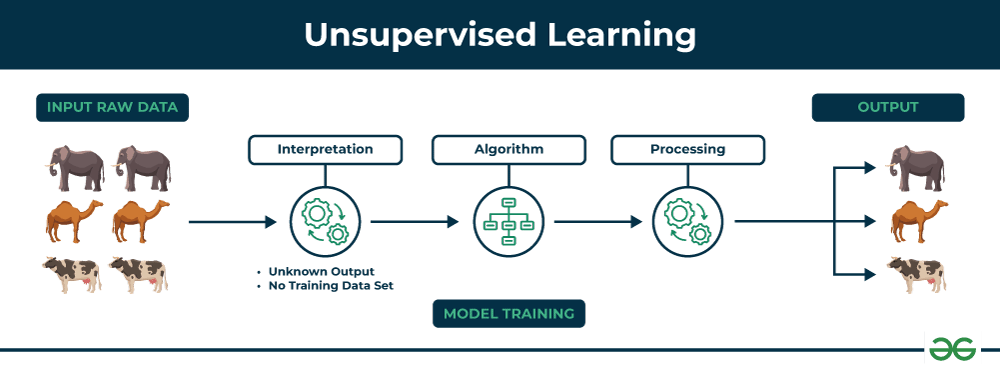
UNSUPERVISED LEARNING

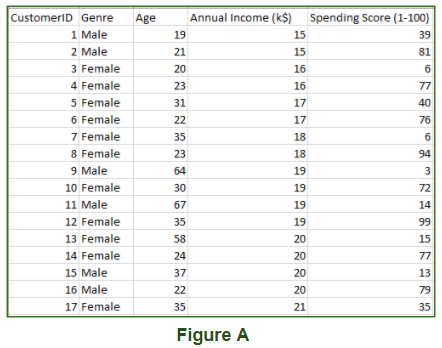
1. Khái niệm unsupervised learning:

* Unsupervised learning là một phân nhánh của học máy, xử lý dữ liệu không có nhãn, khác với học có giám sát mà dữ liệu có nhãn cụ thể.
* Trong unsupervised learning, các thuật toán được thiết kế để tự mình tìm hiểu mẫu và mối quan hệ từ dữ liệu mà không cần hướng dẫn hoặc chỉ dẫn rõ ràng từ con người.
* Các thuật toán unsupervised learning không nhận bất kỳ thông tin đầu ra cụ thể nào từ người dùng mà chỉ phân tích cấu trúc dữ liệu để tìm ra các nhóm hoặc mẫu ẩn.
* Phương pháp này thường được sử dụng trong phân tích khám phá dữ liệu, giúp hiểu rõ cấu trúc và đặc điểm cơ bản của dữ liệu.



1. Nguyên lý hoạt động của unsupervised learning

* **Khám phá cấu trúc ẩn:** Tự động phát hiện mối quan hệ và mẫu trong dữ liệu không được gắn nhãn.
* **Phân cụm:** Nhóm dữ liệu thành các cụm dựa trên sự tương đồng, giúp phân loại dữ liệu một cách có ý nghĩa.
* **Giảm chiều:** Áp dụng các kỹ thuật như phân tích thành phần chính (PCA) để giảm số lượng biến, giữ lại thông tin quan trọng nhất.
* **Không sử dụng nhãn dữ liệu:** Làm việc trực tiếp với dữ liệu thô, không cần dữ liệu đã được phân loại hoặc gắn nhãn trước.
* **Phát hiện mẫu:** Tìm các mẫu hoặc xu hướng thường xuyên xuất hiện trong dữ liệu để hiểu rõ hơn về cấu trúc dữ liệu.
* **Tự học (Self-organizing maps):** Sử dụng mạng tự tổ chức để phân loại và đại diện dữ liệu theo cách mà các mẫu tương tự nhau được nhóm gần nhau.
* **Khám phá độ tương quan:** Xác định và phân tích mối liên hệ giữa các biến để hiểu biết sâu hơn về dữ liệu.
* Ta có ví dụ sau:
* Tập dữ liệu trong Hình A là dữ liệu của Mall chứa thông tin về các khách hàng đăng ký chúng. Sau khi đăng ký, họ sẽ được cung cấp thẻ thành viên và trung tâm mua sắm có thông tin đầy đủ về khách hàng và mỗi lần mua hàng của họ. Giờ đây, bằng cách sử dụng dữ liệu này và các kỹ thuật học tập không giám sát, trung tâm mua sắm có thể dễ dàng nhóm khách hàng dựa trên các thông số mà chúng tôi đang cung cấp.



* Đầu vào của các mô hình unsupervised learning như sau:
* **Dữ liệu phi cấu trúc:** Có thể chứa dữ liệu nhiễu (vô nghĩa), thiếu giá trị hoặc dữ liệu không xác định.
* **Dữ liệu chưa được gắn nhãn :** Dữ liệu chỉ chứa giá trị cho tham số đầu vào, không có giá trị mục tiêu (đầu ra). Dễ dàng thu thập so với nhãn được dán nhãn trong phương pháp Được giám sát.

