CISC 271 Class 24

Patterns – Linear Discriminant Analysis, or LDA

Text Correspondence: $\sim$

Main Concepts:

- *PCA for means of labeled data*
- *PCA and the Rayleigh quotient*
- *Linear discriminant analysis – optimization*

Sample Problem, Machine Inference: How can we separate labeled data?

Principal components analysis (PCA) is a mathematical method in linear data analysis. So far in this course, we have used PCA for dimensionality reduction. We understand that the scores of data that are provided by PCA are orthogonal vectors that span the column space of the zero-mean form of data.

How can PCA help us to analyze data that have labels? We can explore this idea by considering the labels of observations. Suppose that the i^{th} observation in our data has the label y_i. This label is categorical, meaning that y_i can take exactly one of a finite set of values.

For this class, we will suppose that each observation has a *binary* label. Following the convention in machine learning, we will use the values 1 and 2; that is, we will require that the i^{th} label is $y_i \in \{1, 2\}$.

Sir Ronald Fisher, in his 1936 paper [5] that described the Iris data set, introduced a powerful concept for managing observations that have labels. Since then, his concept has been formulated using linear algebra and the Rayleigh quotient. It is variously known as Fisher's linear discriminant and as linear discriminant analysis (LDA).

We will explore LDA by using PCA and scatter matrices. As an introduction, consider the data in Figure 24.1(A). If we perform PCA on these data, we find that there is a "preferred" coordinate frame that is centered at the mean of the data and that is aligned with the eigenvectors of the scatter matrix of the zero-mean form of the data; this is shown in Figure 24.1(B). We can project the labels onto these axes. Figure 24.1(C) shows the labels projected onto the first axis, which is also called the first loading vector, and Figure 24.1(D) shows the labels projected onto the second loading vector. For these data, the principal axis is relatively good at describing the observations and relatively poor at distinguishing the labels of the observations. We can, however, see that the second or "last" loading vector does a better job of distinguishing the labels.

