On linear discriminant analysis

Raul Adell

May 2023

1 Introduction

This essay aims to provide help in the mathematical formulation of the method, already explained in previous essays. In order to write this essays formulas given in class are used and also supplementary material from San José State University, Math 253: Mathematical Methods for Data Visualization, Dr. Guangliang Chen.

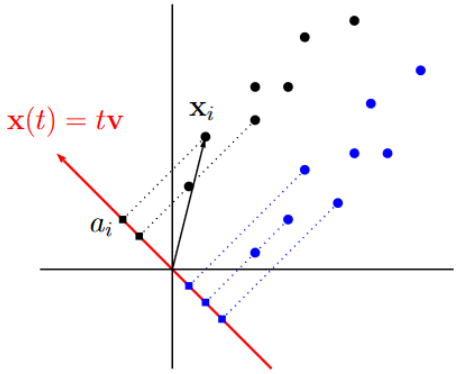
PCA aims to find the most accurate data representation in a lower dimensional space spanned by the maximum variance directions. However, such directions might not work well for tasks like classification. PCA focuses on variance (not always useful or relevant). Here we present a new data reduction method that tries to preserve the discriminatory information between different classes of the data set.

To sum up, this could be useful representatively but not discriminatively.

2 The two-class LDA problem

Given a training data set $\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^d$ consisting of two classes C_1, C_2, find a (unit-vector) direction that "best" discriminates between the two classes.

Consider any unit vector $\boldsymbol{\beta} \in R^d$ (notice that we are changing the notation with respect to the drawing, where β is v) :



First, observe that projections of the two classes onto parallel lines always have "the same amount of separation". This time we are going to focus on lines that pass through the origin.

The 1D projections of the points are

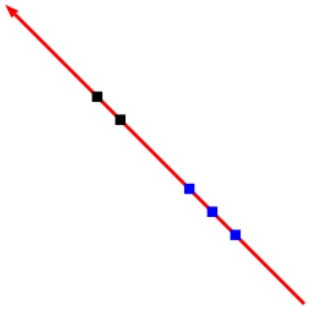
$$a_i = \boldsymbol{\beta}^T \mathbf{x}_i, \quad i = 1, \ldots, n$$

One (naive) idea is to measure the distance between the two class means in the 1D projection space: $|\mu_1 - \mu_2|$, where

$$\begin{aligned} \mu_1 &= \frac{1}{n_1} \sum_{\mathbf{x}_i \in C_1} a_i = \frac{1}{n_1} \sum_{\mathbf{x}_i \in C_1} \boldsymbol{\beta}^T \mathbf{x}_i \\ &= \boldsymbol{\beta}^T \cdot \frac{1}{n_1} \sum_{\mathbf{x}_i \in C_1} \mathbf{x}_i = \boldsymbol{\beta}^T \mathbf{m}_1 \end{aligned}$$

$$\mu_2 = \boldsymbol{\beta}^T \mathbf{m}_2, \quad \mathbf{m}_2 = \frac{1}{n_2} \sum_{\mathbf{x}_i \in C_2} \mathbf{x}_i$$

Thus, the data is currently like this:



It turns out that in order to achieve maximum spearability we should also pay attention to the variances of the projected classes. This way we want the data maximally separated (in terms of the means) while keeping minimal the overlapping.

$$s_1^2 = \sum_{\mathbf{x}_i \in C_1} (a_i - \mu_1)^2, \quad s_2^2 = \sum_{\mathbf{x}_i \in C_2} (a_i - \mu_2)^2$$

This can be achieved through the following modified formulation:

$$\max_{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|=1} \frac{(\mu_1 - \mu_2)^2}{s_1^2 + s_2^2}$$

3 Mathematical formulation

First, we derive a formula for the distance between the two projected centroids:

$$\begin{aligned}(\mu_1 - \mu_2)^2 &= \left(\boldsymbol{\beta}^T \mathbf{m}_1 - \boldsymbol{\beta}^T \mathbf{m}_2\right)^2 = \left(\boldsymbol{\beta}^T (\mathbf{m}_1 - \mathbf{m}_2)\right)^2 \\ &= \boldsymbol{\beta}^T (\mathbf{m}_1 - \mathbf{m}_2) \cdot (\mathbf{m}_1 - \mathbf{m}_2)^T \boldsymbol{\beta} \\ &= \boldsymbol{\beta}^T \mathbf{S}_b \boldsymbol{\beta},\end{aligned}$$

where

$$\mathbf{S}_b = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \in R^{d \times d}$$

is called the between-class scatter matrix.

