo. Đầu tiên, nó cho phép mô hình Gaussian Mixture chọn một cụm cho

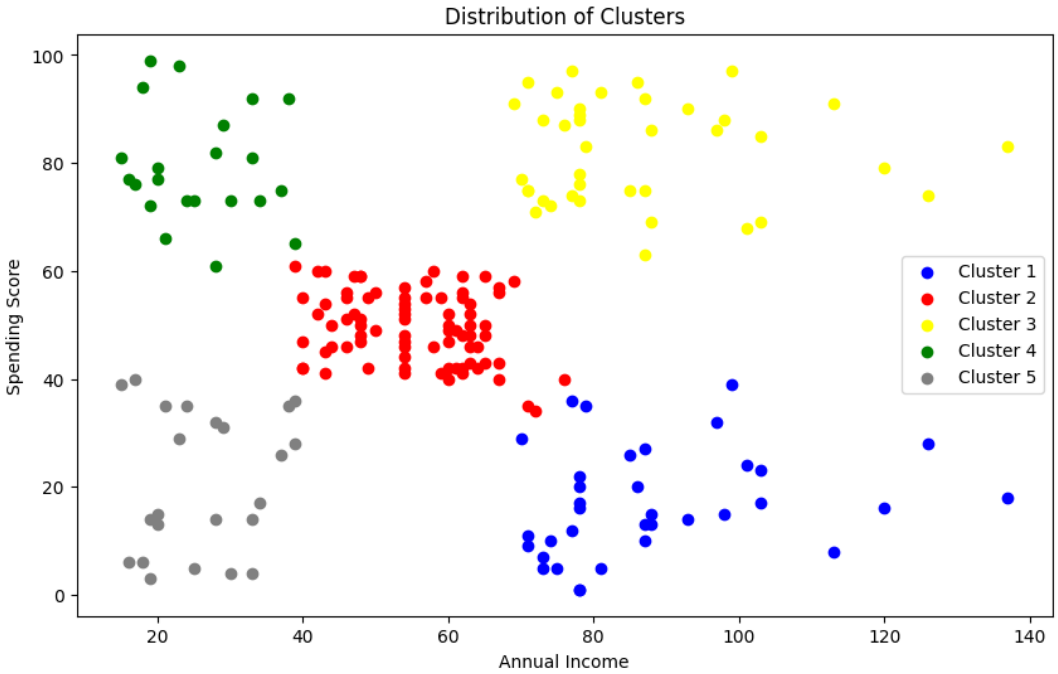
điểm dữ liệu. Sau đó, một nhúng tiềm ẩn được tạo dựa trên cụm được chọn. Sau đó, mạng nơ-ron sâu sẽ giải mã nhúng tiềm ẩn thành một quan sát. Cuối cùng, giới hạn dưới bằng chứng của VaDE được tối đa hóa bằng cách sử dụng mạng mã hóa.



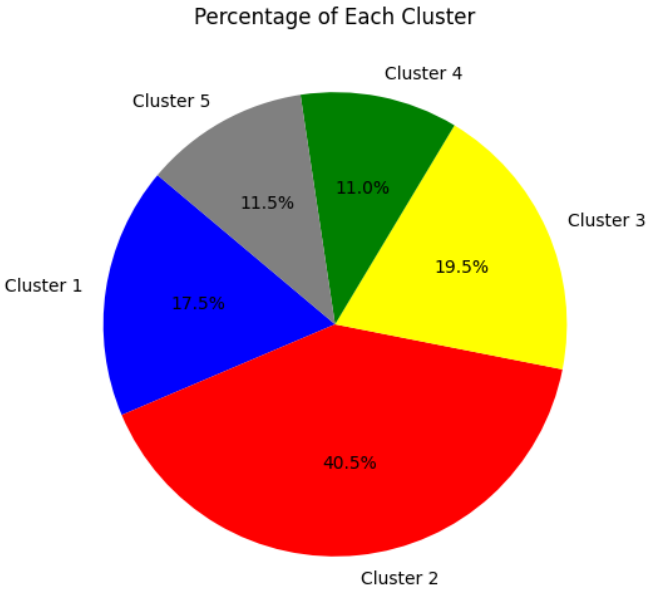
Hình 4.3 Mô hình cho VaDE

CHƯƠNG 5: KẾT QUẢ PHÂN CỤM

Kết quả phân cụm



Hình 5.1. Các cụm dữ liệu trên thuật toán phân cụm



Hình 5.2 Biểu đồ thống kê các cụm

Dựa trên biểu đồ "Percentage of Each Cluster", chúng ta có thể phân tích và diễn giải về 5 phân khúc khách hàng ứng với 5 cụm dữ liệu.

