MACHINE LEARNING

Cao Văn Chung
cvanchung@hus.edu.vn
Informatics Dept., MIM, HUS, VNU Hanoi

Dimensionality Reduction

Principal Component Analysis

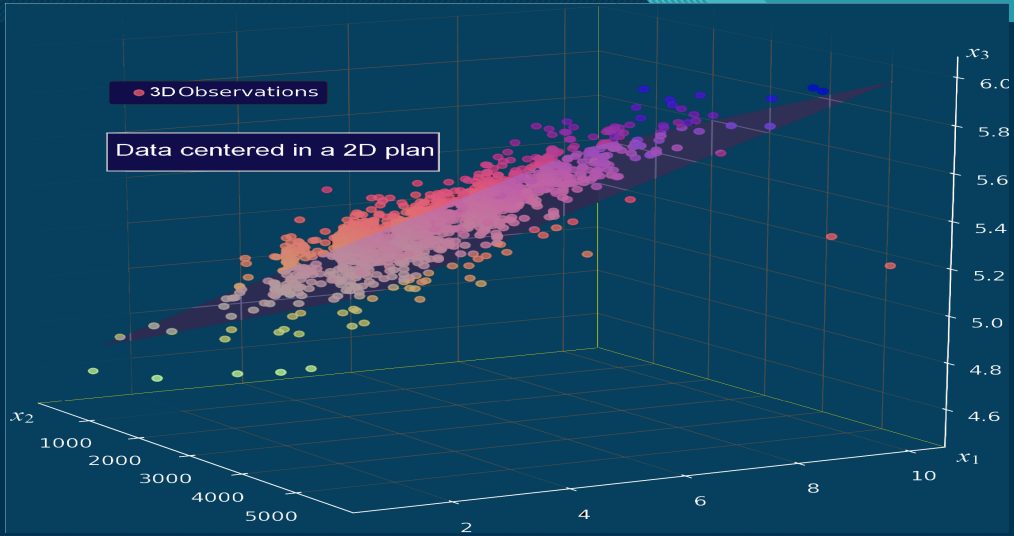
Dimensionality Reduction

- **Dimensionality Reduction** - Giảm chiều dữ liệu, là một trong những kỹ thuật quan trọng trong Machine Learning.
- Các bài toán thực tế có thể có số chiều của feature vectors rất lớn. Bên cạnh đó số lượng điểm dữ liệu cũng rất lớn.
- Lưu trữ & tính toán trực tiếp với dữ liệu có số chiều cao rất khó khăn, tốn kém, tốc độ thấp.

$\Rightarrow$ **Giảm số chiều dữ liệu** là một bước quan trọng và cần thiết.

- Đây cũng được coi là một phương pháp nén dữ liệu.
- Giảm chiều dữ liệu cũng cho phép chúng ta hiển thị trực quan phân bố các mẫu dữ liệu (trên các không gian 2D - 3D).

Dimensionality Reduction



Dimensionality Reduction

- Dimensionality Reduction, có thể quy về việc đi tìm một hàm số với đầu vào là một điểm dữ liệu ban đầu $x \in \mathbb{R}^d$ với d rất lớn, và tạo ra một điểm dữ liệu mới $z \in \mathbb{R}^k$ với $k < d$, mà các yêu cầu về thông tin của x không mất mát quá lớn.
- Trước hết ta tìm hiểu phương pháp đơn giản nhất trong các thuật toán Dimensionality Reduction - Dựa trên một mô hình tuyến tính:

  Principal Component Analysis (PCA) - **Phân tích thành phần chính**.
- PCA dựa trên quan sát rằng dữ liệu thường không phân bố ngẫu nhiên trong không gian mà thường phân bố gần các đường/mặt đặc biệt nào đó.
- PCA xem xét một trường hợp đặc biệt khi các mặt đặc biệt đó có dạng tuyến tính là các không gian con (subspace).

Dimensionality Reduction

- **Ý tưởng**: Đơn giản nhất để giảm chiều dữ liệu từ d chiều về $k < d$ chiều là chỉ giữ lại k chiều (features) phần tử quan trọng nhất.
- Tuy nhiên để hiện thực hóa, cần vượt qua:
  - Làm thế nào để xác định thành phần nào là quan trọng hơn.
  - Trường hợp xấu: lượng thông tin mà mỗi thành phần mang là như nhau, bỏ đi thành phần nào cũng dẫn đến việc mất một lượng thông tin lớn.

Giải pháp với mô hình tuyến tính:

Tìm một hệ cơ sở mới mà trong đó, biểu diễn của các vector dữ liệu ban đầu sẽ có sự khác nhau rõ rệt về tầm quan trọng giữa các thành phần. Lúc đó ta có thể bỏ qua những thành phần ít quan trọng nhất và thông tin ít bị mất mát.

Fundamental knowledge

Người đọc cần bổ sung những kiến thức sau

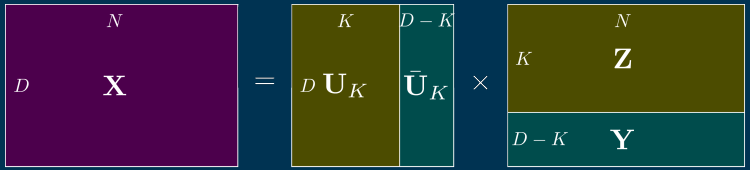
- Chuẩn 2 của ma trận ($\|A\|_2$ với $A \in \mathbb{R}^{m \times n}$).
- Hệ cơ sở của không gian tuyến tính định chuẩn, có tích vô hướng.
  - Biểu diễn vector trong các hệ cơ sở khác nhau.
  - Cơ sở trực chuẩn.
- Trace - vết của ma trận.
- Kỳ vọng và ma trận hiệp phương sai.
- Khai triển kì dị - Singular Value Decomposition (SVD)
  - Vector riêng, giá trị riêng; Khai triển kì dị - Singular Value Decomposition
  - Khai triển kì dị chặt cụt - Truncated SVD
  - Khai triển kì dị thu gọn - Compact SVD
  - Số chiều tối ưu - Optimal Rank k Approximation

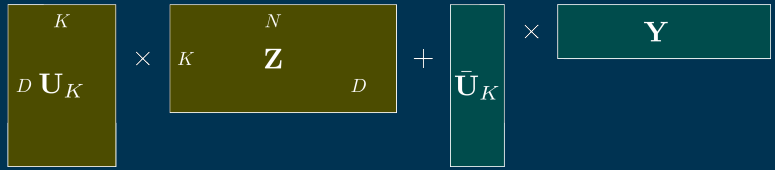
Principal Component Analysis

- Principal Component Analysis - PCA: Tìm một hệ cơ sở mới sao cho: thông tin của dữ liệu tập trung ở một số toạ độ, phần còn lại có ít thông tin.

  Để đơn giản trong tính toán, PCA sẽ tìm một hệ trực chuẩn để làm cơ sở mới.
- Giả sử hệ cơ sở trực chuẩn mới là $U = \{u_1, u_2, \ldots, u_d\}$ và chúng ta muốn giữ lại $k < d$ toạ độ trong hệ cơ sở mới này.
- Không mất tính tổng quát, ta luôn có thể giả sử đó là k thành phần đầu tiên.

Principal Component Analysis





Principal Component Analysis

- Trong hình minh họa, hệ cơ sở mới $\mathbf{U} = [\mathbf{U}_k, \bar{\mathbf{U}}_k]$ là hệ trực chuẩn với $\mathbf{U}_k$ là ma trận con tạo bởi k cột đầu tiên của $\mathbf{U}$.