Remark: Clearly, $\mathbf{S}_b$ is square, symmetric and positive semidefinite. Moreover, $\text{rank}(\mathbf{S}_b) = 1$, which implies that it only has 1 positive eigenvalue!

This can be shown as following. Let $\mathbf{X}$ be the input data matrix of size n x p, where n is the number of samples and p is the number of features. Let $\mathbf{Y}$ be the target variable vector of size n x 1, where Y_i denotes the class label of the $i-th$ sample. The between-class scatter matrix $\mathbf{S}_b$ is defined as:

$$\mathbf{S}_b = \frac{1}{K} \sum_{k=1}^{K} (\mathbf{m_k} - \mathbf{m})(\mathbf{m_k} - \mathbf{m})^T$$

where K is the number of classes, $\mathbf{m_k}$ is the mean vector of class k and $\mathbf{m}$ is the overall mean vector.

If we expand the above equation, we get:

$$\mathbf{S}_b = \frac{1}{K} \sum_{k=1}^{K} (\mathbf{m_k}\mathbf{m_k}^T - \mathbf{m}\mathbf{m_k}^T - \mathbf{m_k}\mathbf{m}^T + \mathbf{m}\mathbf{m}^T)$$

Notice that the first term $\mathbf{m_k}\mathbf{m_k}^T$ is a rank-1 matrix because it can be written as the outer product of two vectors of size p x 1. Notice that the columns

of the matrix ($\mathbf{m_k}$ - $\mathbf{m}$) ($\mathbf{m_k}$ - $\mathbf{m}$)^T are all multiples of the vector ($\mathbf{m_k}$ - $\mathbf{m}$), and therefore are linearly dependent. In fact, any column can be obtained as a scalar multiple of any other column. This implies that the rank of the matrix is at most 1. Similarly, the other terms are also rank-1 matrices. Therefore, the between-class scatter matrix $\mathbf{S}_b$ can be written as the sum of K rank-1 matrices. Since the rank of a sum of matrices is at most the sum of their ranks, it follows that the rank of $\mathbf{S}_b$ is at most $K-1$.

In the case of two classes, the rank of $\mathbf{S}_b$ is at most 1 because there are only two terms in the sum. In the case of more than two classes, the rank of $\mathbf{S}_b$ is at most K-1, but since K is greater than or equal to 2, the rank of $\mathbf{S}_b$ is still at most 1.

Next, for each class $j = 1, 2$, the variance of the projection (onto $\boldsymbol{\beta}$) is:

$$\begin{aligned}
s_j^2 &= \sum_{\mathbf{x}_i \in C_j} (a_i - \mu_j)^2 \\
&= \sum_{\mathbf{x}_i \in C_j} \left(\boldsymbol{\beta}^T \mathbf{x}_i - \boldsymbol{\beta}^T \mathbf{m}_j\right)^2 \\
&= \sum_{\mathbf{x}_i \in C_j} \boldsymbol{\beta}^T (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T \boldsymbol{\beta} \\
&= \boldsymbol{\beta}^T \left[\sum_{\mathbf{x}_i \in C_j} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T \right] \boldsymbol{\beta} \\
&= \boldsymbol{\beta}^T \mathbf{S}_j \boldsymbol{\beta}
\end{aligned}$$

where

$$\mathbf{S}_j = \sum_{\mathbf{x}_i \in C_j} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T \in R^{d \times d}$$

is called the within-class scatter matrix for class j.

The total within-class scatter of the two classes in the projection space is

$$s_1^2 + s_2^2 = \mathbf{v}^T \mathbf{S}_1 \boldsymbol{\beta} + bm\beta^T \mathbf{S}_2 \boldsymbol{\beta} = \boldsymbol{\beta}^T (\mathbf{S}_1 + \mathbf{S}_2) \boldsymbol{\beta} = \boldsymbol{\beta}^T \mathbf{S}_w \boldsymbol{\beta}$$

where

$$\mathbf{S}_w = \mathbf{S}_1 + \mathbf{S}_2 = \sum_{\mathbf{x}_i \in C_1} (\mathbf{x}_i - \mathbf{m}_1)(\mathbf{x}_i - \mathbf{m}_1)^T + \sum_{\mathbf{x}_i \in C_2} (\mathbf{x}_i - \mathbf{m}_2)(\mathbf{x}_i - \mathbf{m}_2)^T$$

is called the total within-class scatter matrix of the original data.

Remark: $\mathbf{S}_w \in R^{d \times d}$ is also square, symmetric, and positive semidefinite.