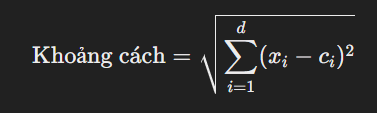
1. Các thuật toán Unsupervised Learning
   1. **Clustering (Phân cụm)**

* Phân cụm trong học máy không giám sát là quá trình phân loại dữ liệu không có nhãn vào các nhóm dựa trên sự giống nhau giữa chúng. Dưới đây là các ý chính liên quan đến phân cụm:
* Mục tiêu: Xác định các mẫu và mối quan hệ không rõ ràng trong dữ liệu mà không cần biết trước ý nghĩa của dữ liệu.
* Ứng dụng: Phân loại dữ liệu dựa trên các mẫu như điểm tương đồng hoặc khác biệt mà thuật toán nhận thấy.
* Dữ liệu đầu vào: Thuật toán xử lý các đối tượng dữ liệu thô, không được phân loại trước, và nhóm chúng dựa trên các tham số đầu vào mà không có chỉ định đầu ra.
* Các thuật toán phân cụm phổ biến bao gồm:
* K-means Clustering: Phân chia dữ liệu thành K cụm dựa trên khoảng cách đến trung tâm cụm.
* Phân cụm theo cấp bậc: Xây dựng cấu trúc phân cấp của các cụm thông qua liên kết liên tục giữa các đối tượng hoặc nhóm đối tượng.
* DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Phân cụm dựa trên mật độ, xác định các cụm dựa trên vùng lân cận có mật độ cao.
* Phân cụm dịch chuyển trung bình (Mean Shift Clustering): Tìm cụm dựa trên tìm kiếm chế độ, điều chỉnh vị trí các cụm theo mật độ dữ liệu.
* Phân cụm phổ (Spectral Clustering): Sử dụng lý thuyết đồ thị phổ để nhóm các đối tượng dựa vào cấu trúc và liên kết của chúng trong không gian đa chiều.

Thông qua các phương pháp này, phân cụm giúp khai thác thông tin từ dữ liệu lớn hoặc phức tạp mà không cần đến sự giám sát trực tiếp, cho phép tìm hiểu sâu hơn về cấu trúc và đặc điểm ẩn của dữ liệu.

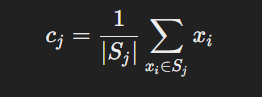
1. **Thuật toán K-means Clustering**

* Thuật toán K-means là một trong những thuật toán phổ biến nhất trong lĩnh vực học không giám sát (unsupervised learning) được sử dụng rộng rãi để thực hiện phân cụm dữ liệu (clustering). Mục tiêu của thuật toán này là phân chia **n** quan sát thành **k** nhóm (hay cụm) sao cho mỗi quan sát thuộc về cụm với trung tâm gần nhất (trung tâm cụm thường được gọi là centroid).
* Công thức của thuật toán K-means
* **Khởi tạo**: Chọn ngẫu nhiên **K** điểm dữ liệu ban đầu làm các centroid.
* **Gán nhãn cụm**: Mỗi điểm dữ liệu được gán vào cụm có centroid gần nhất. Khoảng cách thường được tính bằng khoảng cách Euclidean, được tính bằng công thức:



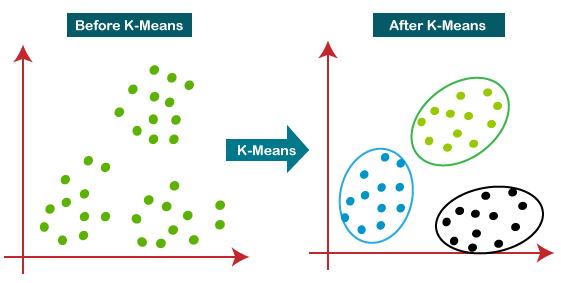
trong đó xi là giá trị của điểm dữ liệu trên chiều i, ci là giá trị của centroid trên chiều i, và d là số chiều của không gian dữ liệu.

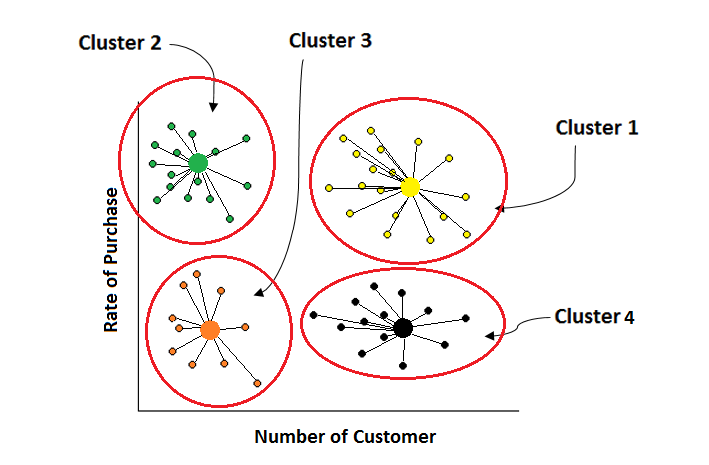
* **Cập nhật centroid**: Sau khi tất cả các điểm đã được gán vào các cụm, centroid mới của mỗi cụm được tính toán lại bằng cách lấy trung bình cộng các điểm trong cụm đó:



trong đó *Sj*​ là tập hợp các điểm dữ liệu trong cụm *j* và |*Sj*​| là số lượng điểm trong cụm đó.