Cụm 1(17,5%): Khách hàng có thu nhập hàng năm cao nhưng chi tiêu hàng năm thấp đối với cụm này ta đưa ra chiến lược: Tăng cường chiến dịch tiếp thị để nâng cao nhận thức về sản phẩm và dịch vụ cao cấp hơn có sẵn. Tập trung vào việc tạo ra các gói sản phẩm và dịch vụ có giá trị cao để thu hút họ chi tiêu hơn. Cụm khách hàng này, tuy có thu nhập hàng năm cao nhưng lại chi tiêu hàng năm thấp, đòi hỏi một chiến lược tiếp thị đặc biệt để khuyến khích họ chi tiêu hơn. Tăng cường chiến dịch tiếp thị để tạo sự nhận thức về sản phẩm và dịch vụ cao cấp và tập trung vào việc tạo ra các gói sản phẩm và dịch vụ có giá trị cao có thể là một phương pháp hiệu quả. Đặc biệt, việc tôn trọng và nâng cao giá trị đầu tư từ phía họ, cung cấp thông tin giáo dục và tư vấn chất lượng cao, đồng thời tạo ra các gói sản phẩm và dịch vụ linh hoạt nhằm đáp ứng nhu cầu riêng biệt của họ có thể là các yếu tố quan trọng giúp kích thích họ chi tiêu nhiều hơn, đồng thời tạo sự hài lòng và tăng cường lòng trung thành của khách hàng.

Cụm 2(40,5%): Khách hàng có thu nhập hàng năm và chi tiêu hàng năm ở mức trung bình đối với cụm này ta đưa ra chiến lược: Tạo các chương trình khuyến mãi và giảm giá hấp dẫn dựa trên lịch sử mua sắm của họ. Tập trung vào việc nâng cao trải nghiệm mua sắm để khuyến khích họ chi tiêu nhiều hơn. Chiến lược tạo các chương trình khuyến mãi và giảm giá hấp dẫn dựa trên lịch sử mua sắm của khách hàng có thu nhập hàng năm và chi tiêu hàng năm ở mức trung bình là một cách tiếp cận hiệu quả. Bằng cách tập trung vào việc nâng cao trải nghiệm mua sắm của họ và tạo ra những ưu đãi hấp dẫn dựa trên thông tin lịch sử mua sắm, chúng ta có thể kích thích họ chi tiêu nhiều hơn. Việc đảm bảo rằng chương trình khuyến mãi và giảm giá được thiết kế để phản ánh sở thích và hành vi mua sắm cũng như tạo ra trải nghiệm mua sắm đáng nhớ có thể góp phần trong việc tạo sự hài lòng và tăng cường mức độ hài lòng của khách hàng.

Cụm 3(19,5%): Khách hàng có thu nhập hàng năm và chi tiêu hàng năm cao, chiến lược đưa ra: Cung cấp dịch vụ cao cấp và sản phẩm đắt tiền hơn để đáp ứng nhu cầu mua sắm của họ. Tập trung vào việc tạo ra trải nghiệm mua sắm sang trọng và đẳng cấp. Để phục vụ cụm khách hàng này, một chiến lược tập trung vào việc cung cấp dịch vụ cao cấp và sản phẩm đắt tiền hơn là một phương pháp tiếp cận hiệu quả. Việc tạo ra trải nghiệm mua sắm sang trọng và đẳng cấp sẽ tạo ra sự hài lòng và lòng trung thành từ phía khách hàng. Ngoài ra, việc tạo ra các sự kiện độc đáo và thú vị, kèm theo dịch vụ chăm sóc khách hàng tận tâm và chất lượng cao có thể tạo nên sự khác biệt đáng kể. Đồng thời, việc nghiên cứu và hiểu rõ sâu hơn về nhu cầu và sở thích của cụm khách hàng này cũng rất quan trọng để hiểu rõ hơn về những gì họ thực sự muốn và cần. Cách tiếp cận này có thể giúp tối ưu hóa kế hoạch tiếp thị và dịch vụ sản phẩm, từ đó tạo ra môi trường mua sắm lôi cuốn và hấp dẫn hơn