Putting everything together, we have derived the following optimization problem:

$$\max_{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|=1} \frac{\boldsymbol{\beta}^T \mathbf{S}_b \boldsymbol{\beta}}{\boldsymbol{\beta}^T \mathbf{S}_w \boldsymbol{\beta}} \longleftarrow \text{ Where did we see this? Sound familiar?}$$

Theorem: Suppose $\mathbf{S}_w$ is nonsingular. The maximizer of the problem is given by the largest eigenvector $\boldsymbol{\beta}_1$ of $\mathbf{S}_w^{-1}\mathbf{S}_b$, i.e.,

$$\mathbf{S}_w^{-1}\mathbf{S}_b\boldsymbol{\beta}_1 = \lambda_1\boldsymbol{\beta}_1$$

Proof. Let $f(\beta) = \frac{\beta^T S_b \beta}{\beta^T S_w \beta}$. We need to maximize $f(\beta)$ subject to $\beta^T S_w \beta = 1$. Using the method of Lagrange multipliers, we introduce a Lagrange multiplier λ and form the function

$$F(\beta, \lambda) = \frac{\beta^T S_b \beta}{\beta^T S_w \beta} - \lambda(\beta^T S_w \beta - 1).$$

Taking the derivative with respect to β and setting it equal to zero gives

$$S_w^{-1} S_b \beta = \lambda \beta.$$

Thus, β is an eigenvector of $S_w^{-1} S_b$ and λ is the corresponding eigenvalue. Note that since S_w is nonsingular, S_w^{-1} exists.

We want to find the eigenvector corresponding to the largest eigenvalue, so we take $\beta = \beta_1$, the eigenvector corresponding to the largest eigenvalue λ_1. Then, we have

$$S_w^{-1} S_b \beta_1 = \lambda_1 \beta_1.$$

Therefore, β_1 is the maximizer of $f(\beta)$, subject to the constraint $\beta^T S_w \beta = 1$.

To show that β_1 is the unique maximizer, suppose that there exists another maximizer $\tilde{\beta}$ such that $\tilde{\beta} \neq \beta_1$. Then, we have

$$\frac{\tilde{\beta}^T S_b \tilde{\beta}}{\tilde{\beta}^T S_w \tilde{\beta}} = \frac{\beta_1^T S_b \beta_1}{\beta_1^T S_w \beta_1} = \lambda_1.$$

Multiplying both sides by $\tilde{\beta}^T S_w \tilde{\beta}$ and using the fact that $\tilde{\beta}^T S_w \tilde{\beta} = 1$ gives

$$\tilde{\beta}^T S_b \tilde{\beta} = \lambda_1 \tilde{\beta}^T S_w \tilde{\beta}.$$

This means that $\tilde{\beta}$ is also an eigenvector of $S_w^{-1} S_b$ corresponding to the eigenvalue λ_1, which contradicts the fact that β_1 is the eigenvector corresponding to the largest eigenvalue. Therefore, β_1 is the unique maximizer of $f(\beta)$ subject to the constraint $\beta^T S_w \beta = 1$. □

Lets make a brief summary of how the problem is formulated and solved. Rayleigh's quotient is used to define the optimization problem that Fisher's LDA solves. The problem involves finding the eigenvectors of a matrix involving both

within-class and between-class covariance matrices. The eigenvector with the largest eigenvalue is used as the projection vector that maximizes the ratio of between-class variance to within-class variance.

We describe briefly the mathematical steps that we would have to follow in order to proof this. To obtain the solution, we first derive the expression for the Rayleigh quotient and show that it can be expressed in terms of the eigenvectors of the generalized eigenvalue problem. We then use the Lagrange multiplier method to optimize the Rayleigh quotient subject to the constraint that the eigenvectors are orthonormal. This leads to a generalized eigenvalue problem, which can be solved to obtain the eigenvectors and eigenvalues.

Finally, we show that the first eigenvector of the generalized eigenvalue problem corresponds to the direction in which the data is most separable by computing the projection of the data onto this direction and showing that it maximizes the class separability criterion.

Remark: $\operatorname{rank}\left(\mathbf{S}_w^{-1}\mathbf{S}_b\right) = \operatorname{rank}\left(\mathbf{S}_b\right) = 1$, so λ_1 is the only nonzero (positive) eigenvalue that can be found. It represents the the largest amount of separation between the two classes along any single direction.

Several ways of rewriting the problem may give us some intuition on how to solve it more efficiently.