  Trong hệ cơ sở mới, ma trận dữ liệu có thể được viết thành

$$\mathbf{X} = \mathbf{U}_k\mathbf{Z} + \bar{\mathbf{U}}_k\mathbf{Y}$$

- Từ đó suy ra

$$\begin{bmatrix}\mathbf{Z}\\ \mathbf{Y}\end{bmatrix} = \begin{bmatrix}\mathbf{U}_k^T\\ \bar{\mathbf{U}}_k^T\end{bmatrix}\mathbf{X} \Rightarrow \begin{matrix}\mathbf{Z} = \mathbf{U}_k^T\mathbf{X}\\ \mathbf{Y} = \bar{\mathbf{U}}_k^T\mathbf{X}\end{matrix} \quad (1)$$

- Mục đích của PCA là đi tìm ma trận trực giao $\mathbf{U}$ sao cho phần lớn thông tin được giữ lại ở phần $\mathbf{U}_k\mathbf{Z}$ và có thể lược bỏ phần $\bar{\mathbf{U}}_k\mathbf{Y}$.
- Cụ thể tìm cách thay $\mathbf{Y}$ bằng một ma trận xấp xỉ $\bar{\mathbf{Y}}$ có toàn bộ các cột như nhàu, không phụ thuộc dữ liệu test (có thể phụ thuộc dữ liệu training).

Principal Component Analysis

- Gọi mỗi cột đó là $\mathbf{b}$ và có thể coi nó là **bias**, khi đó, ta sẽ xấp xỉ $\mathbf{Y} \approx \mathbf{b}\mathbf{1}^T$, với $\mathbf{1}^T \in \mathbb{R}^{1\times N}$ là vector hàng có toàn bộ các phần tử bằng 1.

  Giả sử đã tìm được $\mathbf{U}$, ta cần tìm $\mathbf{b}$ thoả mãn

$$\mathbf{b} = \text{argmin}_{\mathbf{b}} \|\mathbf{Y} - \mathbf{b}\mathbf{1}^T\|_F^2 = \text{argmin}_{\mathbf{b}} \|\bar{\mathbf{U}}_k^T\mathbf{X} - \mathbf{b}\mathbf{1}^T\|_F^2.$$

- Giải phương trình đạo hàm theo $\mathbf{b}$ của hàm mục tiêu bằng 0:

$$(\mathbf{b}\mathbf{1}^T - \bar{\mathbf{U}}_k^T\mathbf{X})\mathbf{1} = 0 \Rightarrow N\mathbf{b} = \bar{\mathbf{U}}_k^T\mathbf{X}\mathbf{1} \Rightarrow \mathbf{b} = \bar{\mathbf{U}}_k^T\bar{\mathbf{x}}.$$

- Dễ thấy, nếu vector kỳ vọng $\bar{\mathbf{x}} = \mathbf{0}$, việc tính toán sẽ thuận tiện hơn nhiều.

Principal Component Analysis

- Có thể đạt được kỳ vọng $\bar{\mathbf{x}} = \mathbf{0}$, nếu ngay từ đầu, ta trừ mỗi vector dữ liệu đi vector kỳ vọng $\bar{\mathbf{x}}$ của toàn bộ dữ liệu.

  Với giá trị $\mathbf{b}$ tìm được này, dữ liệu ban đầu sẽ được xấp xỉ với:

$$\mathbf{X} \approx \tilde{\mathbf{X}} = \mathbf{U}_k\mathbf{Z} + \bar{\mathbf{U}}_k\bar{\mathbf{U}}_k^T\bar{\mathbf{x}}\mathbf{1}^T.$$

- Kết hợp các phương trình trên, ta định nghĩa hàm tổn thất như sau:

$$J = \frac{1}{N}||\mathbf{X} - \tilde{\mathbf{X}}||_F^2 = \frac{1}{N}||\bar{\mathbf{U}}_k\bar{\mathbf{U}}_k^T\mathbf{X} - \bar{\mathbf{U}}_k\bar{\mathbf{U}}_k^T\bar{\mathbf{x}}\mathbf{1}^T||_F^2. \quad (2)$$

- Chú ý: Nếu các cột của một ma trận $\mathbf{V}$ tạo thành một hệ trực chuẩn thì với một ma trận $\mathbf{W}$ bất kỳ, ta luôn có:

$$\|\mathbf{VW}\|_F^2 = \text{trace}(\mathbf{W}^T\mathbf{V}^T\mathbf{VW}) = \text{trace}(\mathbf{W}^T\mathbf{W}) = \|\mathbf{W}\|_F^2.$$

Principal Component Analysis

- Vì vậy hàm tổn thất trong công thức (2) có thể viết lại thành:

$$J = \frac{1}{N}\|\bar{\mathbf{U}}_k^T(\mathbf{X} - \bar{\mathbf{x}}\mathbf{1})^T\|_F^2 = \frac{1}{N}\|\bar{\mathbf{U}}_k^T\hat{\mathbf{X}}\|_F^2 = \frac{1}{N}\|\hat{\mathbf{X}}^T\bar{\mathbf{U}}_k\|_F^2$$
$$= \frac{1}{N}\sum_{i=K+1}^{D}\|\hat{\mathbf{X}}^T\mathbf{u}_i\|_2^2 = \frac{1}{N}\sum_{i=K+1}^{D}\mathbf{u}_i^T\hat{\mathbf{X}}\hat{\mathbf{X}}^T\mathbf{u}_i = \sum_{i=K+1}^{D}\mathbf{u}_i^T\mathbf{S}\mathbf{u}_i. \quad (3)$$

- Ở đây $\mathbf{S}$ là ma trận hiệp phương sai của dữ liệu và $\hat{\mathbf{X}} = \left(\hat{x}_1, \ldots, \hat{x}_N\right)$, trong đó $\hat{\mathbf{x}}_n = \mathbf{x}_n - \bar{\mathbf{x}}_n$, $n = 1, \ldots, N$.
- Viết dạng ma trận $\hat{\mathbf{X}} = \mathbf{X} - \bar{\mathbf{x}}\mathbf{1}^T$ và $\mathbf{S} = \frac{1}{N}\hat{\mathbf{X}}\hat{\mathbf{X}}^T$. Chú ý $\mathbf{S}$ là ma trận bán xác định dương.

Principal Component Analysis

- Chú ý hệ $\{\mathbf{u}_i\}$, $i = 1, \ldots, d$ trực chuẩn, cùng với tính đối xứng, nửa xác định dương của $\mathbf{S}$, ta có

$$\begin{aligned} L &= \sum_{i=1}^{D} \mathbf{u}_i^T \mathbf{S} \mathbf{u}_i = \frac{1}{N} \|\hat{\mathbf{X}}^T \mathbf{U}\|_F^2 = \frac{1}{N} \text{trace}(\hat{\mathbf{X}}^T \mathbf{U} \mathbf{U}^T \hat{\mathbf{X}}) \\ &= \frac{1}{N} \text{trace}(\hat{\mathbf{X}}^T \hat{\mathbf{X}}) = \frac{1}{N} \text{trace}(\hat{\mathbf{X}} \hat{\mathbf{X}}^T) = \text{trace}(\mathbf{S}) = \sum_{i=1}^{D} \lambda_i \end{aligned} \tag{4}$$

- Ở đây $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D \geq 0$ là các giá trị riêng của ma trận đối xứng nửa xác định dương $\mathbf{S}$, do đó chúng là các số thực không âm.

Principal Component Analysis

- Từ đẳng thức cuối của (4) suy ra L không phụ thuộc vào hệ cơ sở $\mathbf{U}$ mà chỉ phụ thuộc bản thân dữ liệu. Cụ thể hơn, L chính là tổng của các phương sai theo từng thành phần của dữ liệu ban đầu.
- Bài toán cực tiểu hóa hàm tổn thất J trong (2) tương đương với cực đại hóa

$$F = L - J = \sum_{i=1}^{k} \mathbf{u}_i \mathbf{S} \mathbf{u}_i^T.$$

Principal Component Analysis

Định lý

F đạt giá trị lớn nhất $\max F = \sum_{i=1}^{k} \lambda_i$ khi các vector riêng $\{\mathbf{u}_i\}$ ứng với giá trị riêng $\{\lambda_i\}$ lập thành một hệ trực chuẩn: $\{\mathbf{u}_i\}$ trực giao và $\|\mathbf{u}_i\|_2 = 1$, $i = 1, \ldots, k$.