Cụm 4(11%): Khách hàng có thu nhập hàng năm thấp nhưng chi tiêu hàng năm cao. Chiến lược: Tập trung vào việc cung cấp các sản phẩm và dịch vụ có giá trị cao mặc

dù giá thành thấp hơn để duy trì mức chi tiêu hiện tại của họ và tạo sự trung thành. Tạo các chương trình giảm giá và khuyến mãi có thể hấp dẫn họ mua sắm hơn. Để tăng số lượng khách hàng trong cụm đối tượng này, chúng ta có thể xem xét triển khai chương trình khuyến mãi đặc biệt dành riêng cho những khách hàng có thu nhập thấp nhưng chi tiêu cao. Chương trình này có thể bao gồm việc cung cấp các ưu đãi đặc biệt, giảm giá hoặc gói sản phẩm dịch vụ có giá trị cao với mức giá ưu đãi đặc biệt. Điều này có thể tạo động lực và đồng thời thu hút được sự quan tâm của cụm khách hàng này, từ đó tăng cường số lượng khách hàng mua sắm. Ngoài ra, việc xây dựng chương trình khách hàng thân thiết cũng là một cách hiệu quả để tăng sự động viên và giữ chân khách hàng có thu nhập thấp nhưng chi tiêu cao. Bằng cách cung cấp các ưu đãi và quyền lợi đặc biệt dành riêng cho họ, chúng ta có thể tạo sự cam kết và lòng trung thành từ phía cụm khách hàng này, từ đó tăng cường số lượng khách hàng và doanh số bán hàng.

Cụm 5(11,5%): Khách hàng có thu nhập hàng năm và chi tiêu hàng năm thấp. Chiến lược: Tạo ra các chương trình khuyến mãi, giảm giá và sản phẩm giá rẻ để thu hút khách hàng này. Tập trung vào việc tối ưu hóa chi phí và tăng cường giá trị cho khách hàng. Để thu hút khách hàng có thu nhập hàng năm và chi tiêu hàng năm thấp, chúng ta có thể tạo ra một loạt các chương trình giảm giá và khuyến mãi hấp dẫn. Điều này có thể bao gồm việc tạo ra sản phẩm và dịch vụ với mức giá thấp hơn, cùng với việc cung cấp các ưu đãi đặc biệt cho cụm khách hàng này. Chúng ta cũng cần tập trung vào việc tối ưu hóa chi phí để đảm bảo rằng chúng ta có thể cung cấp giá trị cao nhất có thể cho khách hàng mà vẫn duy trì mức giá hấp dẫn. Việc cung cấp các sản phẩm và dịch vụ giá rẻ và hữu ích có thể tạo sự thu hút đến cụm khách hàng này và tăng cường giá trị cho họ. Việc áp dụng chiến lược này có thể giúp tối ưu hóa tác động đối với cụm khách hàng có thu nhập hàng năm và chi tiêu hàng năm thấp, từ đó tăng cường sự hấp dẫn và đồng thời tạo ra lợi ích lâu dài cho doanh nghiệp.

KẾT LUẬN

Trong nghiên cứu này, chúng tôi thảo luận và đánh giá các kỹ thuật phân đoạn khách hàng chính được sử dụng cho mục đích phân tích khách hàng. Chúng ta thấy rằng không có phương pháp hoàn hảo cho việc phân đoạn khách hàng vì kết quả của quá trình phân đoạn phụ thuộc vào nhiều yếu tố, chẳng hạn như màu sắc điểm ảnh, cấu trúc, độ sáng, sự tương đồng của hình ảnh, nội dung hình ảnh và lĩnh vực vấn đề. Do đó, không thể xem xét một phương pháp duy nhất cho tất cả loại hình ảnh, cũng như tất cả các phương pháp có thể hoạt động tốt cho một loại hình ảnh cụ thể. Vì vậy, việc sử dụng giải pháp phức hợp bao gồm nhiều phương pháp cho vấn đề phân đoạn khách hàng là cần thiết. Mặt khác chúng tôi cũng đảm bảo có giới thiệu hơn nhiều hơn các phương pháp trong deep learning nhằm thể hiện đảm bảo tính phát triển của việc giải quyết vấn đề trong tương lai.

