MDS-FDA: Extending Fisher-Discriminant Analysis for Non-vector Data*

Author: Jong-Hyun Won

Introduction

This short article explains how to extend the Fisher-Discriminant Aanlysis (FDA) for non-vector data. The Fisher-Discriminant Analysis (FDA) is a classical dimensionality reduction algorithm using variances of data with classes. Using the separation obtained from covariance information, this algorithm has been utilized heavily in various applications, particularly those using standard vector data. However, there is little information about whether this algorithm directly use non-vector data as well. In this short article, we show the algorithm can be applied to non-vector data directly, by utilizing the multidimensional scaling (MDS) with pairwise distances of non-euclidean data. We refer to this algorithm as the MDS-FDA.

For the rest of this article, we first review the original Fisher-Discriminant Analysis (FDA) and explain how can we apply the FDA algorithm to non-vector data.

The Fisher-Discriminant Analysis (FDA). The Fisher-Discriminant Analysis (FDA) is a linear dimensionality reduction algorithm using class-covariances and the variance of whole data. For a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ with $\mathbf{x}_i \in \mathbb{R}^D$ and labels $y_i \in \{1, \cdots, c\}$, the FDA finds the projection matrix $W \in \mathbb{R}^{D\times k}$ with $k \leq c-1$ by solving the following objective funtion:

$$\mathcal{J}(W) := \max_{W\in\mathbb{R}^{D\times k}} \frac{tr(W^\top S_b W)}{tr(W^\top S_w W)}, \tag{1}$$

where the matrix $S_b \in \mathbb{R}^{D\times D}$ is the variance of the means of data with classes:

$$S_b = \sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^\top, \tag{2}$$

with the number of i-th class data N_i, mean of i-th class data μ_i, and the mean of whole data μ. The matrix S_w is the sum of each class conditioned covariance matrix $S_j \in \mathbb{R}^{D\times D}$:

$$S_w = \sum_{j=1}^{c} S_j. \tag{3}$$

Thus, what the objective function $\mathcal{J}(W)$ does is simple as follows; It attempts to find the matrix W maximizing the means of each class means (i.e., separating out each class means) and, at the same time, minimizing the within-variances of class data.

The closed-form solution. The solution of the maximization problem of $\mathcal{J}(W)$ is obtained by a closed-form expression. To show this, first, we show the objective $\mathcal{J}(W)$ can be rewritten as:

$$\mathcal{J}(W) = \max_{W} \frac{tr(W^\top S_b W)}{tr(W^\top S W)}, \tag{4}$$

with the global covariance $S = S_b + S_w$. Then, by using the fact that it is a Rayleigh quotient form, we can find the optimal solution with a generalized eigenvalue problem:

$$S_b W = S W \Lambda, \tag{5}$$

with the eigenvalue-diagonal matrix $\Lambda \in \mathbb{R}^{D \times D}$. When S_b is invertible, the solution is found by the eigenvalue problem:

$$S^{-1} S_b W = W \Lambda. \tag{6}$$

The optimal solution $W^* \in \mathbb{R}^{D \times k}$ is composed of the $k \leq c-1$ eigenvectors corresponding to the k largest eigenvalues. Finally, using the optimal W^* and the data matrix $X \in \mathbb{R}^{N \times D}$, we obtain the following linear subspace $U \in \mathbb{R}^{N \times k}$ by projecting X onto W^*:

$$U = X W^* \in \mathbb{R}^{N \times k}. \tag{7}$$

Because the FDA algorithm uses variances of data which can be expressed as inner products of data, kernelizing this algorithm is also applicable and well-known. By its simple interpretation and extension with kernels, the FDA has been recognized as standard dimensionality reduction method in literature, but there has not been dealt with thoroughly whether this algorithm can be used for non-vector data. In the following, we show this can be done by using the gram matrix in multidimensional scaling (MDS).