- Định lý trên có thể được chứng minh bằng quy nạp theo k.
- Giá trị riêng lớn nhất λ_1 được gọi là *Thành phần chính thứ nhất* (First Principal Component); trị riêng thứ hai λ_2 còn được gọi là *Thành phần chính thứ hai*.v.v.
- Chính vì vậy, phương pháp này có tên gọi là *Phân tích thành phần chính* - Principal Component Analysis.

Principal Component Analysis

- **Các bước của phương pháp PCA**
  1. Tính vector kỳ vọng của toàn bộ dữ liệu:

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$

  2. Tính dữ liệu chuẩn hóa $\hat{\mathbf{x}}_n$

$$\hat{\mathbf{x}}_n = \mathbf{x}_n - \bar{\mathbf{x}}; \qquad n = 1, 2, \ldots, N.$$

  3. Tính ma trận hiệp phương sai:

$$\mathbf{S} = \frac{1}{N}\hat{\mathbf{X}}\hat{\mathbf{X}}^T$$

Principal Component Analysis

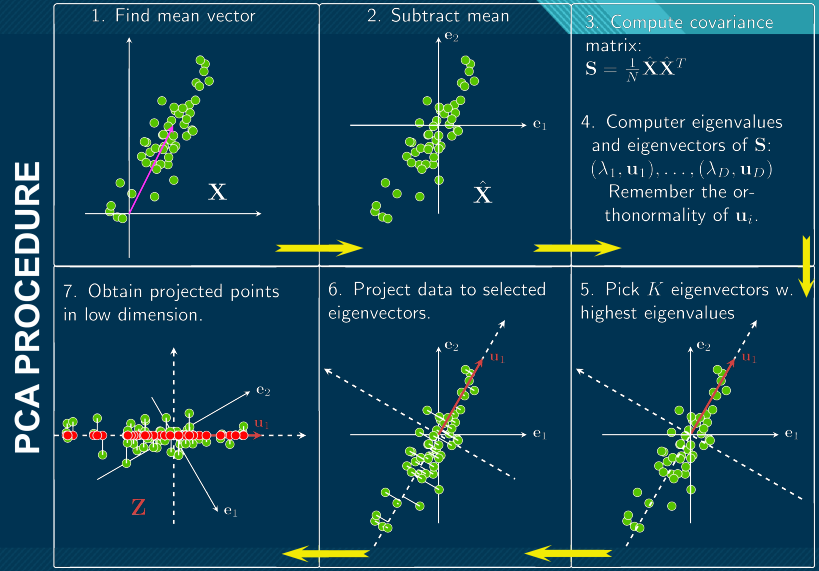
- **Các bước của phương pháp PCA**
  4. Tính các trị riêng λ_i và vector riêng $\mathbf{u}_i$ có $\|\mathbf{u}_i\|_2 = 1$ của ma trận này, sắp xếp chúng theo thứ tự giảm dần của trị riêng.
  5. Chọn k giá trị riêng lớn nhất và k vector riêng trực chuẩn tương ứng. Xây dựng ma trận $\mathbf{U}_k$.
  6. Các cột của $\{\mathbf{U}_i\}_{i=1}^{k}$ là hệ cơ sở trực chuẩn, tạo thành không gian con của k thành phần chính, it gần với phân bố của dữ liệu ban đầu đã chuẩn hoá.
  7. Chiếu dữ liệu ban đầu đã chuẩn hoá $\hat{\mathbf{X}}$ xuống không gian con nói trên. Dữ liệu mới chính là toạ độ của các điểm dữ liệu trên không gian mới

$$\mathbf{Z} = \mathbf{U}_k^T \hat{\mathbf{X}}.$$

- Dữ liệu ban đầu được xấp xỉ theo dữ liệu mới

$$\mathbf{x} \approx \mathbf{U}_k \mathbf{Z} + \bar{\mathbf{x}}.$$

Principal Component Analysis



PCA & Truncated SVD

- Cho ma trận $\mathbf{X} \in \mathbb{R}^{d \times N}$, xét bài toán xấp xỉ X bởi A với $\text{rank}(A) \leq k$

$$\min_{\mathbf{A}} \|\mathbf{X} - \mathbf{A}\|_F \qquad \text{s.t.} \quad \text{rank}(\mathbf{A}) = K.$$

- Giả sử SVD của $\mathbf{X}$ là $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$, với $\mathbf{U} \in \mathbb{R}^{D \times D}$ và $\mathbf{V} \in \mathbb{R}^{N \times N}$ là trực giao; $\Sigma \in \mathbb{R}^{D \times N}$ là ma trận đường chéo (không nhất thiết vuông) và các phần tử trên đường chéo không âm, giảm dần.
- Lúc đó nghiệm của bài toán xấp xỉ trên sẽ là

$$\mathbf{A} = \mathbf{U}_k \Sigma_k \mathbf{V}_k^T$$

với $\mathbf{U} \in \mathbb{R}^{d \times k}$ và $\mathbf{V} \in \mathbb{R}^{N \times k}$ là k cột đầu tiên của $\mathbf{U}$, $\mathbf{V}$; $\Sigma_k \in \mathbb{R}^{k \times k}$ là ma trận đường chéo con ứng với k hàng, k cột đầu tiên của Σ.

PCA & Truncated SVD

Xét phương pháp PCA: Xét dữ liệu chuẩn hóa $\hat{\mathbf{X}}$, lúc đó kỳ vọng $\bar{\mathbf{x}} = 0$.

- Nghiệm xấp xỉ của PCA trở thành

$$\hat{\mathbf{X}} \approx \tilde{\mathbf{X}} = \mathbf{U}_k \mathbf{Z}.$$

- Bài toán tối ưu của PCA sẽ trở thành

$$[\mathbf{U}_k, \mathbf{Z}] = \min_{\mathbf{U}_k, \mathbf{Z}} \|\hat{\mathbf{X}} - \mathbf{U}_k \mathbf{Z}\|_F \quad \text{s.t.:} \quad \mathbf{U}_k^T \mathbf{U}_k = \mathbf{I}_k$$

  với $\mathbf{I}_k \in \mathbb{R}^{k \times k}$ là ma trận đơn vị trong không gian các ma trận vuông k chiều; điều kiện ràng buộc chính là hệ $\mathbf{U}_k$ trực chuẩn.

Dễ dàng thấy rằng, với dữ liệu được chuẩn hóa $\hat{\mathbf{X}}$ ($\bar{\mathbf{x}} = 0$), nghiệm bài toán tối ưu trong phương pháp PCA chính là nghiệm Truncated SVD của ma trận dữ liệu (với các điểm dữ liệu ứng với các cột, các điểm được viết thành hàng trong ma trận).

Optimal Dim k

- Số chiều k được xác định dựa trên lượng thông tin muốn giữ lại.
- PCA còn được gọi là phương pháp *tối đa tổng phương sai được giữ lại*.
  - Có thể coi tổng các phương sai được giữ lại là lượng thông tin được giữ lại.
  - Với phương sai càng lớn, tức dữ liệu có độ phân tán cao, thể hiện lượng thông tin càng lớn.
- Nhắc lại

$$L = \sum_{i=1}^{D} \mathbf{u}_i^T \mathbf{S} \mathbf{u}_i = \frac{1}{N} \text{trace}(\hat{\mathbf{X}}^T \hat{\mathbf{X}}) = \frac{1}{N} \text{trace}(\hat{\mathbf{X}} \hat{\mathbf{X}}^T) = \text{trace}(\mathbf{S}) = \sum_{i=1}^{D} \lambda_i$$

tức là trong mọi hệ trục toạ độ, tổng phương sai của dữ liệu là như nhau và bằng tổng các trị riêng của ma trận hiệp phương sai $\sum_{i=1}^{d} \lambda_i$.