© R E Ellis 2024

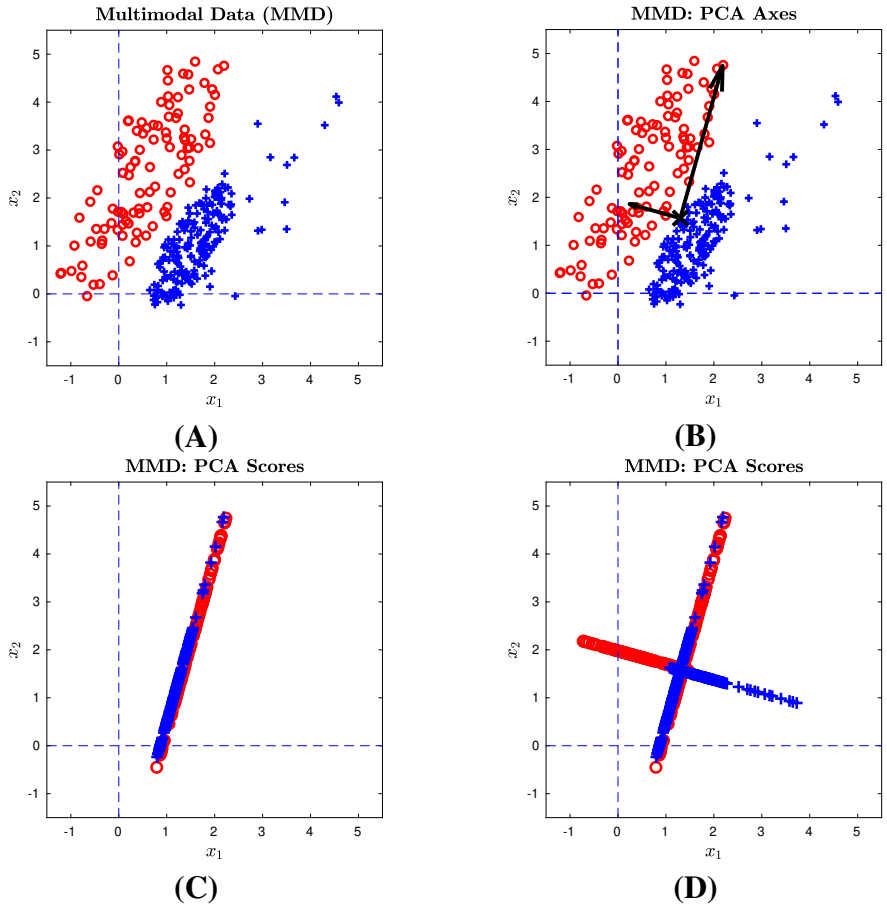

Figure 24.1: Artificially generated data with two labels. (A) The data shown as red for one label and as blue for the other label. (B) The data and PCA axes; the data mean is the black cross and the loading vectors are relatively scaled by the eigenvalues of the scatter matrix of the zero-mean form of the data. (C) Labels projected onto the first loading vector, slightly offset for visualization. (D) Labels projected onto both loading vectors, slightly offset for visualization.

24.1 Scatter Matrices For Labelled Data

We assume that our data are provided as a "tall thin" data matrix $A \in \mathbb{R}^{m \times n}$, with $m > n$. Previously, we transformed these data to a zero-mean matrix $M = A - \vec{1}^T \bar{A}$ and a symmetric positive semidefinite scatter matrix S. For this class, we will follow a convention in data analysis and write the *total* scatter as

$$S_T \stackrel{\text{def}}{=} M^T M \tag{24.1}$$

Next, we will partition the data into distinct data matrices. The observations that have label $y_j = 1$ are gathered into a matrix A_1 and the observations that have label $y_j = 2$ are gathered into

© R E Ellis 2024

a matrix A_2. If we permute the observations, the original data matrix A can be written, in terms of these partitions, as

$$A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \tag{24.2}$$

The zero-mean versions of these matrices must be calculated with care. The observations with label $y_j = 1$ have a mean observation $\bar{A}_1$ and those with label $y_j = 2$ have a mean observation $\bar{A}_2$. These means are related to the mean of the original data as

$$\bar{A} = \bar{A}_1 + \bar{A}_2 \tag{24.3}$$

The zero-mean matrices can be found from Equation 24.2 as

$$M_1 \stackrel{\text{def}}{=} A_1 - \vec{1}^T \bar{A}_1 \qquad M_2 \stackrel{\text{def}}{=} A_2 - \vec{1}^T \bar{A}_2 \tag{24.4}$$

There are four scatter matrices associated with the partitioning of A into A_1 and A_2. The first three are the *within-label* scatter; these are defined, from Equation 24.4, as

$$\begin{aligned} S_1 &\stackrel{\text{def}}{=} M_1^T M_1 \\ S_2 &\stackrel{\text{def}}{=} M_2^T M_2 \\ S_W &\stackrel{\text{def}}{=} S_1 + S_2 \end{aligned} \tag{24.5}$$

The fourth scatter matrix is the *between-label* scatter. This is the scatter of the zero-mean means, which we defined from Equation 24.3 as

$$S_B \stackrel{\text{def}}{=} \begin{bmatrix} \bar{A}_1 - \bar{A} \\ \bar{A}_2 - \bar{A} \end{bmatrix}^T \begin{bmatrix} \bar{A}_1 - \bar{A} \\ \bar{A}_2 - \bar{A} \end{bmatrix} \tag{24.6}$$

If the observations have p partitions, that is, if each label is $y_j \in \mathbb{Z}_{++} : y_j \leq p$, then Equation 24.5 and Equation 24.6 can be simply extended.

