Dimensionality Reduction

Ziping Zhao

School of Information Science and Technology
ShanghaiTech University, Shanghai, China

CS182: Introduction to Machine Learning (Fall 2022)
http://cs182.sist.shanghaitech.edu.cn

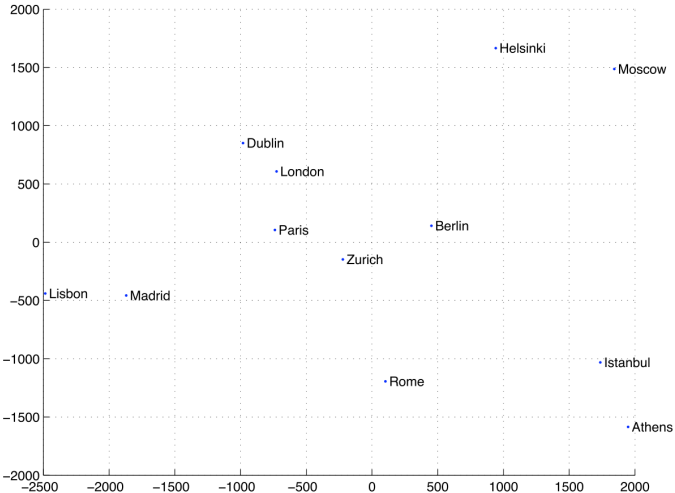
Ch. 6 of I2ML (Secs. 6.4, 6.6, and 6.12 – 6.13 excluded)

Outline

Multidimensional Scaling

- Problem formulation:
  - Given the pairwise distances between pairs of points in some space (but the exact coordinates of the points and their dimensionality are unknown).
  - We want to embed the points in a lower-dimensional space (e.g., two-dimensional space) such that the pairwise Euclidean distances in this space are as close as possible to those in the original space.
- The projection to the lower-dimensional space is not unique because the pairwise distances are invariant to such operations as translation, rotation, and reflection.
- MDS is closely related to the Euclidean distance matrix (EDM) problem.

MDS Embedding of Cities



Derivation – I

- Sample $\mathcal{X} = \{\mathbf{x}^t \in \mathbb{R}^d\}_{t=1}^N$ which is not available as feature vectors
- Squared Euclidean distance between points r and s:

$$\begin{aligned} d_{rs}^2 = \|\mathbf{x}^r - \mathbf{x}^s\|^2 &= \sum_{j=1}^{d} (x_j^r - x_j^s)^2 \\ &= \sum_{j=1}^{d} (x_j^r)^2 + \sum_{j=1}^{d} (x_j^s)^2 - 2\sum_{j=1}^{d} x_j^r x_j^s \\ &= b_{rr} + b_{ss} - 2b_{rs} \end{aligned} \tag{1}$$

where

$$b_{rs} = \sum_{j=1}^{d} x_j^r x_j^s = (\mathbf{x}^r)^T \mathbf{x}^s \qquad \text{(dot product of } \mathbf{x}^r \text{ and } \mathbf{x}^s\text{)}$$

or in matrix form with $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^N]$:

$$\mathbf{B} = \mathbf{X}^T \mathbf{X}$$

Derivation – II

- Centering of data to constrain the solution:

$$\sum_{t=1}^{N} x_j^t = 0, \ \forall j = 1, \ldots, d$$

- Summing up equation (1) on r, s, and both r and s, we get

$$\sum_{r=1}^{N} d_{rs}^2 = T + Nb_{ss}, \quad \sum_{s=1}^{N} d_{rs}^2 = Nb_{rr} + T, \quad \sum_{r=1}^{N}\sum_{s=1}^{N} d_{rs}^2 = 2NT$$

where

$$T = \sum_{t=1}^{N} b_{tt} = \sum_{t=1}^{N}\sum_{j=1}^{d} (x_j^t)^2$$

Derivation – III

- By defining

$$d_{*s}^2 = \frac{1}{N}\sum_r d_{rs}^2 \qquad d_{r*}^2 = \frac{1}{N}\sum_s d_{rs}^2 \qquad d_{**}^2 = \frac{1}{N^2}\sum_r\sum_s d_{rs}^2$$