The MDS-FDA algorithm for non-euclidean data embedding

Problem setup. We consider the set $\mathcal{Q} = \{(s_i, y_i)\}_{i=1}^N$, a set of non-euclidean data s_i and their classes $y_i \in \{1, 2, \cdots, c\}$. We also assume that we know the pairwise distance matrix $D \in \mathbb{R}_+^{N \times N}$ of the data where

$$D_{ij} = d(s_i, s_j), \tag{8}$$

with some metric $d(\cdot, \cdot) \in \mathbb{R}_+$. Using this information, the goal is to find the linear embedding $U \in \mathbb{R}^{N \times k}$ maximizing between-class covariance and minimizing within-class covariance of s_i, such as what FDA does.

The core idea and the derivation for the solution. We assume the matrix $X \in \mathbb{R}^{N \times D}$ as the most *similar* vector representation for $\mathcal{Q}$. When we run FDA for this X, the objective is equal to the equation (4). We want to find the projection matrix $W \in \mathbb{R}^{D \times k}$ by

$$\mathcal{J}(W) = \max_{W} \frac{tr(W^\top S_b W)}{tr(W^\top S W)}. \tag{9}$$

Here, by using the Representer theorem, we can rewrite the W as:

$$W = X^\top V, \tag{10}$$

with a coefficient matrix $V \in \mathbb{R}^{N\times k}$. Now, by substituting W as $X^\top V$ in the FDA objective function, we get the objective:

$$\mathcal{J}(V) := \max_V \frac{tr(V^\top X S_b X^\top V)}{tr(V^\top X S X^\top V)}. \tag{11}$$

Assuming the data mean μ for X is zero vector, we can rewrite the denominator of $\mathcal{J}(V)$ as:

$$\begin{aligned} tr(V^\top X S X^\top V) &= tr(V^\top X X^\top X X^\top V) && (12)\\ &= tr(V^\top G^2 V), && (13)\end{aligned}$$

with the Gram matrix $G = XX^\top$. Likewise, the numerator is

$$\begin{aligned} tr(V^\top X S_b X^\top V) &= tr\left(V^\top \sum_{i=1}^{c} N_i X \mu_i \mu_i^\top X^\top V\right) && (14)\\ &= tr\left(V^\top \sum_{i=1}^{c} N_i K^{(i)} K^{(i)\top} V\right), && (15)\end{aligned}$$

with $K^{(i)} \in \mathbb{R}^{N\times 1}$ and $K_l^{(i)} = \mathbf{x}_l^\top \mu_i = \sum_k^{N_j} \mathbf{x}_l^\top \mathbf{x}_k$. Let $M = \sum_{i=1}^c N_i K^{(i)} K^{(i)\top} \in \mathbb{R}^{N\times N}$. Then, finally, we rewrite the objective function $\mathcal{J}(V)$ as:

$$\mathcal{J}(V) = \max_V \frac{tr(V^\top M V)}{tr(V^\top G^2 V)}. \tag{16}$$

Because the M and the gram matrix G are positive-(semi)definite, we obtain the closed form solution as the following eigenvalue problem:

$$(G^2)^{-1} M V = V\Lambda \tag{17}$$

with an eigenvalue matrix $\Lambda \in \mathbb{R}^{D\times D}$, and we pick k eigenvectors with largest eigenvalues as we do in FDA. The closed-form expression in Eq. (17) clearly shows that we need the gram matrix G and M for obtaining the solution. But, how can we know them without knowing the X? In our problem (and in practice), we only access the dataset $\mathcal{Q}$ and the pairwise distance matrix D. How do we obtain G and M by using D? The answer is to derive the G and M from Multidimensional scaling of $\mathcal{Q}$. Recall that the gram matrix G is obtained by the pairwise distance D:

$$G = -\frac{1}{2} C D^2 C, \tag{18}$$

with a centering matrix C[1]. When we get to obtain the gram matrix, the matrix M can be easily obtained from G, since it is a subset of G (since G contains all inner products of data). Now we know both G and M, we find the optimal projection matrix V^* by solving the eigenvalue problem:

$$(G^2)^{-1} M V = V\Lambda, \tag{19}$$

and we get the optimal embedding $U \in \mathbb{R}^{N\times k}$:

$$\begin{aligned} U &= XW^* && (20)\\ &= XX^\top V^* && (21)\\ &= GV^* && (22)\\ &= -\frac{1}{2} C D^2 C V^*. && (23)\end{aligned}$$

[1]This is not important one, it is a matrix keeping the center of data $\mu = \mathbf{0}$.