EECS 553: Machine Learning (ECE) **University of Michigan**

Linear Discriminant Analysis

Winter 2023 **Clayton Scott**

1 Linear Classifiers

Linear Discriminant Analysis (LDA) is a classification method that produces a *linear classifier*. What is a linear classifier? When there are two classes, and the labels are -1 and 1, a linear classifier has the form

$$f(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^T\boldsymbol{x} + b), \tag{1}$$

where $\boldsymbol{w} \in \mathbb{R}^d$ is called the *weight vector* and $b \in \mathbb{R}$ is called the *bias* or *offset*, and $\text{sign}(t)$ is 1 if $t \geq 0$ and -1 otherwise.[1] The set $\{\boldsymbol{x} \mid \boldsymbol{w}^T\boldsymbol{x} + b = 0\}$ is the equation of a *hyperplane*.

For multiclass classification with labels $1, 2, \ldots, K$, a linear classifier has the form

$$f(\boldsymbol{x}) = \underset{k=1,\ldots,K}{\arg\max}\ \boldsymbol{w}_k^T\boldsymbol{x} + b_k, \tag{2}$$

where $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K \in \mathbb{R}^d$ and $b_1, \ldots, b_K \in \mathbb{R}$. LDA is our first example of a linear classifier but we will see others. As an exercise one can easily show that the multiclass definition of linear classifier when $K = 2$ agrees with the definition in (1).

2 The LDA Assumption

Unless otherwise noted, let the class labels be $1, 2, \ldots, K$. LDA is a plug-in classifier based on the formula

$$f^*(\boldsymbol{x}) = \underset{k}{\arg\max}\ \pi_k g_k(\boldsymbol{x})$$

for a Bayes classifier. In LDA, it is assumed that $\boldsymbol{X}|Y = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ for $k \in \{1, \ldots, K\}$. In other words,

$$\begin{aligned} g_k(\boldsymbol{x}) &= \phi(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \\ &:= (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)). \end{aligned}$$

Thus, the class conditional distributions are multivariate Gaussians with a common covariance matrix $\boldsymbol{\Sigma}$.

3 The LDA Classifier

Suppose we have training data $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)$. LDA is the classifier obtained by estimating $\pi_k, \boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$ by

$$\begin{aligned} \widehat{\pi}_k &= \frac{n_k}{n}, \qquad n_k = |\{i : y_i = k\}| \\ \widehat{\boldsymbol{\mu}}_k &= \frac{1}{n_k} \sum_{i:y_i=k} \boldsymbol{x}_i \\ \widehat{\boldsymbol{\Sigma}} &= \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_{y_i})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_{y_i})^T. \end{aligned}$$

[1] A better name for a linear classifier would be an *affine classifier*, since $\boldsymbol{w}^T\boldsymbol{x} + b$ is an affine function of $\boldsymbol{x}$, but the term "linear" is very entrenched.

and plugging these estimates into the Bayes classifier formula. In other words, the LDA classifier is

$$\widehat{f}(\boldsymbol{x}) = \underset{1\le k\le K}{\arg\max}\ \widehat{\pi}_k \cdot \phi(\boldsymbol{x}; \widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}})$$

The estimates $\widehat{\pi}_k, \widehat{\boldsymbol{\mu}}_k$, and $\widehat{\boldsymbol{\Sigma}}$ can be shown to be maximum likelihood estimates. The estimate $\widehat{\boldsymbol{\Sigma}}$ is known as *the pooled sample covariance matrix*. This estimate can be written

$$\widehat{\boldsymbol{\Sigma}} = \sum_{k=1}^{K} \frac{n_k}{n} \widehat{\boldsymbol{\Sigma}}_k,$$

where $\widehat{\boldsymbol{\Sigma}}_k$ is the sample covariance matrix based on data from class k.

Since the natural logarithm is a strictly increasing function, the LDA classifier can also be expressed

$$\widehat{f}(\boldsymbol{x}) = \underset{1\le k\le K}{\arg\max}\ \log \widehat{\pi}_k + \log \phi(\boldsymbol{x}; \widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}).$$

3.1 Binary Case

Let's show that $\widehat{f}$ is a linear classifier when $K = 2$. For this subsection only, let the labels be -1 and 1. Then the LDA classifier is