  and using equation (1), we get

$$b_{rs} = \frac{1}{2}(d_{r*}^2 + d_{*s}^2 - d_{**}^2 - d_{rs}^2)$$

- We have obtained the form of matrix $\mathbf{B}$.
- $\mathbf{B} = \mathbf{X}^T\mathbf{X}$ is p.s.d. with positive eigenvalues, so it can be expressed as its spectral decomposition:

$$\mathbf{B} = \mathbf{C}\mathbf{D}\mathbf{C}^T = \mathbf{C}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{C}^T = (\mathbf{C}\mathbf{D}^{1/2})(\mathbf{C}\mathbf{D}^{1/2})^T$$

  where $\mathbf{C}$ is the matrix whose columns are the eigenvectors of $\mathbf{B}$ and $\mathbf{D}^{1/2}$ is the diagonal matrix whose diagonal elements are the square roots of the eigenvalues.

Projection to Lower-Dimensional Space

- If we ignore the eigenvectors of $\mathbf{B}$ with very small eigenvalues (the eigenvalues of $\mathbf{B} = \mathbf{X}^T\mathbf{X}$ are the same as the eigenvalues of $\mathbf{X}\mathbf{X}^T$) and keep the largest k ones, $\mathbf{C}\mathbf{D}^{1/2}$ will only be a low-rank approximation of $\mathbf{X}^T$.
- Let $\mathbf{c}_j \in \mathbb{R}^N$ be the k eigenvectors chosen with corresponding eigenvalues λ_j.
- New dimensions in k-dimensional embedding space:

$$z_j^t = \sqrt{\lambda_j} c_j^t, \quad j = 1, \ldots, k, \ t = 1, \ldots, N$$

So the new coordinates of instance t are given by the t-th elements of the eigenvectors after normalization, i.e.,

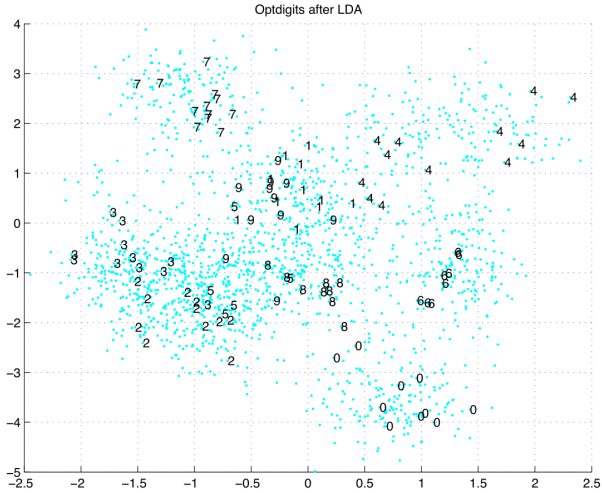
$$[\sqrt{\lambda_1} c_1^t, \ldots, \sqrt{\lambda_k} c_k^t]^T \in \mathbb{R}^k$$

Outline

Linear Discriminant Analysis

- Unlike PCA, FA, and MDS, LDA is a supervised dimensionality reduction method.
- LDA is typically used with a classifier for classification problems.
- Goal: the classes are well-separated after projecting to a low-dimensional space by utilizing the label information (output information).

Example



2-Class Case

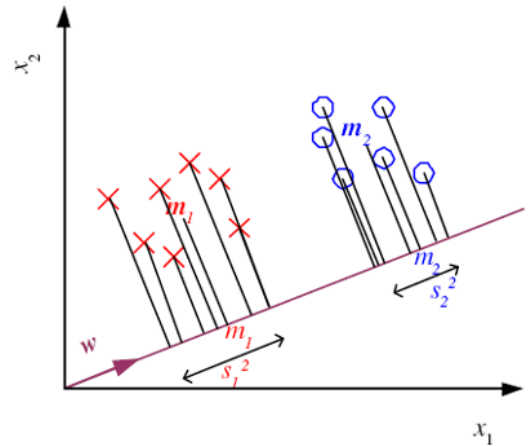
- Given sample $\mathcal{X} = \{(\mathbf{x}^t, r^t)\}$, where $r^t = 1$ if $\mathbf{x}^t \in C_1$ (class 1) and $r^t = 0$ if $\mathbf{x}^t \in C_2$ (class 2).
- Find vector $\mathbf{w}$ on which the data are projected such that the examples from C_1 and C_2 are as well separated as possible.
- Projection of $\mathbf{x}$ onto $\mathbf{w}$ (dimensionality reduced from d to 1):