© R E Ellis 2024

24.2 Fisher's Linear Discriminant

Sir Ronald Fisher [5] observed that, for a wide variety of data, we have two potentially conflicting goals:

- Maximize the between-label scatter, and
- Minimize the within-label scatter

We will follow conventions in linear data analysis and write these goals as "argument maximum" problems. They are:

$$\vec{w}_B = \underset{\vec{u}\in\mathbb{R}:\vec{u}\neq\vec{0}}{\operatorname{argmax}} R(S_B, \vec{u}) \tag{24.7}$$

$$\vec{w}_W = \underset{\vec{u}\in\mathbb{R}:\vec{u}\neq\vec{0}}{\operatorname{argmin}} R(S_W, \vec{u}) \tag{24.8}$$

Fisher's linear discriminant elegantly combines Equation 24.7 with Equation 24.8. The concept is to maximize the *ratio* of the Rayleigh quotients.

A crucial assumption that we will make is that the within-label scatter matrix, S_W, is symmetric positive definite. This usually occurs in empirical problems and there are technical ways of managing data that do not meet this criterion. For our purposes, $S_W \succ 0$ implies that the quadratic form $\vec{u}^T S_W \vec{u}$ is always positive for a non-zero vector $\vec{u}$, and therefore that the Rayleigh quotient $R(S_W, \vec{u})$ is always non-zero for a non-zero vector argument $\vec{u}$.

Fisher's linear discriminant can be written as

$$\vec{w} = \underset{\vec{u}\in\mathbb{R}:\vec{u}\neq\vec{0}}{\operatorname{argmax}} \frac{R(S_B, \vec{u})}{R(S_W, \vec{u})} = \underset{\vec{u}\in\mathbb{R}:\vec{u}\neq\vec{0}}{\operatorname{argmax}} \frac{\vec{u}^T S_B \vec{u}}{\vec{u}^T S_W \vec{u}} \tag{24.9}$$

As derived in the extra notes for this class – provided that $S_W \succ 0$ – Equation 24.9 has the solution

$$\vec{w} = \vec{v}_{\mathrm{MAX}}(S_W^{-1} S_B) \tag{24.10}$$

Equation 24.10 provides us with the direction vector $\vec{w}$ that simultaneously optimizes the between-label scatter and the within-label scatter. Using this vector is called *linear discriminant analysis*, or LDA.

Strang [14], in pp 81–82, observes that $S_W^{-1} S_B$ is not necessarily a symmetric positive matrix and recommends using $S_W^{-1/2} S_B S_W^{-1/2}$ instead. His modification is correct but goes beyond our needs for the LDA direction vector.

© R E Ellis 2024

24.3 LDA Using Test Data

We can repeat the previous PCA tests on out artificial multimodal data. Using LDA, we can superimpose on the data the S_W axis, the S_B axis, and the LDA axis. These are shown in Figure 24.2.