- A slight better way: Rewrite as a generalized eigenvalue problem

$$\mathbf{S}_b\boldsymbol{\beta}_1 = \lambda_1\mathbf{S}_w\boldsymbol{\beta}_1,$$

- The smartest way is to rewrite as

$$\begin{aligned}\lambda_1\boldsymbol{\beta}_1 &= \mathbf{S}_w^{-1}\underbrace{\left(\mathbf{m}_1-\mathbf{m}_2\right)\left(\mathbf{m}_1-\mathbf{m}_2\right)^T}_{\mathbf{S}_b}\boldsymbol{\beta}_1 \\ &= \mathbf{S}_w^{-1}\left(\mathbf{m}_1-\mathbf{m}_2\right)\cdot\underbrace{\left(\mathbf{m}_1-\mathbf{m}_2\right)^T\boldsymbol{\beta}_1}_{\text{scalar}}\end{aligned}$$

This implies that

$$\boldsymbol{\beta}_1 \propto \mathbf{S}_w^{-1}\left(\mathbf{m}_1-\mathbf{m}_2\right)$$

and it can be computed from $\mathbf{S}_w^{-1}\left(\mathbf{m}_1-\mathbf{m}_2\right)$ through rescaling!

4 Multiclass extension

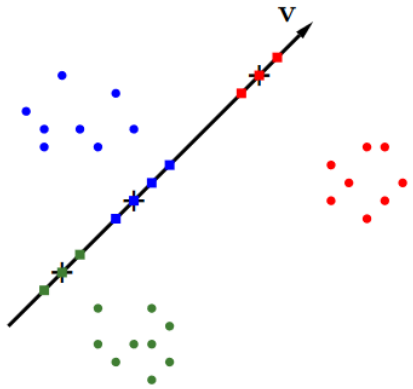
The previous procedure only applies to 2 classes. When there are $c \geq 3$ classes, what is the "most discriminatory" direction?

It will be based on the same intuition that the optimal direction $\boldsymbol{\beta}$ should project the different classes such that

- each class is as tight as possible

- their centroids are as far from each other as possible

Both are actually about variances.



For any unit vector $\boldsymbol{\beta}$, the tightness of the projected classes (of the training data) is still described by the total within-class scatter:

$$\sum_{j=1}^{c} s_j^2 = \sum \boldsymbol{\beta}^T \mathbf{S}_j \boldsymbol{\beta} = \boldsymbol{\beta}^T \left(\sum \mathbf{S}_j \right) \boldsymbol{\beta} = \boldsymbol{\beta}^T \mathbf{S}_w \boldsymbol{\beta}$$

where the $\mathbf{S}_j, 1 \leq j \leq c$ are defined in the same way as before:

$$\mathbf{S}_j = \sum_{\mathbf{x} \in C_j} (\mathbf{x} - \mathbf{m}_j) (\mathbf{x} - \mathbf{m}_j)^T$$

and $\mathbf{S}_w = \sum \mathbf{S}_j$ is the total within-class scatter matrix.

To make the class centroids μ_j (in the projection space) as far from each other as possible, we can just maximize the variance of the centroids set $\{\mu_1, \ldots, \mu_k\}$:

$$\sum_{j=1}^{c} (\mu_j - \bar{\mu})^2 = \frac{1}{c} \sum_{j<\ell} (\mu_j - \mu_\ell)^2, \quad \text{where} \quad \bar{\mu} = \frac{1}{c} \sum_{j=1}^{c} \mu_j \longleftarrow \text{ simple average.}$$

We actually use a weighted mean of the projected centroids to define the betweenclass scatter:

$$\sum_{j=1}^{c} n_j (\mu_j - \mu)^2, \quad \text{where} \quad \mu = \frac{1}{n} \sum_{j=1}^{c} n_j \mu_j \longleftarrow \text{weighted average}$$

because the weighted mean (μ) is the projection of the global centroid ($\mathbf{m}$) of the training data onto $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}^T \mathbf{m} = \boldsymbol{\beta}^T \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \right) = \boldsymbol{\beta}^T \left(\frac{1}{n} \sum_{j=1}^{c} n_j \mathbf{m}_j \right) = \frac{1}{n} \sum_{j=1}^{c} n_j \mu_j = \mu$$

In contrast, the simple mean does not have such a geometric interpretation:

$$\bar{\mu} = \frac{1}{c} \sum_{j=1}^{c} \mu_j = \frac{1}{c} \sum_{j=1}^{c} \boldsymbol{\beta}^T \mathbf{m}_j = \boldsymbol{\beta}^T \left(\frac{1}{c} \sum_{j=1}^{c} \mathbf{m}_j \right)$$