TÀI LIỆU THAM KHẢO

[1] Lee, Y., Kwon, O., & Lee, Y. (2015). Customer segmentation using purchase data for an online retailer. Expert Systems with Applications, 42(1), 332-341. DOI: 10.1016/j.eswa.2014.08.0

[2] Kumar, A., Jain, A., Jain, S., & Jain, S. (2016). Comparative study of clustering algorithms for customer segmentation in e-commerce. In 2016 International Conference on Computing, Analytics and Security Trends (CAST) (pp. 300-305). IEEE. DOI: 10.1109/CAST.2016.79

[3] Xiang, Y., & Gong, Y. (2018). Online Shopping Behavior Analysis Based on K-means, GMM and DBSCAN Clustering Algorithm. In 2018 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 1122-1126). IEEE. DOI: 10.1109/CSCI46756.2018.00209

[4] Chen, X., Zuo, X., Wu, Z., & Liu, X. (2020). A Comparative Study of Customer Segmentation Methods Based on Online Shopping Behavior. In 2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) (pp. 791-795). IEEE. DOI: 10.1109/IEEM47687.2020.9378743

[5] Jadhav, M., & Sonawane, K. (2021). Credit card fraud detection using clustering techniques. In 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 1-6). IEEE. DOI: 10.1109/CONFLUENCE51715.2021.9461624

[6] Yeh, Y.-L., & Huang, C.-C. (2018). Customer segmentation of bicycle-sharing users based on their usage behavior. Sustainability, 10(5), 1579. DOI: 10.3390/su10051579

[7] Dolničar, S. & Leisch, F. (2004) Segmenting markets by bagged clustering. Australasian Marketing Journal, 12, 2, pp. 51–65.

[8] Decker, R., Wagner, R. & Scholz, S.W. (2005) Growing clustering algorithms in market segmentation: defining target groups and related marketing communication. In H.-H. Bock, W. Gaul & M. Vicki (eds) Data Analysis, Classification and the Forward Search. Berlin: Springer, pp. 23–30.

[9] Aldenderfer, M.S. & Blashfield, R.K. (1984) Cluster Analysis. Beverly Hills: Sage Publications.

[10]Prateek Majumder, "K-Means clustering with Mall Customer Segmentation Data | Full Detailed Code and Explanation", 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/05/k-means-clustering-with-mall-customer-segmentation-data-full-detailed-code-and-explanation/

[11]Dhiraj Kumar, “Implementing Customer Segmentation Using Machine Learning [Beginners Guide]”, 2023, [Online]. Available: https://neptune.ai/blog/customer-segmentation-using-machine-learning

[12]Pham Dinh Khanh, “16. Gaussian Mixture Model”,” 16.1. Ước lượng MLE cho phân phối Gaussian đa chiều”,2021, [Online]. Available: https://phamdinhkhanh.github.io/deepai-book/ch_ml/index_GMM.html
[13]Pham Dinh Khanh ,“13.1. Các bước của thuật toán k-Means Clustering”, [Online]. Available: https://phamdinhkhanh.github.io/deepai-book/ch_ml/KMeans.html
[14]Pham Dinh Khanh ,“ 15. DBSCAN”, [Online]. Available: https://phamdinhkhanh.github.io/deepai-book/ch_ml/index_DBSCAN.html
[15]Lightrun,” 5 Approaches to Deep Learning Clustering You Really Need to Know”, [Online]. Available: https://lightrun.com/approaches-to-deep-learning-clustering/?zarsrc=30&utm_source=zalo&utm_medium=zalo&utm_campaign=zalo&fbclid=IwAR2Ua8RAblIYwDweF94QQmlr9iKdhdYEkJzBywEGh04Ct7VX-YlBl5kr54w