* **Lặp lại:** Quá trình gán nhãn cụm và cập nhật centroid được lặp lại cho đến khi đạt được điều kiện dừng (ví dụ như các centroid không thay đổi nữa giữa các lần lặp, hoặc đã đạt số lần lặp tối đa, hoặc sự thay đổi trong tổng khoảng cách giữa các điểm dữ liệu và centroid của chúng dưới một ngưỡng nhất định).
* Hình ảnh về thuật toán K-means:





1. **Thuật toán phân cụm theo cấp bậc**

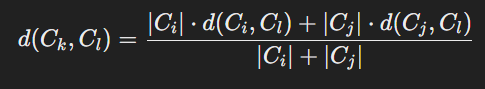
* Trong lĩnh vực học không giám sát (unsupervised learning), thuật toán phân cụm theo cấp bậc (hierarchical clustering) được sử dụng rộng rãi để phân chia dữ liệu thành các cụm dựa trên độ tương tự của chúng. Có hai hướng tiếp cận chính: Agglomerative (tổng hợp) và Divisive (chia tách). Trong bối cảnh học không giám sát, mục tiêu là tổ chức dữ liệu thành các cấu trúc có ý nghĩa mà không có bất kỳ nhãn nào được cung cấp trước.
* **Thuật toán Agglomerative Clustering**
* Bước 1: Khởi tạo
* Bắt đầu bằng cách coi mỗi điểm dữ liệu là một cụm riêng lẻ. Vì vậy, nếu có N điểm dữ liệu, ta có N cụm.
* Bước 2: Tính toán khoảng cách
* Tính toán khoảng cách giữa các cụm bằng cách sử dụng một số tiêu chí như khoảng cách Euclidean, khoảng cách Manhattan, hoặc bất kỳ phương pháp nào khác phù hợp với dữ liệu và yêu cầu phân tích.
* Bước 3: Gộp cụm
* Tìm cặp cụm có khoảng cách thấp nhất và gộp chúng lại với nhau, tạo thành một cụm mới.
* Bước 4: Cập nhật khoảng cách
* Khi hai cụm *Ci*​ và *Cj*​ được gộp thành một cụm mới *Ck*​, ta cần cập nhật khoảng cách giữa *Ck*​ và tất cả các cụm còn lại. Có nhiều phương pháp để cập nhật khoảng cách này, bao gồm:
* Single Linkage: Khoảng cách giữa hai cụm là khoảng cách ngắn nhất giữa bất kỳ thành viên nào của cụm này đến bất kỳ thành viên nào của cụm kia.



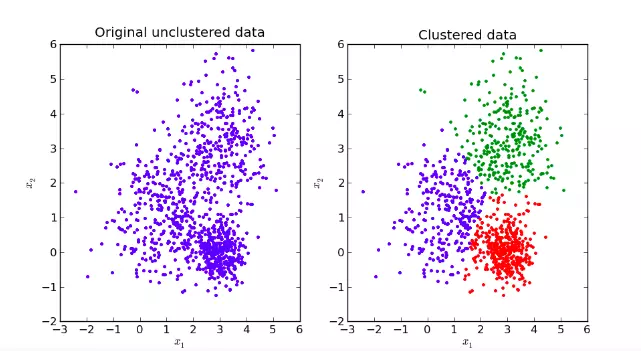
* Complete Linkage: Khoảng cách giữa hai cụm là khoảng cách lớn nhất giữa bất kỳ thành viên nào của cụm này đến bất kỳ thành viên nào của cụm kia.

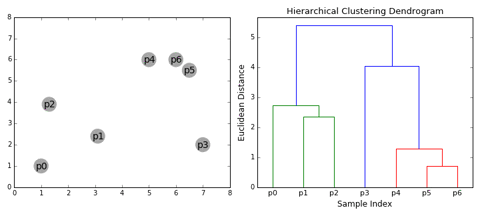


* Average Linkage: Khoảng cách giữa hai cụm là trung bình cộng của khoảng cách giữa mọi cặp thành viên của hai cụm.