Optimal Dim k

- PCA với số chiều k giữ lại lượng thông tin (tổng các phương sai) là $\sum_{i=1}^{k} \lambda_i$.
- Tỷ lệ thông tin được giữ lại trong PCA có thể tính theo

$$r_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{d} \lambda_j}.$$

- Ví dụ: để giữ lại 99% thông tin, cần chọn số chiều k là số tự nhiên nhỏ nhất sao cho $r_k \geq 0.99$.
- Chú ý khi dữ liệu phân bố quanh một không gian con, các giá trị phương sai lớn nhất ứng với các λ_i đầu tiên lớn hơn nhiều so với các phương sai còn lại. Do đó ta có thể chọn được k khá nhỏ để đạt được $r_k \geq 0.99$.

Applying of PCA in practice

Trong phần này ta giả thiết dữ liệu đã được chuẩn hóa. Lúc đó

$$\bar{\mathbf{x}} = 0 \quad \text{và} \quad \mathbf{S} = \frac{1}{N}\mathbf{X}\mathbf{X}^T.$$

- Có hai trường hợp trong thực tế mà chúng ta cần lưu ý về PCA
  - Lượng dữ liệu quan sát nhỏ hơn rất nhiều so với số chiều dữ liệu: $N < d$;
  - Lượng dữ liệu trong tập quan sát là rất lớn (massive dataset).
- Hậu quả: Tính toán trực tiếp ma trận hiệp phương sai và các giá trị riêng có thể trở nên bất khả thi.

Dim. $d > N$ - cardinal of data

- $\mathbf{X} \in \mathbb{R}^{d \times N}$ với $d > N$. Trước hết cần chọn $K < \text{rank}(\mathbf{X}) \leq N$.
- Để tính các giá trị riêng và vector riêng của $\mathbf{S}$ cần đến một số tính chất
  - **Tính chất 1**: Giá trị riêng của A cũng là giá trị riêng của kA với $k \neq 0$ bất kỳ.
  - **Tính chất 2**: Với $d_1, d_2 > 0$ là các số tự nhiên bất kỳ; $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$ và $\mathbf{B} \in \mathbb{R}^{d_2 \times d_1}$, lúc đó tập các giá trị riêng của ma trận $\mathbf{AB}$ cũng là tập giá trị riêng của $\mathbf{BA}$[1].
  - **Tính chất 3**: Giả sử $(\lambda, \mathbf{u})$ là một cặp trị riêng - vector riêng của $\mathbf{T} = \mathbf{X}^T\mathbf{X}$, lúc đó $(\lambda, \mathbf{Xu})$ là cặp trị riêng - vector riêng của $\mathbf{S} = \mathbf{X}\mathbf{X}^T$. Thật vậy

$$\mathbf{X}^T\mathbf{X}\mathbf{u} = \lambda\mathbf{u} \Rightarrow (\mathbf{X}\mathbf{X}^T)(\mathbf{X}\mathbf{u}) = \lambda\mathbf{X}\mathbf{u}.$$

- Vậy thay vì tìm trị riêng của ma trận hiệp phương sai $\mathbf{S} \in \mathbb{R}^{d \times d}$, ta có thể tìm trị riêng của $\mathbf{T} = \mathbf{X}^T\mathbf{X} \in \mathbb{R}^{N \times N}$ có số chiều $N < d$.

[1] Theorem 1.3.22, "Matrix Analysis", Horn and Johnson - the second edition.

Chuẩn hóa hệ vector riêng

Không gian con riêng: ứng với trị riêng của một ma trận là không gian sinh (*span subspace*) tạo bởi toàn bộ các vector riêng ứng với trị riêng đó.

- Ta cần chuẩn hoá các vector riêng $\{u_i\}$ tìm được sao cho chúng tạo thành một hệ trực chuẩn.
- Việc này có thể được thực hiện dựa trên một số tính chất của không gian con riêng ứng với các giá trị riêng.

Chuẩn hóa hệ vector riêng

- Tính chất của hệ riêng $\{(\lambda_i, u_i)\}$
  - **Chú ý 1**: Nếu $\mathbf{A}$ là một ma trận đối xứng, $(\lambda_1, \mathbf{x}_1), (\lambda_2, \mathbf{x}_2)$ là các căp trị riêng - vector riêng của $\mathbf{A}$ với $\lambda_1 \neq \lambda_2$, thì $\mathbf{x}_1^T \mathbf{x}_2 = 0$. Thật vậy

$$\mathbf{x}_2^T \mathbf{A} \mathbf{x}_1 = \mathbf{x}_1^T \mathbf{A} \mathbf{x}_2 = \lambda_1 \mathbf{x}_2^T \mathbf{x}_1 = \lambda_2 \mathbf{x}_1^T \mathbf{x}_2 \Rightarrow \mathbf{x}_1^T \mathbf{x}_2 = 0 \quad \text{do} \quad \lambda_1 \neq \lambda_2.$$

  - Tính chất trên suy ra, hai vector bất kỳ trong hai không gian riêng ứng với hai trị riêng khác nhau của một ma trận đối xứng thì trực giao với nhau.
  - **Chú ý 2**: với các trị riêng độc lập tìm được trong một không gian riêng, ta có thể dùng thuật toán Gram-Schmit để chuẩn hoá chúng về một hệ trực chuẩn.
- Kết hợp hai điểm trên, ta có thể thu được các vector riêng tạo thành một hệ trực chuẩn, chính là ma trận $\mathbf{U}_k$ trong phương pháp PCA.

Trường hợp large-scale

- Thực tế có những bài toán mà d, N rất lớn. Lúc đó cần giải tìm hệ riêng của ma trận với kích thước lớn.
- **Ví dụ:** Dữ liệu là có 1 triệu bức ảnh 1000×1000 pixels (số chiều là 10^6 - rất lớn).
- Trực tiếp tính toán giá trị riêng và vector riêng cho ma trận hiệp phương sai là không khả thi.
- Có thể sử dụng phương pháp lặp **Power Method**[2] để xấp xỉ giá trị riêng λ lớn nhất và vector riêng (chuẩn hóa) ứng với nó.

Thuật toán Power Method có các bước như trang sau.

[2] https://www.cs.huji.ac.il/ csip/tirgul2.pdf

Appendix: Power Method

$\mathbf{A} \in \mathbb{R}^{n \times n}$ - Ma trận nửa xác định dương.

1. Chọn vector $\mathbf{q}^{(0)} \in \mathbb{R}^n$ với $\|\mathbf{q}^{(0)}\|_2 = 1$ bất kỳ.
2. Tại các bước $i = 1, 2, \ldots$, tính $\mathbf{z} = \mathbf{A}\mathbf{q}^{(i-1)}$.
3. Chuẩn hóa $\mathbf{q}^{(k)} = \mathbf{z}/\|\mathbf{z}\|_2$.
4. Nếu $\|\mathbf{q}^{(k)} - \mathbf{q}^{(k-1)}\|_2$ đủ nhỏ thì sang bước 5. Ngược lại, đặt $k := k + 1$ và quay lại bước 2.)
5. Tính: $\lambda_1 = (\mathbf{q}^{(k)})^T \mathbf{A}\mathbf{q}^{(k)}$, đây là giá trị riêng lớn nhất của $\mathbf{A}$; $\mathbf{q}^{(k)}$ là vector riêng ứng với λ_1. Kết thúc.