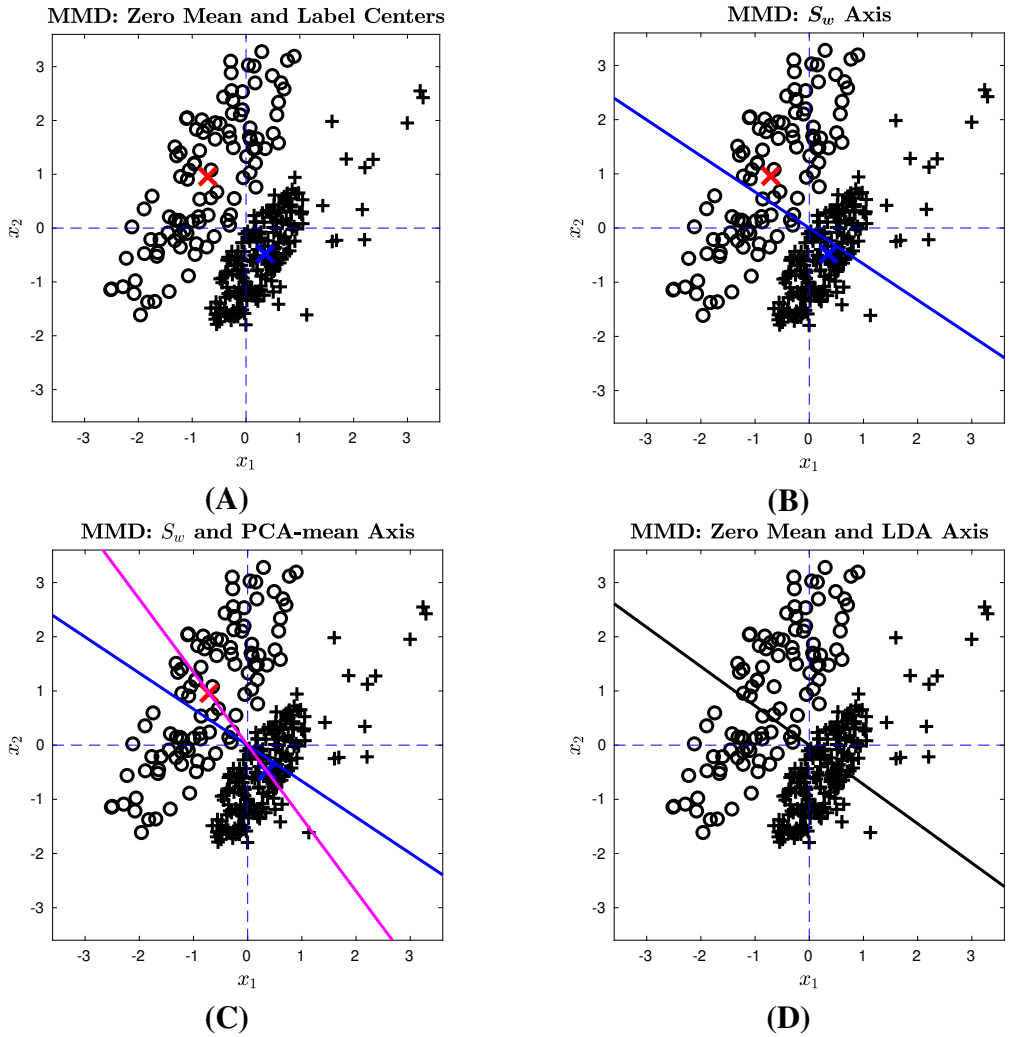

Figure 24.2: Artificially generated data with two labels. (A) The data shown in black, with red for mean of one label and blue for mean of other label. (B) The data and smallest within-label scatter S_w axis, shown in blue. (C) The data and largest between-label scatter S_B, shown in magenta. (D) The data and the LDA axis, shown in black.

These data are linearly separable. The LDA axis, plus additional information, can be used to deduce a separating hyperplane. However, most uses of LDA do not go beyond finding an axis along which the labels of observations are optimally separated by Fisher's linear discriminant.

© R E Ellis 2024

24.4 Extra Notes – Maximum of Rayleigh Quotient

Theorem: Maximum of Rayleigh quotient

For any $B \in \mathbb{R}^{n \times n}$ such that $B = B^T$ and $B \succeq 0$, for the largest eigenvalue of B that is $\lambda_{\text{MAX}}(B)$, the maximum of the Rayleigh quotient of B is

$$\begin{aligned} \lambda_{\text{MAX}}(B) &= \max_{\vec{u} \in \mathbb{R}^n : \vec{u} \neq \vec{0}} R(B, \vec{u}) \\ \vec{v}_{\text{MAX}}(B) &= \operatorname*{argmax}_{\vec{u} \in \mathbb{R}^n : \vec{u} \neq \vec{0}} R(B, \vec{u}) \end{aligned} \tag{24.11}$$

Proof:

Because $\|\vec{u}\| \neq 0$, we can transform the vector $\vec{u}$ to a unit vector $\vec{w}$ as

$$\vec{w} = \frac{\vec{u}}{\|\vec{u}\|} \tag{24.12}$$

The unconstrained Rayleigh quotient of Equation 24.11 is transformed, using the substitution of Equation 24.12, to a constrained Rayleigh quotient that can be written as

$$\max_{\vec{u} \in \mathbb{R}^n : \vec{u} \neq \vec{0}} R(B, \vec{u}) \max_{\vec{u} \in \mathbb{R}^n : \vec{u} \neq \vec{0}} \frac{\vec{u}^T B \vec{u}}{\vec{u}^T \vec{u}} = \max_{\vec{w} \in \mathbb{R}^n : \|\vec{w}\| = 1} \vec{w}^T B \vec{w} \tag{24.13}$$

Consider representing the gradient of a function that has a vector argument as a 1-form. It is straightforward to demonstrate that, for any symmetric matrix $K = K^T$,

$$\frac{\partial}{\partial \vec{w}} \left[\vec{w}^T K \vec{w} \right] = 2 \vec{w}^T K \tag{24.14}$$

The constrained optimization problem of Equation 24.13 can be solved by forming the Lagrangian function $\mathcal{L}(\vec{w}, \lambda)$ from the objective and the constraint that $\vec{w}^T \vec{w} - 1 = 0$, so

$$\mathcal{L}(\vec{w}, \lambda) = \vec{w}^T B \vec{w} - \lambda (\vec{w}^T \vec{w} - 1) \tag{24.15}$$

Differentiating Equation 24.15 with respect to $\vec{w}$ and λ, and setting the transposes equal to the zero vector and zero respectively, give the Lagrange equations

$$\begin{aligned} \left[\frac{\partial}{\partial \vec{w}} \mathcal{L}(\vec{w}, \lambda) \right]^T &= 2B\vec{w} - 2\lambda \vec{w} &= \vec{0} \\ \left[\frac{\partial}{\partial \lambda} \mathcal{L}(\vec{w}, \lambda) \right]^T &= \vec{w}^T \vec{w} - 1 &= 0 \end{aligned} \tag{24.16}$$

© R E Ellis 2024

The solutions to Equation 24.16 are found as

$$\begin{aligned} B\vec{w}^* &= \lambda^* \vec{w}^* && (24.17) \\ \|\vec{w}^*\| &= 1 && (24.18) \end{aligned}$$

The solution of Equation 24.18 requires that $\vec{w}^*$ have a unit norm. The solution of Equation 24.17 requires that λ^* be an eigenvalue of B and that $\vec{w}^*$ be the associated eigenvector. The problem in Equation 24.13 is maximized by

$$\begin{aligned} \vec{w}^* &= \vec{v}_{\text{MAX}}(B) \\ \lambda^* &= \lambda_{\text{MAX}}(B) \end{aligned} \tag{24.19}$$

The constrained solutions of Equation 24.19 can be substituted into Equation 24.13 to find Equation 24.11.

24.5 Extra Notes – Fisher's Linear Discriminant

Theorem: Fisher's Linear Discriminant

> For any $S_B \in \mathbb{R}^{n \times n}$ such that $S_B = S_B^T$ and $S_B \succeq 0$, and for any $S_W \in \mathbb{R}^{n \times n}$ such that $S_W = S^T$ and $S_W \succ 0$, for the largest eigenvalue of B that is $\lambda_{\text{MAX}}(B)$, the maximum of the ratio $R(S_B, \vec{u})/R(S_W, \vec{u})$ is
>
> $$\begin{aligned} \lambda_{\text{MAX}}(S_W^{-1} S_B) &= \max_{\vec{u} \in \mathbb{R}^n : \vec{u} \neq \vec{0}} \frac{R(S_B, \vec{u})}{R(S_W, \vec{u})} \\ \vec{v}_{\text{MAX}}(S_W^{-1} S_B) &= \operatorname*{argmax}_{\vec{u} \in \mathbb{R}^n : \vec{u} \neq \vec{0}} \frac{R(S_B, \vec{u})}{R(S_W, \vec{u})} \end{aligned} \tag{24.20}$$