* Bước 5: Lặp lại
* Lặp lại quá trình từ Bước 3 cho đến khi tất cả dữ liệu được gộp vào một cụm duy nhất hoặc đạt đến số lượng cụm mong muốn.
* Bước 6: Tạo Dendrogram
* Dendrogram là một biểu đồ biểu diễn quá trình gộp các cụm và giúp xác định số lượng cụm phù hợp bằng cách cắt dendrogram ở mức độ phù hợp.
* Hình ảnh về thuật toán phân cụm cấp bậc





* 1. **Association Rule Learning**
* Học quy tắc kết hợp là một kỹ thuật trong học máy không giám sát dùng để khám phá mối liên kết giữa các tham số trong tập dữ liệu lớn. Dưới đây là những điểm chính về kỹ thuật này:
* Mục đích: Khám phá và hiểu các mối quan hệ hữu ích giữa các sản phẩm trong các tình huống như phân tích giỏ hàng mua sắm.
* Ứng dụng: Giúp các cửa hàng và doanh nghiệp phân tích mối quan hệ giữa các sản phẩm dựa trên hành vi mua sắm của khách hàng, như việc mua sữa thường đi kèm với mua bánh mì, trứng hoặc bơ.
* Tăng doanh số: Các mô hình này được sử dụng để lên kế hoạch cho các chương trình khuyến mãi và ưu đãi, tăng doanh số bán hàng.
* Các thuật toán phổ biến cho khai thác quy tắc kết hợp bao gồm:
* Thuật toán Apriori: Phương pháp cổ điển để quy nạp quy tắc, tìm các tập hợp mục thường xuyên và quy tắc từ đó.
* Thuật toán tăng trưởng FP (FP-Growth): Giải pháp thay thế hiệu quả hơn cho Apriori không yêu cầu tạo ra các tập con lặp đi lặp lại.
* Thuật toán Eclat: Khai thác tập mục đóng để tìm quy tắc kết hợp một cách hiệu quả.
* Thuật toán dựa trên cây hiệu quả: Dùng cho tập dữ liệu lớn, cải thiện khả năng mở rộng và hiệu suất xử lý.

Những thuật toán này giúp phát hiện các mối quan hệ không rõ ràng và phức tạp giữa các mục, hỗ trợ quyết định trong kinh doanh và tiếp thị.

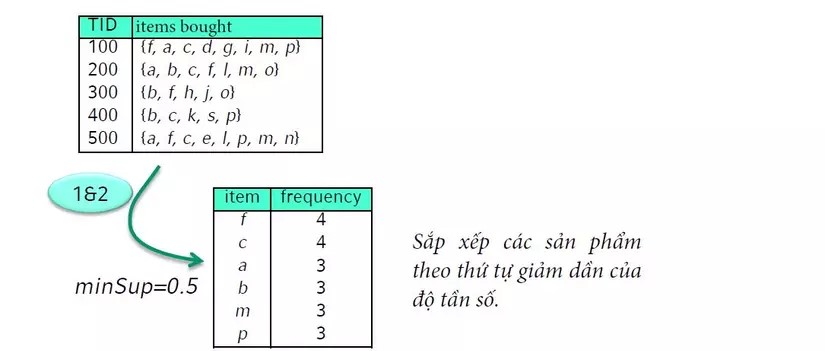
* **Thuật toán tăng trưởng FP (FP-Growth):**
* Thuật toán FP-Growth (Frequent Pattern Growth) thường được sử dụng trong khai phá quy tắc kết hợp, đặc biệt là trong các tình huống cần phát hiện các mẫu thường xuyên (frequent patterns) trong một tập dữ liệu lớn. Thuật toán này được phát triển để giải quyết các vấn đề liên quan đến hiệu suất của thuật toán Apriori, thuật toán cổ điển trong cùng lĩnh vực. FP-Growth tập trung vào việc giảm số lần truy cập cơ sở dữ liệu và không sử dụng phương pháp sinh các ứng viên, điều làm chậm quá trình của Apriori.
* FP-Growth hoạt động theo hai bước chính:

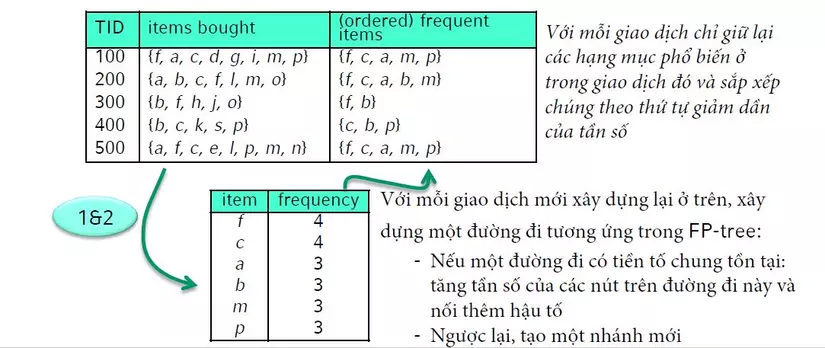
1. Xây dựng cây FP-Tree: Đây là bước xây dựng một cây để lưu trữ thông tin về các mẫu thường xuyên. Cây FP (Frequent Pattern Tree) là một cấu trúc dữ liệu dạng cây, nơi mỗi nút đại diện cho một mục hàng và mỗi đường đi từ nút gốc đến một nút bất kỳ biểu diễn một tập phổ biến của các mục hàng.