Phương pháp Power hội tụ nhanh![2]
Sau khi tìm được trị riêng lớn nhất λ_1, có thể tìm các giá trị riêng tiếp theo dựa vào tính chất

[2] https://www.cs.huji.ac.il/ csip/tirgul2.pdf

Appendix: Power Method

Định lý

Nếu ma trận nửa xác định dương $\mathbf{A}$ có các trị riêng $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n (\geq 0)$ và các vector riêng tương ứng $\mathbf{v}_1, \ldots, \mathbf{v}_n$ tạo thành một hệ trực chuẩn, thì ma trận

$$\mathbf{B} = \mathbf{A} - \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T$$

có các trị riêng $\lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n \geq 0$ tương ứng với các vector riêng $\mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_n$.

- Thật vậy, với $i = 1$: $\mathbf{B}\mathbf{v}_1 = (\mathbf{A} - \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T)\mathbf{v} = \mathbf{A}\mathbf{v}_1 - \lambda_1 \mathbf{v}_1 = \mathbf{0}$.
- Với $i > 1$: $\mathbf{B}\mathbf{v}_i = (\mathbf{A} - \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T)\mathbf{v}_i = \mathbf{A}\mathbf{v}_i - \lambda_1 \mathbf{v}_1 (\mathbf{v}_1^T \mathbf{v}_i) = \mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$.

Appendix: Power Method

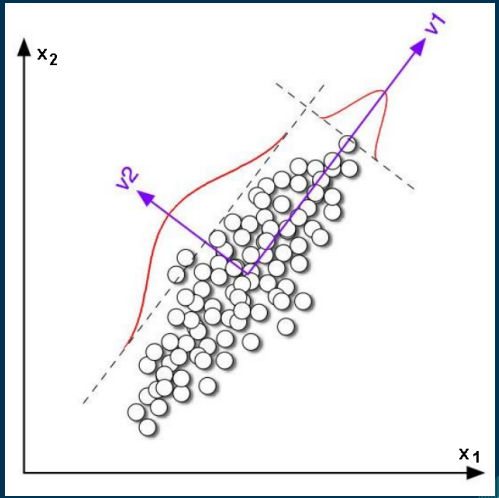
- Đến đây ta có thủ tục lặp để tìm k cặp trị riêng - vector riêng tương ứng $(\lambda_i, u_i)_{i=1}^k$ với $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_k \geq \lambda_j \geq 0$, $\forall j > k$, như sau
  0. Bước đầu, $CountPair = 1$;
  1. Áp dụng Power Method để tìm cặp $(\bar{\lambda}_1, \bar{u}_1)$ với $\bar{\lambda}_1$ lớn nhất.
  2. Gán $\lambda_{CountPair} := \bar{\lambda}_1$; $u_{CountPair} := \bar{u}_1$. Nếu $CountPair = k$, sang bước 4. Ngược lại, sang bước 3.
  3. Áp dụng kết quả định lý trước, đặt: $\mathbf{A} = \mathbf{A} - \bar{\lambda}_1 \bar{u}_1 \bar{u}_1^T$.
     Đặt $CountPair = CountPair + 1$, quay lại bước 1.
  4. Hệ $\{(\lambda_i, u_i)\}_{i=1}^k$ thu được chính là hệ cần tìm.
- Với quy trình trên, ta sẽ tìm được (xấp xỉ) k trị riêng và vector riêng tương ứng của ma trận hiệp phương sai $\mathbf{S}$, với $k \leq \text{rank}(\mathbf{S})$ tùy ý.

Remarks

- PCA là một phương pháp Unsupervised.
- Việc thực hiện PCA trên toàn bộ dữ liệu không phụ thuộc vào class(nếu có) của mỗi dữ liệu.
- PCA đôi khi không hiệu quả thậm chí gây sai lệch khi áp dụng cho các bài toán classification.
- Với các bài toán Large-scale, đôi khi việc tính toán trên toàn bộ dữ liệu là không khả thi vì còn có vấn đề về bộ nhớ.
- Giải pháp là thực hiện PCA lần đầu trên một tập con dữ liệu vừa với bộ nhớ, sau đó lấy một tập con khác để (incrementally) cập nhật nghiệm của PCA tới khi nào hội tụ.
- Có nhiều hướng mở rộng của PCA, có thể tìm kiếm theo từ khoá: Sparse PCA, Kernel PCA, Robust PCA.

PCA vs. Labeled Data

- PCA tìm hệ cơ sở mới của không gian, trong đó các thuộc tính (chiều) dữ liệu có chênh lệch lớn về "độ quan trọng".
- **Độ quan trọng** - xác định qua độ lớn của **phương sai** thuộc tính (chiều) dữ liệu đó
- PCA không quan tâm đến nhãn/phân loại của dữ liệu.

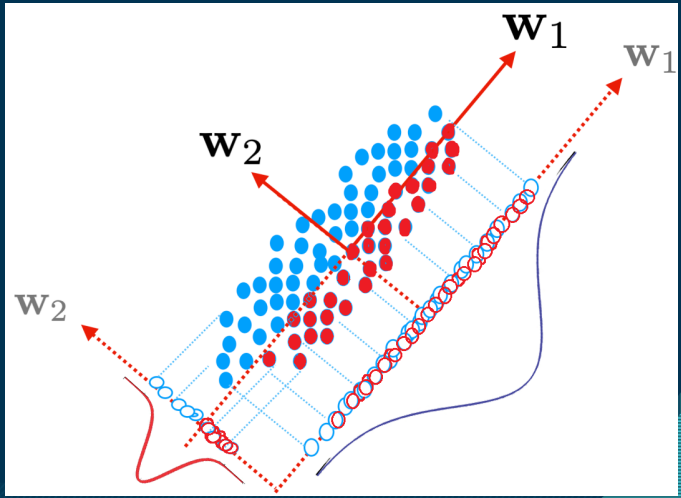


PCA vs. Labeled Data

- Trường hợp dữ liệu có nhãn, việc chỉ giữ lại thành phần có phương sai lớn có thể làm giảm tính phân biệt của các lớp nhãn dữ liệu.

⇒ Giảm hiệu quả nếu sử dụng dữ liệu mới sau PCA cho các bài toán phân loại!

- Để bảo toàn tính tách biệt giữa các lớp dữ liệu ⇒ Cần phương pháp giảm chiều khác!



Linear Discriminant Analysis

Linear Discriminant Analysis

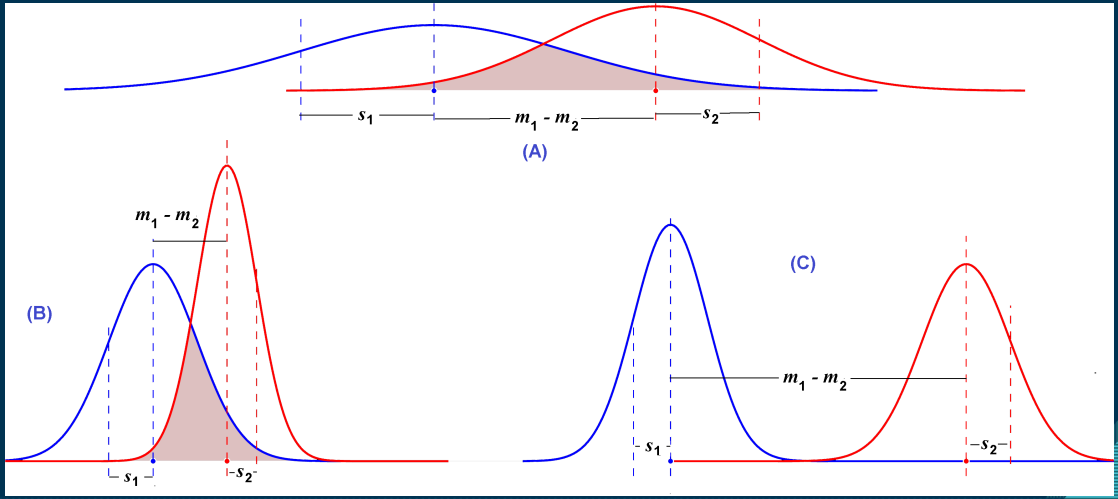
Linear Discriminant Analysis - Phân tích phân biệt tuyến tính

- Linear: Là phương pháp tuyến tính;
- Discriminant Analysis: Phân tích độ phân biệt.
- Như tên gọi, LDA là phương pháp giảm chiều dựa trên việc chiếu xuống không gian tuyến tính con (SubSpace).
- Phương pháp có tính đến việc áp dụng cho các loại dữ liệu có nhãn (dữ liệu có phân loại).
- Phương pháp ưu tiên **độ phân biệt (Discriminant)** giữa các nhóm dữ liệu tách được tuyến tính.