We simplify the between-class scatter (in the $\boldsymbol{\beta}$ space) as follows:

$$\begin{aligned} \sum_{j=1}^{c} n_j \left(\mu_j - \mu \right)^2 &= \sum n_j \left(\boldsymbol{\beta}^T \left(\mathbf{m}_j - \mathbf{m} \right) \right)^2 \\ &= \sum n_j \mathbf{v}^T \left(\mathbf{m}_j - \mathbf{m} \right) \left(\mathbf{m}_j - \mathbf{m} \right)^T \boldsymbol{\beta} \\ &= \boldsymbol{\beta}^T \left(\sum n_j \left(\mathbf{m}_j - \mathbf{m} \right) \left(\mathbf{m}_j - \mathbf{m} \right)^T \right) \boldsymbol{\beta} \\ &= \boldsymbol{\beta}^T \mathbf{S}_b \boldsymbol{\beta}. \end{aligned}$$

We have thus arrived at the same kind of problem

$$\max_{\boldsymbol{\beta}: \|\boldsymbol{\beta}\|=1} \frac{\boldsymbol{\beta}^T \mathbf{S}_b \boldsymbol{\beta}}{\boldsymbol{\beta}^T \mathbf{S}_w \boldsymbol{\beta}} \longleftarrow \frac{\sum n_j \left(\mu_j - \mu \right)^2}{\sum s_j^2}$$

Remark: When $c = 2$, it can be verified that

$$\sum_{j=1}^{2} n_j \left(\mu_j - \mu \right)^2 = \frac{n_1 n_2}{n} \left(\mu_1 - \mu_2 \right)^2, \quad \text{where} \quad \mu = \frac{1}{n} \left(n_1 \mu_1 + n_2 \mu_2 \right)$$

and

$$\sum_{j=1}^{2} n_j \left(\mathbf{m}_j - \mathbf{m} \right) \left(\mathbf{m}_j - \mathbf{m} \right)^T = \frac{n_1 n_2}{n} \left(\mathbf{m}_2 - \mathbf{m}_1 \right) \left(\mathbf{m}_2 - \mathbf{m}_1 \right)^T, \mathbf{m} = \frac{1}{n} \left(n_1 \mathbf{m}_1 + n_2 \mathbf{m}_2 \right)$$

This shows that when there are only two classes, the weighted definitions are just a scalar multiple of the unweighted definitions.

Therefore, the multiclass LDA $\sum n_j \left(\mu_j - \mu \right)^2 / \sum s_j^2$ is a natural generalization of the two-class LDA $\left(\mu_1 - \mu_2 \right)^2 / \left(s_1^2 + s_2^2 \right)$.

Regarding the computations is pretty much the same as last section. The solution is given by the largest eigenvector of $\mathbf{S}_w^{-1} \mathbf{S}_b$ (when $\mathbf{S}_w$ is nonsingular):

$$\mathbf{S}_w^{-1} \mathbf{S}_b \boldsymbol{\beta}_1 = \lambda_1 \boldsymbol{\beta}_1$$

However, the formula $\boldsymbol{\beta}_1 \propto \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$ is no longer valid:

$$\lambda_1 \boldsymbol{\beta}_1 = \mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{v}_1 = \mathbf{S}_w^{-1}\sum_j n_j(\mathbf{m}_j - \mathbf{m})\underbrace{(\mathbf{m}_j - \mathbf{m})^T\boldsymbol{\beta}_1}_{\text{scalar}}$$

So we have to find $\boldsymbol{\beta}_1$ by solving a generalized eigenvalue problem:

$$\mathbf{S}_b\boldsymbol{\beta}_1 = \lambda_1\mathbf{S}_w$$

One question may arise as a result of the proof of the first remark. In fact, it is a very significant question in terms of results. **How many discriminatory directions can we find?**

To answer this question, we just need to count the number of nonzero eigenvalues

$$\mathbf{S}_w^{-1}\mathbf{S}_b$$

since only the nonzero eigenvectors will be used as the discriminatory directions.

In order to apply the Theorem previously described, it is assumed that $\mathbf{S}_w$ is nonsingular. As a result of the proof in the first remark we have seen that

$$rank(\mathbf{S}_w) \leq c - 1$$

(where c is the number of training classes). Therefore, one can only find at most $c - 1$ discriminatory directions.

5 Comparison with PCA

Finally, a great table comparing the main features of the last two methods seen in class. A recap of some differences between both techniques is also given in the introduction.

This resource has been obtained from the lecture notes previously mentioned. Note again the change in notation in V and

Comparison between PCA and LDA

	PCA	LDA
Use labels?	no (unsupervised)	yes (supervised)
Criterion	variance	discrimination
#dimensions (k)	any	$\leq c-1$
Computing	SVD	generalized eigenvectors
Linear projection?	yes $((\mathbf{x}-\mathbf{m})^T\mathbf{V})$	yes $(\mathbf{x}^T\mathbf{V})$
Nonlinear boundary	can handle*	cannot handle