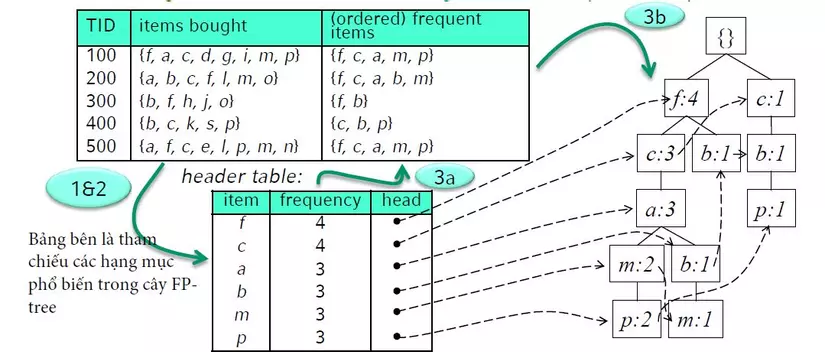
* Bước 1: Đếm tần suất xuất hiện của mỗi mục hàng trong cơ sở dữ liệu và sắp xếp chúng theo thứ tự giảm dần của tần suất xuất hiện.
* Bước 2: Xây dựng cây FP bằng cách thêm các giao dịch vào cây. Mỗi giao dịch được sắp xếp theo thứ tự tần suất và được thêm vào cây. Nếu một đường đi của giao dịch đã tồn tại, tần số của nó được tăng lên; nếu không, một đường mới sẽ được tạo.

1. Tìm kiếm các mẫu thường xuyên: Sau khi cây được xây dựng, thuật toán sẽ tìm kiếm các mẫu thường xuyên từ cây FP.

* Bước 1: Bắt đầu từ mỗi mục hàng (nút lá) và truy vết ngược lại đến gốc để tạo ra một tập con, được gọi là "điều kiện mẫu cơ sở".
* Bước 2: Sử dụng tập con này để xây dựng một cây FP-Tree điều kiện. Quá trình này lặp lại cho đến khi không còn mẫu nào có thể được tạo ra.
* Thuật toán FP-Growth có hiệu quả cao do nó không cần phải sinh các tập ứng viên mà trực tiếp xây dựng các mẫu thường xuyên từ dữ liệu, giảm đáng kể số lượng truy cập cơ sở dữ liệu so với Apriori. Thuật toán này đặc biệt hiệu quả khi xử lý các cơ sở dữ liệu lớn với số lượng mục và giao dịch lớn.
* Cần lưu ý rằng FP-Growth là một phần của học máy không giám sát (unsupervised learning) trong bối cảnh phân tích mẫu thường xuyên không cần đến thông tin nhãn của dữ liệu, mà tập trung vào việc phát hiện các cấu trúc và mối quan hệ tự nhiên trong dữ liệu.
* Một số hình ảnh về thuật toán:







* 1. **Dimensionality Reduction**
* Giảm kích thước là quá trình giảm số lượng tính năng hoặc đối tượng trong tập dữ liệu, đồng thời cố gắng lưu giữ càng nhiều thông tin càng tốt. Kỹ thuật này có tác dụng cải thiện hiệu suất của các thuật toán học máy và hỗ trợ trong việc trực quan hóa dữ liệu.
* Dưới đây là các thuật toán giảm kích thước phổ biến:
* Phân tích thành phần chính (PCA): Sử dụng chuyển đổi tuyến tính để giảm số chiều dữ liệu, tập trung vào việc lưu giữ thông tin quan trọng nhất qua các thành phần chính.
* Phân tích phân biệt tuyến tính (LDA): Giảm kích thước dữ liệu trong khi tối ưu hóa sự phân biệt giữa các lớp khác nhau, làm nổi bật các đặc điểm phân biệt.
* Hệ số ma trận không âm (NMF): Phân tách dữ liệu thành các thành phần không âm, giúp dễ dàng giải thích hơn, thường được ứng dụng trong phân tích văn bản và xử lý tín hiệu.
* Nhúng tuyến tính cục bộ (LLE): Bảo toàn cấu trúc hình học cục bộ của dữ liệu khi giảm kích thước, phù hợp cho dữ liệu có cấu trúc phức tạp.
* Isomap: Giảm kích thước bằng cách bảo toàn khoảng cách địa lý toàn cầu, thích hợp khi cấu trúc dữ liệu phức tạp và mang tính không tuyến tính.
* Thuật toán phân tích thành phần chính(PCA)