Linear Discriminant Analysis

- Xét dữ liệu với 02 nhãn: Để đơn giản, trước hết ta xét dữ liệu với 2 nhãn (VD: binary classification): $\{01; 02\}$;
- Ký hiệu: Tập dữ liệu có nhãn 01 là X_1; Tập dữ liệu có nhãn 02 là X_2.
- Giả sử chiếu xuống 01 chiều bất kỳ của không gian dữ liệu, $X_1; X_2$ có kỳ vọng $m_1; m_2$ - phương sai $s_1; s_2$, tương ứng.
- Phân tích độ phân biệt trong các trường hợp của (m_1, m_2) và (s_1^2, s_2^2) như trong hình minh họa sau.

Linear Discriminant Analysis



Linear Discriminant Analysis

Trong hình minh họa, ta dùng đồ thị hình chuông (phân bố chuẩn) chỉ để dễ hình dung - Dữ liệu thực tế không nhất thiết phải có phân bố chuẩn)

Phân tích các trường hợp:

(A) Khoảng cách giữa các kỳ vọng $|m_2 - m_2|$ lớn nhưng phương sai (s_1^2, s_2^2) cũng lớn: Dữ liệu quá phân tán dẫn đến vùng chồng lấn có thể lớn - **độ phân biệt (Discriminant)** thấp.

(B) Phương sai (s_1^2, s_2^2) nhỏ nhưng khoảng cách $|m_2 - m_2|$ cũng nhỏ: Hình chiếu của các tập dữ liệu phân bố quá gần nhau dẫn đến có thể vùng chồng lấn lớn - **độ phân biệt (Discriminant)** thấp.

(C) Phương sai (s_1^2, s_2^2) nhỏ và khoảng cách $|m_2 - m_2|$ lớn: Hình chiếu của các tập dữ liệu phân bố ít chồng lấn - **độ phân biệt (Discriminant)** cao.

Linear Discriminant Analysis

Từ phân tích các trường hợp của khoảng cách giữa các kỳ vọng và phương sai của hai tập $X_1; X_2$:

- Phương sai s_i $(s = 1, 1)$ nhỏ thể hiện việc dữ liệu ít phân tán. Điều này nghĩa là dữ liệu trong mỗi tập $X_1; X_2$ có xu hướng tập trung. Đặt $N_j = |X_j|, j = 1, 2$ - Các tổng phương sai $(N_j s_j^2)$ còn được gọi là các **within-class variances**.
- Khoảng cách giữa các kỳ vọng $|m_1 - m_2|$ lớn chứng tỏ rằng hai classes $X_1; X_2$ nằm xa nhau, tức dữ liệu giữa các classes có độ phân biệt (Discriminant) cao. Bình phương khoảng cách giữa hai kỳ vọng $|m_1 - m_2|^2$ còn được gọi là **between-class variance**.

Các class dữ liệu sẽ có độ phân biệt cao nếu **within-class variances** nhỏ (dữ liệu trong mỗi class có tính tập trung cao) và **between-class variance** lớn (khoảng cách giữa các class lớn).

Linear Discriminant Analysis - Loss function

Linear Discriminant Analysis (LDA) tìm phép chiếu xuống không gian (tuyến tính) con sao cho trên đó, tỉ lệ **between-class variance**/**within-class variances** lớn nhất có thể.

- Sắp xếp lại chỉ số để:

$$X_1 = \{x_n | label(x_n) = 1; 1 \leq n \leq N_1\} \quad \text{và}$$
$$X_2 = \{x_m | label(x_m) = 2; N_1 + 1 \leq m \leq N\}.$$

- Phép chiếu (điểm) dữ liệu $x \in \mathbb{R}^d$ xuống đường thẳng có thể biểu diễn dạng

$$z = Prj(x) = \mathbf{w}^T \mathbf{x}, \quad \mathbf{w} \in \mathbb{R}^d.$$

Linear Discriminant Analysis - Loss function

- Kỳ vọng của các class X_1, X_2:

$$\mathbf{m}_j = \frac{1}{N_j} \sum_{n \in \mathcal{X}_j} \mathbf{x}_n, \quad j = 1, 2.$$

- Hình chiếu của các kỳ vọng trên đường thẳng xác định bởi

$$\bar{m}_j = Prj(\mathbf{m}_j) = \mathbf{w}^T \mathbf{m}_j = \frac{1}{N_j} \sum_{n \in \mathcal{X}_j} \mathbf{w}^T \mathbf{x}_n$$

- Khoảng cách giữa các hình chiếu của kỳ vọng của X_1, X_2

$$\bar{m}_1 - \bar{m}_2 = \mathbf{w}^T (\mathbf{m}_1 - \mathbf{m}_2) = \mathbf{w}^T \left(\frac{1}{N_2} \sum_{x_n \in \mathcal{X}_1} \mathbf{x}_n - \frac{1}{N_1} \sum_{x_m \in \mathcal{X}_2} \mathbf{x}_m \right).$$

Linear Discriminant Analysis - Loss function

- Tổng phương sai hình chiếu, tức within-class variances của các class X_1, X_2:

$$S_k^2 = \sum_{n \in \mathcal{X}_k} \left(\mathbf{w}^T x_n - \mathbf{w}^T \mathbf{m}_k\right)^2 = \sum_{n \in \mathcal{X}_k} (w^T x_n - \bar{m}_k)^2, \quad k = 1, 2$$

  Các within-class variances ở đây được tính là tổng bình phương độ lệch của hình chiếu dữ liệu trong mỗi class so với kỳ vọng.

- LDA đi tìm giá trị lớn nhất của tỷ lệ sau theo w:

$$J(w) := \frac{(\bar{m}_1 - \bar{m}_2)^2}{S_1^2 + S_2^2} \longrightarrow \max_w.$$

Linear Discriminant Analysis - Loss function

- Tử thức:

$$
\begin{aligned}
(\bar{m}_1 - \bar{m}_2)^2 &= \left(\mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_2)\right)^2 = \\
&= \left(\mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_2)\right)\left(\mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_2)\right)^T = \\
&= \mathbf{w}^T \underbrace{(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T}_{\mathbf{S}_B} \mathbf{w} = \mathbf{w}^T \mathbf{S}_B \mathbf{w}
\end{aligned} \tag{5}
$$

  Ma trận $\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$ đối xứng, nửa xác định dương - còn gọi là **between-class covariance matrix**.

- Ở đây chú ý $\left(\mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_2)\right)$ là vô hướng (01 chiều) nên

$$
\left(\mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_2)\right) = \left(\mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_2)\right)^T.
$$

Linear Discriminant Analysis - Loss function

- Mẫu thức:

$$
\begin{aligned}
S_1^2 + S_2^2 =& \sum_{n \in \mathcal{C}_1} \left(\mathbf{w}^T(\mathbf{x}_n - \mathbf{m}_1)\right)^2 + \sum_{n \in \mathcal{C}_2} \left(\mathbf{w}^T(\mathbf{x}_n - \mathbf{m}_2)\right)^2 \\
=& \mathbf{w}^T \underbrace{\sum_{k=1}^{2} \sum_{n \in \mathcal{C}_k} (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^T}_{\mathbf{S}_W} \mathbf{w} = \mathbf{w}^T \mathbf{S}_W \mathbf{w}
\end{aligned} \tag{6}
$$

Do ma trận $\mathbf{S}_W = \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^T + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_2)^T$ là tổng của 02 ma trận đối xứng, nửa xác định dương - nên nó cũng là đối xứng, nửa xác định dương, còn gọi là **within-class covariance matrix**.