$$z = \mathbf{w}^T \mathbf{x}$$

- $\mathbf{m}_i \in \mathbb{R}^d$ and $m_i \in \mathbb{R}$ are sample means of C_i before and after projection:

$$m_1 = \frac{\sum_t \mathbf{w}^T \mathbf{x}^t r^t}{\sum_t r^t} = \mathbf{w}^T \mathbf{m}_1$$
$$m_2 = \frac{\sum_t \mathbf{w}^T \mathbf{x}^t (1 - r^t)}{\sum_t (1 - r^t)} = \mathbf{w}^T \mathbf{m}_2$$

Projection



Between-Class Scatter

- Between-class scatter (in a form of normalized sample variance/covariance matrix):

$$\begin{aligned}(m_1 - m_2)^2 =& (\mathbf{w}^T\mathbf{m}_1 - \mathbf{w}^T\mathbf{m}_2)^2 \\ =& \mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{w} \\ =& \mathbf{w}^T\mathbf{S}_B\mathbf{w}\end{aligned}$$

where the between-class scatter matrix

$$\begin{aligned}\mathbf{S}_B =& (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \\ =& \frac{4}{N_1 + N_2}(N_1(\mathbf{m}_1 - \mathbf{m})(\mathbf{m}_1 - \mathbf{m})^T + N_2(\mathbf{m}_2 - \mathbf{m})(\mathbf{m}_2 - \mathbf{m})^T)\end{aligned}$$

where $\mathbf{m}$ is the mean of the class means, i.e., $\mathbf{m} = \frac{1}{2}(\mathbf{m}_1 + \mathbf{m}_2)$, and N_1 and N_2 are the # of instances in C_1 and C_2, respectively.

Within-Class Scatter

- Within-class scatter:

$$\begin{aligned} s_1^2 &= \sum_t (\mathbf{w}^T \mathbf{x}^t - m_1)^2 r^t \\ &= \sum_t \mathbf{w}^T (\mathbf{x}^t - \mathbf{m}_1)(\mathbf{x}^t - \mathbf{m}_1)^T \mathbf{w} r^t \\ &= \mathbf{w}^T \mathbf{S}_1 \mathbf{w} \end{aligned}$$

  where the within-class scatter matrix $\mathbf{S}_1 = \sum_t (\mathbf{x}^t - \mathbf{m}_1)(\mathbf{x}^t - \mathbf{m}_1)^T r^t$. Also,

$$s_2^2 = \mathbf{w}^T \mathbf{S}_2 \mathbf{w}$$

  with $\mathbf{S}_2 = \sum_t (\mathbf{x}^t - \mathbf{m}_2)(\mathbf{x}^t - \mathbf{m}_2)^T (1 - r^t)$.

- So, the total within-class scatter

$$s_1^2 + s_2^2 = \mathbf{w}^T \mathbf{S}_W \mathbf{w}$$

  where

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

Fisher's Linear Discriminant

- Fisher's linear discriminant refers to the vector $\mathbf{w}$ that maximizes the Fisher criterion (a.k.a. generalized Rayleigh quotient):

$$\underset{\mathbf{w}}{\text{maximize}} \quad J(\mathbf{w}) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2} = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

  which is equivalent to solve

$$\begin{aligned} \underset{\mathbf{w}}{\text{maximize}} \quad & \mathbf{w}^T \mathbf{S}_B \mathbf{w} \\ \text{subject to} \quad & \mathbf{w}^T \mathbf{S}_W \mathbf{w} = 1 \end{aligned}$$