Phân tích thành phần chính (PCA - Principal Component Analysis) là một kỹ thuật quan trọng trong học máy không giám sát (unsupervised learning). PCA được sử dụng để giảm số chiều của dữ liệu, trong khi vẫn giữ lại những thông tin quan trọng nhất. Đây là cách làm phổ biến để khám phá và phân tích dữ liệu, và thường được dùng để trực quan hóa dữ liệu nhiều chiều. Dưới đây là công thức và bước thực hiện cơ bản của PCA:

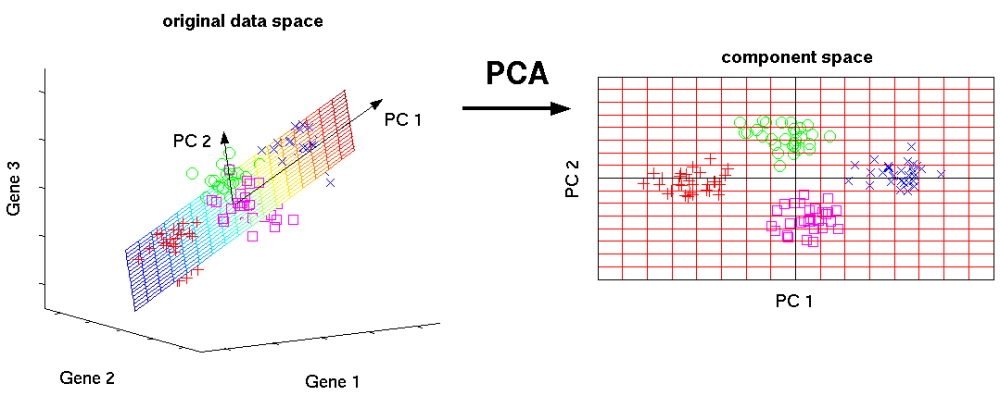
* Bước 1: Chuẩn bị dữ liệu
* Cho X là ma trận dữ liệu n×d với n mẫu và d thuộc tính.
* Chuẩn hóa dữ liệu: Trừ đi giá trị trung bình của mỗi thuộc tính từ dữ liệu tương ứng để dữ liệu có trung bình là 0.
* Bước 2: Tính ma trận hiệp phương sai
* Tính ma trận hiệp phương sai C của X:

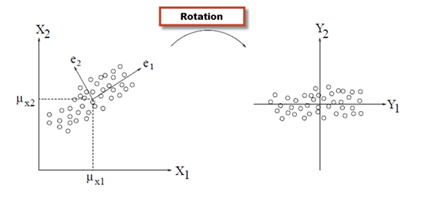


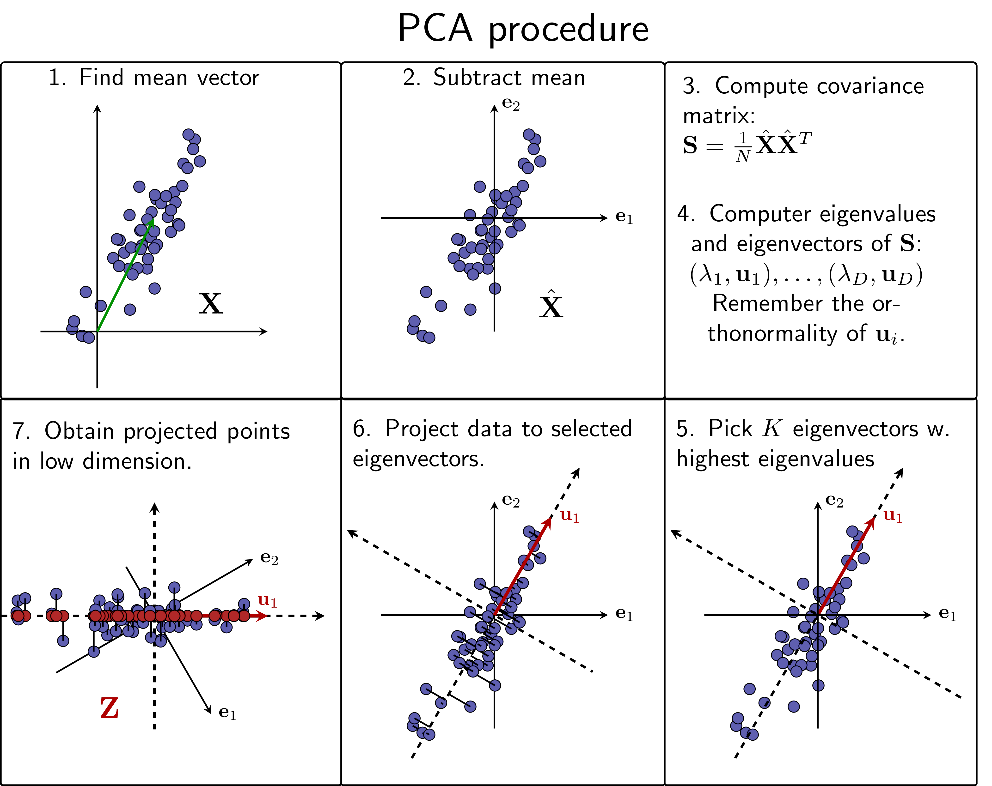
* Lưu ý rằng XTX là ma trận hiệp phương sai khi các dòng của X đã được chuẩn hóa (có trung bình là 0).
* Bước 3: Tính trị riêng và vectơ riêng
* Tìm các trị riêng λ và vectơ riêng v của ma trận hiệp phương sai C. Các vectơ riêng này sẽ là các thành phần chính.
* Sắp xếp các trị riêng theo thứ tự giảm dần và chọn k vectơ riêng tương ứng với k trị riêng lớn nhất để làm thành phần chính
* Bước 4: Chuyển đổi dữ liệu sang không gian mới
* Chuyển đổi dữ liệu ban đầu X sang không gian có số chiều nhỏ hơn bằng cách nhân X với k vectơ riêng đã chọn:

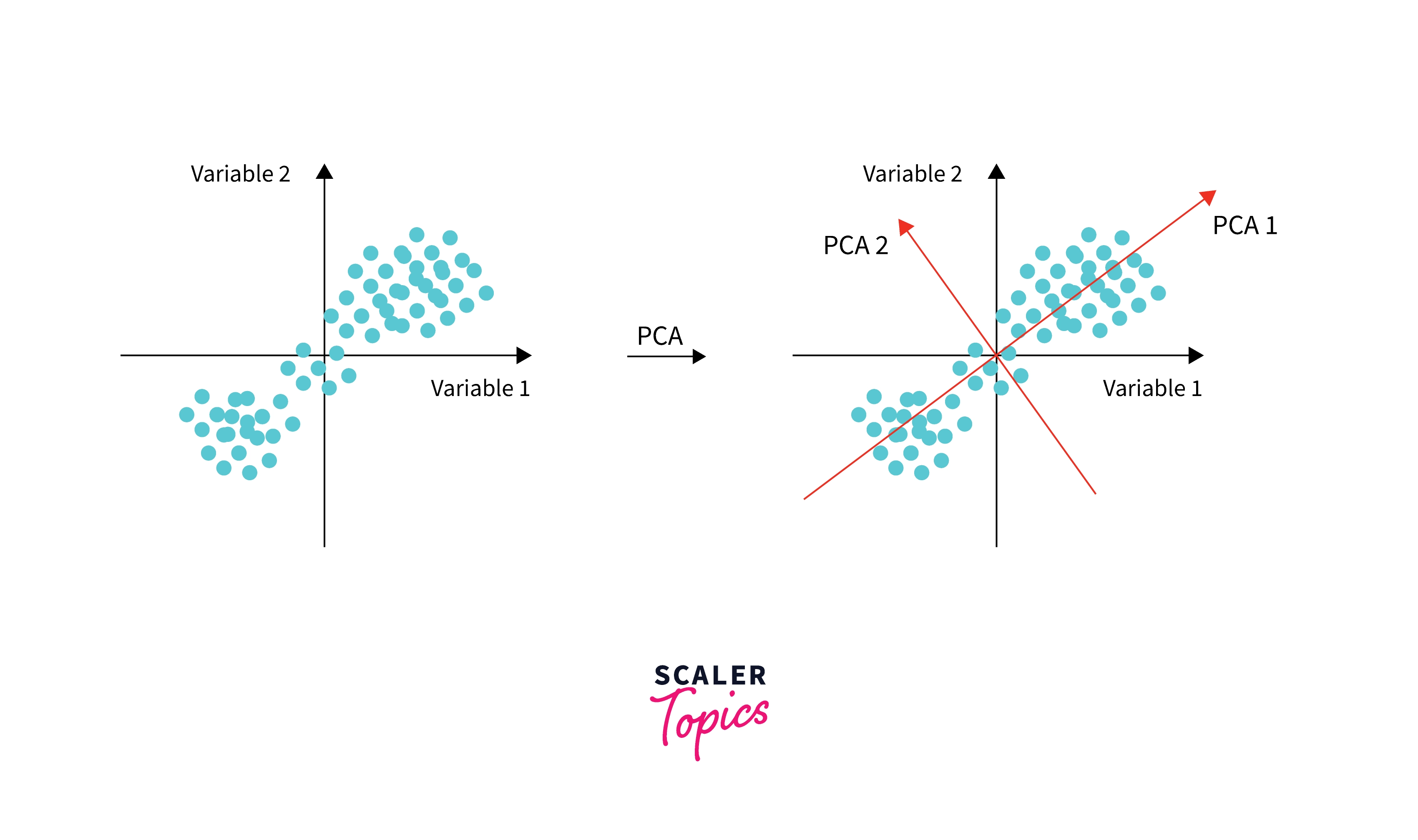
*Y = XVk*

* Trong đó Vk là ma trận chứa k vecto riêng của các cột.
* Hình ảnh về thuật toán PCA