Linear Discriminant Analysis - Optimization

- Từ (5) và (6) suy ra bài toán tối ưu của LDA:

$$\mathbf{w} = \arg\max_{\mathbf{w}} J(w) = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \tag{7}$$

- Viết bài toán dạng (7) ta mặc định $\mathbf{S}_W$ không suy biến, nhưng $\mathbf{S}_W$ chỉ nửa xác định dương nên điều này thực tế có thể không thỏa mãn.
- Lúc đó có thể sử dụng $\bar{\mathbf{S}}_W := \mathbf{S}_W + \lambda \mathbf{I}$ với $\lambda > 0$ đủ bé thay thế cho $\mathbf{S}_W$.

Linear Discriminant Analysis - Optimization

- Tìm nghiệm (7) từ các điểm tới hạn $\nabla_{\mathbf{w}} J(\mathbf{w}) = 0$:

$$\begin{aligned} \nabla_{\mathbf{w}} J(\mathbf{w}) &= \frac{1}{(\mathbf{w}^T \mathbf{S}_W \mathbf{w})^2} \left(2\mathbf{S}_B \mathbf{w}(\mathbf{w}^T \mathbf{S}_W \mathbf{w}) - 2\mathbf{w}^T \mathbf{S}_B \mathbf{w} \mathbf{w}^T \mathbf{S}_W \mathbf{w}\right) = \mathbf{0} \\ \Leftrightarrow \mathbf{S}_B \mathbf{w} &= \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \mathbf{S}_W \mathbf{w} \Leftrightarrow \mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w} = J(\mathbf{w}) \mathbf{w} \end{aligned} \tag{8}$$

  Trong (8) ta áp dụng $\nabla_{\mathbf{w}} \mathbf{w}^T A \mathbf{w} = 2A\mathbf{w}$ với A đối xứng.
- Từ $\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w} = J(\mathbf{w}) \mathbf{w}$ và $J(\mathbf{w}) \in \mathbb{R}$ suy ra $\mathbf{w}$ là vector riêng ứng với giá trị riêng $J(\mathbf{w})$ của ma trận $\mathbf{S}_W^{-1} \mathbf{S}_B$.
- Vậy **giá trị cực đại** của $J(\mathbf{w})$ là **giá trị riêng lớn nhất** của $\mathbf{S}_W^{-1} \mathbf{S}_B$.

Linear Discriminant Analysis - Optimization

- Chú ý: $\mathbf{S}_W^{-1}\mathbf{S}_B$ không phụ thuộc $\mathbf{w}$ - là đối xứng, nửa xác định dương, nên tất cả giá trị riêng là thực, không âm, không phụ thuộc $\mathbf{w}$.
- Đặt giá trị riêng lớn nhất của $\mathbf{S}_W^{-1}\mathbf{S}_B$ là L ($L = J(\mathbf{w})$). Luôn chọn được $\mathbf{w}$ sao cho (do hệ 01 phương trình $d > 1$ ẩn - vô số nghiệm):

$$(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{w} = L = J(\mathbf{w})$$

- Từ (5) : $\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$ và định nghĩa của L, thay vào đẳng thức trên dẫn tới

$$L\mathbf{w} = \mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)\underbrace{(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{w}}_{L} = L\mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

Linear Discriminant Analysis - Optimization

Đẳng thức cuối:

$$L\mathbf{w} = L\mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

suy ra chỉ cần chọn:

$$\mathbf{w} = \beta\mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

với $\beta \neq 0$ bất kỳ, thì

- $\mathbf{w}$ là vector riêng của $\mathbf{S}_W^{-1}\mathbf{S}_B$ tương ứng với trị riêng lớn nhất L

$$L\mathbf{w} = \mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w} = \left[\max_{\mathbf{w}} J(\mathbf{w})\right].\mathbf{w}$$

- Do đó

$$\mathbf{w} = \arg\max_{w} J(w) \quad \text{và} \quad \max_{w} J(\mathbf{w}) = L.$$

Linear Discriminant Analysis for Multinomial Data.

LDA for Multinomial Data

- Giả thiết: Dữ liệu $\mathbf{x} \in \mathbb{R}^d$ thuộc về C nhãn (label) khác nhau với $d > C > 2$, tức là $\forall \mathbf{x} \in \mathbf{X}$, label$(\mathbf{x}) \in \{1, 2, \ldots, C\}$. Đặt:

$$\mathbf{X}_k = \{\mathbf{x} \in \mathbf{X}; \quad \text{label}(\mathbf{x}) = k\}; \quad N_k = |\mathbf{X}_k|$$

- Cần giảm dữ liệu về $d' < d$ chiều - bằng một phép biến đổi tuyến tính. Kết quả dữ liệu tương ứng với $\mathbf{x}$ sẽ có dạng

$$\bar{x} = \mathbf{W}^T\mathbf{x} \quad \text{trong đó} \quad \mathbf{W} \in \mathbb{R}^{d \times d'}.$$

LDA for Multinomial Data

Ký kiệu:

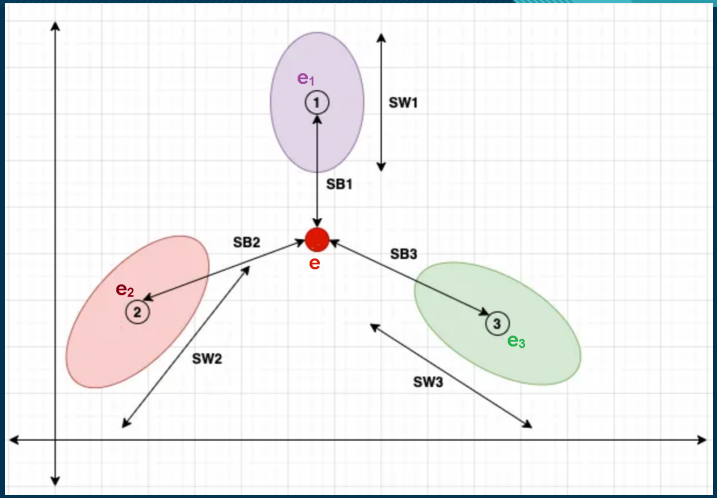
$$\mathbf{m}_k = \frac{1}{N_k} \sum_{x_n \in \mathbf{X}_k} \mathbf{x}_n \in \mathbb{R}^d; \qquad \mathbf{e}_k = \frac{1}{N_k} \sum_{x_n \in \mathbf{X}_k} \bar{x}_n = \mathbf{W}^T \mathbf{m}_k \in \mathbb{R}^{d'}$$

$$\bar{\mathbf{X}}_k = \{\bar{x} = \mathbf{W}\mathbf{x} : \mathbf{x} \in \mathbf{X}_k\} = \mathbf{W}^T \mathbf{X}_k; \qquad \mathbf{e} = \frac{1}{N} \sum_{n=1}^{N} \bar{x}_n = \mathbf{W}^T \mathbf{m}$$

$$\mathbf{m} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n; \qquad \mathbf{E}_k = \mathbf{e}_k \mathbf{1}_k^T = \underbrace{(\mathbf{e}_k \quad \mathbf{e}_k \dots \mathbf{e}_k)}_{N_k - columns}$$

$$\mathbf{M}_k = \mathbf{m}_k \mathbf{1}_k^T = \underbrace{(\mathbf{m}_k \quad \mathbf{m}_k \dots \mathbf{m}_k)}_{N_k - columns}; \qquad \mathbf{E}|_k = \mathbf{e}\mathbf{1}_k^T = \underbrace{(\mathbf{e} \quad \mathbf{e} \dots \mathbf{e})}_{N_k - columns}$$