- We can prove the optimal solution satisfies the following generalized eigenvalue problem:

$$\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}$$

  or, if $\mathbf{S}_W$ is nonsingular, an equivalent eigenvalue problem:

$$\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w} = \lambda \mathbf{w}$$

2-Class Case

- For the 2-class case, we note that

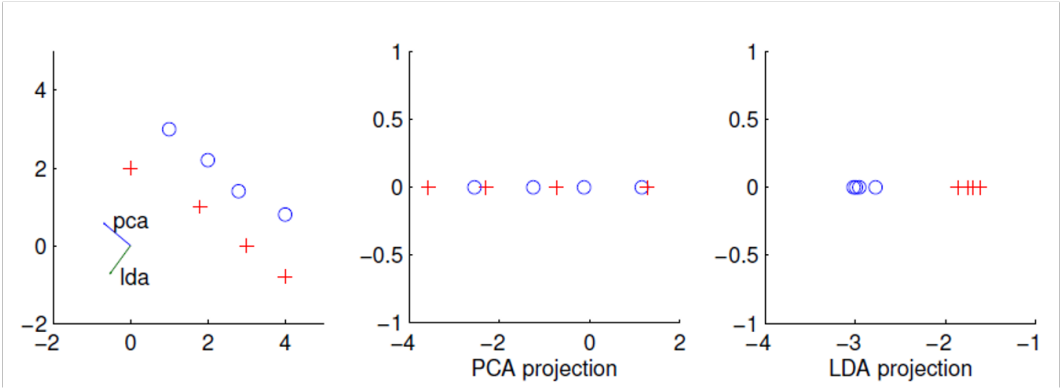
$$\mathbf{S}_B\mathbf{w} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{w} = c(\mathbf{m}_1 - \mathbf{m}_2)$$

  for some constant c and hence $\mathbf{S}_B\mathbf{w}$ (also $\mathbf{S}_W\mathbf{w}$) is in the same direction of $\mathbf{m}_1 - \mathbf{m}_2$.
- So we get

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

  where c is some irrelevant constant factor.
- We have projected the samples from d dimensions to 1, i.e., dimensionality reduction, and any classification method can be used afterward.
- Recall that when $p(\mathbf{x}|C_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ (i.e., homoscedasticity), we have a linear discriminant where $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, and we see that Fisher's linear discriminant is optimal if the classes are normally distributed. But Fisher's linear discriminant can be used even when the classes are not normal.

PCA vs. LDA



$K > 2$ Classes – I

- Find the matrix $\mathbf{W} \in \mathbb{R}^{d \times k}$ such that

$$\mathbf{z} = \mathbf{W}^T \mathbf{x} \in \mathbb{R}^k$$

- Within-class scatter matrix for C_i:

$$\mathbf{S}_i = \sum_t r_i^t (\mathbf{x}^t - \mathbf{m}_i)(\mathbf{x}^t - \mathbf{m}_i)^T$$

  where $r_i^t = 1$ if $\mathbf{x}^t \in C_i$ and 0 otherwise.

- Total within-class scatter matrix:

$$\mathbf{S}_W = \sum_{i=1}^K \mathbf{S}_i$$

$K > 2$ Classes – II

- Between-class scatter matrix:

$$\mathbf{S}_B = \sum_{i=1}^{K} N_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T$$

where $\mathbf{m}$ is the overall mean (i.e., mean of the class means) or sometimes chosen as the mean weighted by sample numbers and $N_i = \sum_t r_i^t$.

- We need to find the matrix $\mathbf{W}$ that maximizes

$$\underset{\mathbf{W}}{\text{maximize}} \quad J(\mathbf{W}) = \frac{\det(\mathbf{W}^T \mathbf{S}_B \mathbf{W})}{\det(\mathbf{W}^T \mathbf{S}_W \mathbf{W})}$$

which corresponds to the eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$ with the largest eigenvalues.

- Take $k \leq K-1$: since $\mathbf{S}_B$ is the sum of K rank-1 matrices and only $K-1$ of them are independent, $\mathbf{S}_B$ has a maximum rank of $K-1$.