1. Những thách thức của unsupervised learning

* **Đánh giá:**  Việc đánh giá hiệu suất của các thuật toán unsupervised learning là khó khăn nếu không có nhãn hoặc danh mục được xác định trước.
* **Khả năng diễn giải:**  Việc hiểu quá trình ra quyết định của các mô hình học tập không giám sát thường là một thách thức.
* **Trang bị quá mức:**  Các thuật toán unsupervised learning có thể quá khớp với tập dữ liệu cụ thể được sử dụng để đào tạo, hạn chế khả năng khái quát hóa dữ liệu mới của chúng.
* **Chất lượng dữ liệu:**  Thuật toán unsupervised learning rất nhạy cảm với chất lượng của dữ liệu đầu vào. Dữ liệu nhiễu hoặc không đầy đủ có thể dẫn đến kết quả sai lệch hoặc không chính xác.
* **Độ phức tạp tính toán:**  Một số thuật toán unsupervised learning, đặc biệt là các thuật toán xử lý dữ liệu nhiều chiều hoặc bộ dữ liệu lớn, có thể tốn kém về mặt tính toán.

1. Ưu điểm và nhược điểm của Unsupervised learning
   1. Ưu điểm

* **Không yêu cầu dữ liệu được gắn nhãn:**  Không giống như học có giám sát, học không giám sát không yêu cầu dữ liệu được gắn nhãn, việc thu thập này có thể tốn kém và mất thời gian.
* **Có thể khám phá các mẫu ẩn:**  Thuật toán học không giám sát có thể xác định các mẫu và mối quan hệ trong dữ liệu mà con người có thể không rõ ràng.
* **Có thể được sử dụng cho nhiều nhiệm vụ khác nhau:**  Học không giám sát có thể được sử dụng cho nhiều nhiệm vụ khác nhau, chẳng hạn như phân cụm, giảm kích thước và phát hiện bất thường.
* **Có thể được sử dụng để khám phá dữ liệu mới:**  Học tập không giám sát có thể được sử dụng để khám phá dữ liệu mới và thu được những hiểu biết sâu sắc mà các phương pháp khác có thể không thực hiện được.
  1. Nhược điểm
* **Khó đánh giá:**  Có thể khó đánh giá hiệu suất của các thuật toán học không giám sát vì không có nhãn hoặc danh mục được xác định trước để so sánh kết quả.
* **Có thể khó diễn giải:**  Có thể khó hiểu được quá trình ra quyết định của các mô hình học tập không giám sát.
* **Có thể nhạy cảm với chất lượng của dữ liệu:**  Thuật toán học không giám sát có thể nhạy cảm với chất lượng của dữ liệu đầu vào. Dữ liệu nhiễu hoặc không đầy đủ có thể dẫn đến kết quả sai lệch hoặc không chính xác.
* **Có thể tốn kém về mặt tính toán:**  Một số thuật toán học không giám sát, đặc biệt là các thuật toán xử lý dữ liệu nhiều chiều hoặc bộ dữ liệu lớn, có thể tốn kém về mặt tính toán

1. Ứng dụng của Unsupervised Learning

* **Phân khúc khách hàng:**  Học tập không giám sát có thể được sử dụng để phân khúc khách hàng thành các nhóm dựa trên nhân khẩu học, hành vi hoặc sở thích của họ. Điều này có thể giúp doanh nghiệp hiểu rõ hơn về khách hàng của mình và nhắm mục tiêu đến họ bằng các chiến dịch tiếp thị phù hợp hơn.
* **Phát hiện gian lận:**  Học tập không giám sát có thể được sử dụng để phát hiện gian lận trong dữ liệu tài chính bằng cách xác định các giao dịch đi chệch khỏi mô hình dự kiến. Điều này có thể giúp ngăn chặn gian lận bằng cách gắn cờ các giao dịch này để điều tra thêm.
* **Hệ thống đề xuất:**  Học tập không giám sát có thể được sử dụng để đề xuất các mục cho người dùng dựa trên hành vi hoặc sở thích trong quá khứ của họ. Ví dụ: hệ thống đề xuất có thể sử dụng phương pháp học không giám sát để xác định những người dùng có cùng sở thích về phim và sau đó đề xuất những bộ phim mà những người dùng đó yêu thích.
* **Xử lý ngôn ngữ tự nhiên (NLP):**  Học không giám sát được sử dụng trong nhiều nhiệm vụ NLP khác nhau, bao gồm mô hình hóa chủ đề, phân cụm tài liệu và gắn thẻ từng phần của lời nói.
* **Phân tích hình ảnh:**  Học không giám sát được sử dụng trong nhiều nhiệm vụ phân tích hình ảnh, bao gồm phân đoạn hình ảnh, phát hiện đối tượng và nhận dạng mẫu hình ảnh.