LDA for Multinomial Data



LDA for Multinomial Data

- Như trên hình minh họa, tương tự trường hợp 2 nhãn (binary), chúng ta muốn hình chiếu thông qua $\mathbf{W}$ có tính chất
  - Các **within-class variance** SW_k nhỏ (xem hình minh họa)
  - Các **between-class variance** SB_k lớn (xem hình minh họa)
- Các within-class variance có thể xác định bởi

$$SW_k^2 = \sum_{x_n \in \mathbf{X}_k} ||\bar{x}_n - \mathbf{e}_k||_2^2 = ||\bar{\mathbf{X}}_k - \mathbf{E}_k||_F^2.$$

- Các between-class variance có thể xác định qua

$$SB^2 = \sum_{k=1}^{C} N_k ||\mathbf{e}_k - \mathbf{e}||_2^2 = \sum_{k=1}^{C} ||\mathbf{E}_k - \mathbf{E}|_k||_F^2$$

LDA for Multinomial Data: Within-class variance

Within-class variance

$$\begin{aligned} SW_k^2 &= ||\bar{\mathbf{X}}_k - \mathbf{E}_k||_F^2 = ||\mathbf{W}^T(\mathbf{X}_k - \mathbf{M}_k)||_F^2 \\ &= \text{trace}\left(\mathbf{W}^T(\mathbf{X}_k - \mathbf{M}_k)(\mathbf{X}_k - \mathbf{M}_k)^T\mathbf{W}\right) \end{aligned}$$

Tương tự trường hợp binary, tổng **within-class variance**

$$\begin{aligned} \Sigma_W = \sum_{k=1}^{C} SW_k^2 &= \sum_{k=1}^{C} \text{trace}\left(\mathbf{W}^T(\mathbf{X}_k - \mathbf{M}_k)(\mathbf{X}_k - \mathbf{M}_k)^T\mathbf{W}\right) \\ &= \text{trace}\left(\mathbf{W}^T\mathbf{S}_W\mathbf{W}\right) \end{aligned}$$

Trong đó: $\mathbf{S}_W = \sum_{k=1}^{C} ||\mathbf{X}_k - \mathbf{M}_k||_F^2 = \sum_{k=1}^{C} \sum_{x_n \in \mathbf{X}_k} (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^T$ là ma trận đối xứng, nửa xác định dương (tổng của C ma trận như vậy).

LDA for Multinomial Data: Between-class variance

Between-class variance Có thể định nghĩa bởi

$$SB^2 = \sum_{k=1}^{C} N_k \|\mathbf{e}_k - \mathbf{e}\|_2^2 = \sum_{k=1}^{C} \|\mathbf{E}_k - \mathbf{E}|_k\|_F^2$$

Tương tự trường hợp SW, ký hiệu $\mathbf{M}|_k$ tương tự $\mathbf{E}|_k$, ta có

$$SB^2 = \sum_{k=1}^{C} \|\mathbf{E}_k - \mathbf{E}|_k\|_F^2 = \sum_{k=1}^{C} \text{trace}\left(\mathbf{W}^T(\mathbf{M}_k - \mathbf{M}|_k)(\mathbf{M}_k - \mathbf{M}|_k)^T\mathbf{W}\right)$$

$$= \text{trace}\left\{\mathbf{W}^T \underbrace{\left[\sum_{k=1}^{C}(\mathbf{M}_k - \mathbf{M}|_k)(\mathbf{M}_k - \mathbf{M}|_k)^T\right]}_{\mathbf{S}_B} \mathbf{W}\right\} = \text{trace}\left(\mathbf{W}^T\mathbf{S}_B\mathbf{W}\right)$$

LDA for Multinomial Data: Loss function

Chú ý

$$\mathbf{S}_B = \sum_{k=1}^{C} (\mathbf{M}_k - \mathbf{M}|_k)(\mathbf{M}_k - \mathbf{M}|_k)^T = \sum_{k=1}^{C} N_k (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T$$

cũng là ma trận đối xứng, nửa xác định dương.
Tương tự trường hợp Binary, bài toán tối ưu cho trường hợp MultiNomial vẫn dựa trên cực đại hóa tỷ lệ $\mathbf{S}_B/\mathbf{S}_W$:

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} J(\mathbf{W}) = \arg\max_{\mathbf{W}} \frac{\text{trace}(\mathbf{W}^T \mathbf{S}_B \mathbf{W})}{\text{trace}(\mathbf{W}^T \mathbf{S}_W \mathbf{W})} \tag{9}$$

Giải (9) qua việc tìm điểm tới hạn: $\nabla_{\mathbf{W}} J(\mathbf{W}) = 0$.

LDA for Multinomial Data: Optimization

Ta có: Với mọi ma trận $\mathbf{A} \in \mathbb{R}^{d \times d}$ đối xứng: $\nabla_{\mathbf{W}} \text{trace}(\mathbf{W}^T \mathbf{A} \mathbf{W}) = 2\mathbf{A}\mathbf{W}$.

Áp dụng các kỹ thuật tương tự trường hợp binary, thu được

$$\nabla_{\mathbf{W}} J(\mathbf{W}) = \frac{2\left[\mathbf{S}_B \mathbf{W} \text{trace}(\mathbf{W}^T \mathbf{S}_W \mathbf{W}) - \text{trace}(\mathbf{W}^T \mathbf{S}_B \mathbf{W}) \mathbf{S}_W \mathbf{W}\right]}{\left[\text{trace}(\mathbf{W}^T \mathbf{S}_W \mathbf{W})\right]^2} = \mathbf{0}$$

Suy ra

$$\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{W} = J(\mathbf{W}) \mathbf{W}$$

Tức là mỗi cột của $\mathbf{W}$ là vector riêng của $\mathbf{S}_W^{-1} \mathbf{S}_B$ ứng với trị riêng $J(\mathbf{W})$. Do đó

$$J(\mathbf{W}) \longrightarrow \max_{\mathbf{W}}$$

đồng nghĩa với nó ($J(\mathbf{W})$) là trị riêng lớn nhất của $\mathbf{S}_W^{-1} \mathbf{S}_B$.

LDA for Multinomial Data: Optimization

- Để các chiều dữ liệu trong không gian mới không phụ thuộc tuyến tính, các cột của $\mathbf{W}$ cần độc lập tuyến tính.
- Giả sử $\lambda_{\max}$ là trị riêng lớn nhất của $\mathbf{S}_W^{-1}\mathbf{S}_B$. Có tối đa bao nhiêu vector riêng độc lập tuyến tính ứng với $\lambda_{\max}$?
- Các vector độc lập tuyến tính ứng với $\lambda_{\max}$ tạo thành không gian con riêng $V_{\lambda_{\max}}$ của không gian dữ liệu $\mathbb{R}^d$ ban đầu.
- Chú ý: Số cột độc lập tuyến tính của $\mathbf{W} = \text{rank}\big(V_{\lambda_{\max}}\big)$ và

$$\text{rank}\big(V_{\lambda_{\max}}\big) \leq \text{rank}\big(\mathbf{S}_W^{-1}\mathbf{S}_B\big) \leq \min\big\{\text{rank}(\mathbf{S}_W^{-1}), \text{rank}(\mathbf{S}_B)\big\}$$

LDA for Multinomial Data: Number of Dimentions

Lemma

Hạng của $\mathbf{S}_B$ *nhỏ hơn số phân lớp* C

$$rank(\mathbf{S}_B) \leq C - 1$$

Vậy

- Số chiều của không gian mới là một số không lớn hơn $C - 1$.
- Cơ sở mới cho bài toán multi-class LDA là các vector riêng độc lập tuyến tính ứng với trị riêng cao nhất của $\mathbf{S}_W^{-1}\mathbf{S}_B$.