Proof:

We can assume that $\vec{u} \neq \vec{0}$ because of the constraints in Equation 24.20. We can abbreviate the numerators of the Rayleigh quotients, and the ratio of the Rayleigh quotients, as

$$\begin{aligned} f_B(\vec{u}) &\stackrel{\text{def}}{=} \vec{u}^T S_B \vec{u} \\ f_W(\vec{u}) &\stackrel{\text{def}}{=} \vec{u}^T S_W \vec{u} \\ f(\vec{u}) &\stackrel{\text{def}}{=} \frac{f_B(\vec{u})}{f_W(\vec{u})} \end{aligned} \tag{24.21}$$

© R E Ellis 2024

Observe that, because $S_W \succ 0$, $(\vec{u} \neq \vec{0}) \rightarrow (f_W(\vec{u}) > 0)$ so $f(\vec{u})$ is well formed. Representing vector derivatives as 1-forms, and for brevity omitting the vector argument $\vec{u}$, recall the Quotient Rule for calculus to find the derivative of $f(\vec{u})$ in Equation 24.21 as

$$\begin{aligned} \frac{\partial f}{\partial \vec{u}} &= \frac{[\partial f_B/\partial \vec{u}]f_W - [\partial f_W/\partial \vec{u}]f_B}{(f_W)^2} \\ &= \frac{[2\vec{u}^T S_B]\vec{u}^T S_W \vec{u} - [2\vec{u}^T S_W]\vec{u}^T S_B \vec{u}}{(f_W)^2} \end{aligned} \tag{24.22}$$

Set $[\frac{\partial f}{\partial \vec{u}}]^T$ of Equation 24.22 to $\vec{0}$, and multiply both sides by $(f_W)^2$, to write

$$[\vec{u}^*]^T S_W \vec{u}^* [S_B^T \vec{u}^*] - [\vec{u}^*]^T S_B \vec{u}^* [S_W^T \vec{u}^*] = \vec{0} \tag{24.23}$$

Because $S_B = S_B^T$ and $S_W = S_W^T$ and $f_W(\vec{u}) > 0$, we can substitute transposes into Equation 24.23 and divide by $[\vec{u}^*]^T S_W \vec{u}^*$ to write

$$S_B \vec{u}^* - \frac{[\vec{u}^*]^T S_B \vec{u}^*}{[\vec{u}^*]^T S_W \vec{u}^*} S_W \vec{u}^* = \vec{0} \tag{24.24}$$

We can abbreviate the ratio in Equation24.24 as a function that has a vector argument, so set

$$\lambda(\vec{u}^*) \stackrel{\text{def}}{=} \frac{[\vec{u}^*]^T S_B \vec{u}^*}{[\vec{u}^*]^T S_W \vec{u}^*} \tag{24.25}$$

Substituting Equation 24.25 into Equation24.24, we can write

$$\begin{aligned} S_B \vec{u}^* - \lambda(\vec{u}^*) S_W \vec{u}^* &= \vec{0} \\ \equiv \quad S_B \vec{u}^* &= \lambda(\vec{u}^*) S_W \vec{u}^* \\ \equiv \quad S_B \vec{u}^* &= S_W \lambda(\vec{u}^*) \vec{u}^* \end{aligned} \tag{24.26}$$

Because $S_W \succ 0$, S_W^{-1} exists. We can solve Equation 24.26 by pre-multiplying both sides by S_W^{-1}, so we can write

$$[S_W^{-1} S_B]\vec{u}^* = \lambda(\vec{u}^*)\vec{u}^* \tag{24.27}$$

The solutions to Equation 24.27 are the unit eigenvectors $\vec{u}^*$ and the associated eigenvalues $\lambda(\vec{u}^*)$. The maximum eigenvalue/eigenvector pair is the solution to Equation 24.20.

© R E Ellis 2024