$$\begin{aligned}
\widehat{f}(\boldsymbol{x}) &= \operatorname{sign}(\log \widehat{\pi}_1 + \log \phi(\boldsymbol{x}; \widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\Sigma}}) - \log \widehat{\pi}_{-1} - \log \phi(\boldsymbol{x}; \widehat{\boldsymbol{\mu}}_{-1}, \widehat{\boldsymbol{\Sigma}})) \\
&= \operatorname{sign}\Bigg(\log \widehat{\pi}_1 - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\widehat{\boldsymbol{\Sigma}}| - \frac{1}{2}(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_1)^T \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_1) \\
&\qquad - \log \widehat{\pi}_{-1} + \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\widehat{\boldsymbol{\Sigma}}| + \frac{1}{2}(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_{-1})^T \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_{-1}) \Bigg) \\
&= \operatorname{sign}\left(\log \frac{\widehat{\pi}_1}{\widehat{\pi}_{-1}} + \frac{1}{2}\Big[(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_{-1})^T \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_{-1}) - (\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_1)^T \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_1) \Big] \right) \\
&= \operatorname{sign}\left(\log \frac{\widehat{\pi}_1}{\widehat{\pi}_{-1}} + \frac{1}{2}\Big[\boldsymbol{x}^T \widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{x} - 2\boldsymbol{x}^T \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_{-1} + \widehat{\boldsymbol{\mu}}_{-1}^T \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_{-1} - \boldsymbol{x}^T \widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{x} + 2\boldsymbol{x}^T \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_1^T \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_1 \Big] \right) \\
&= \operatorname{sign}\left(\log \frac{\widehat{\pi}_1}{\widehat{\pi}_{-1}} + \boldsymbol{x}^T \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1}) + \frac{1}{2}(\widehat{\boldsymbol{\mu}}_{-1}^T \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_{-1} - \widehat{\boldsymbol{\mu}}_1^T \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_1) \right) \\
&= \operatorname{sign}(\boldsymbol{w}^T \boldsymbol{x} + b)
\end{aligned}$$

where $\boldsymbol{w} = \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1})$ and $b = \log(\widehat{\pi}_1/\widehat{\pi}_{-1}) + \frac{1}{2}(\widehat{\boldsymbol{\mu}}_{-1}^T \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_{-1} - \widehat{\boldsymbol{\mu}}_1^T \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_1)$.

3.2 Multiclass Case

In the multiclass case, we can also show that LDA is linear. Because constant terms (that is, terms not depending on k) do not affect which k achieves the maximum, we have

$$\begin{aligned}
\widehat{f}(\boldsymbol{x}) &= \arg\max_k \ \log \widehat{\pi}_k + \log \phi(\boldsymbol{x}; \widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}) \\
&= \arg\max_k \ \log \widehat{\pi}_k - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\widehat{\boldsymbol{\Sigma}}| - \frac{1}{2}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_k)^T\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_k) \\
&= \arg\max_k \ \log \widehat{\pi}_k - \frac{1}{2}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_k)^T\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_k) \\
&= \arg\max_k \ \log \widehat{\pi}_k - \frac{1}{2}\Big[\boldsymbol{x}^T\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{x} - 2\boldsymbol{x}^T\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_k + \widehat{\boldsymbol{\mu}}_k^T\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_k\Big] \\
&= \arg\max_k \ \log \widehat{\pi}_k + \boldsymbol{x}^T\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_k - \frac{1}{2}\widehat{\boldsymbol{\mu}}_k^T\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_k\Big] \\
&= \arg\max_k \ \boldsymbol{w}_k^T\boldsymbol{x} + b_k
\end{aligned}$$

where $\boldsymbol{w}_k = \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_k$ and $b_k = \log\widehat{\pi}_k - \frac{1}{2}\widehat{\boldsymbol{\mu}}_k^T\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_k$.

4 Mahalanobis Distance

Let $\boldsymbol{A}$ be a $d \times d$ positive definite matrix, and define

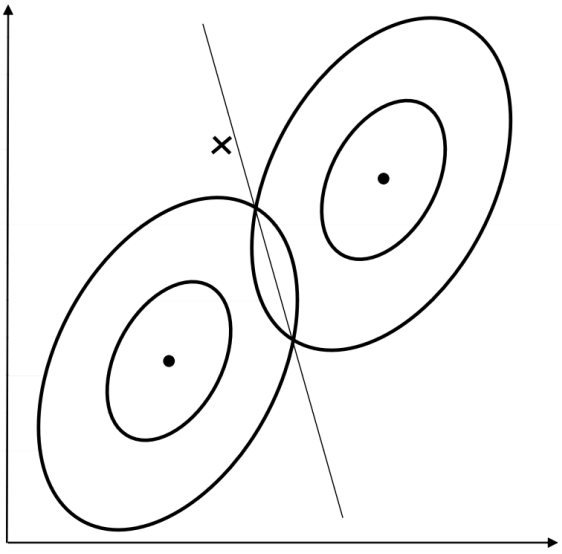
$$D_{\boldsymbol{A}}(\boldsymbol{x}, \boldsymbol{x}') := \sqrt{(\boldsymbol{x}-\boldsymbol{x}')^T\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{x}')},$$

referred to as the *Mahalanobis distance* between $\boldsymbol{x}$ and $\boldsymbol{x}'$. Then the LDA classifier can be expressed

$$\begin{aligned}
\widehat{f}(\boldsymbol{x}) &= \arg\max_k \quad \log\widehat{\pi}_k - \frac{1}{2}D^2_{\widehat{\boldsymbol{\Sigma}}^{-1}}(\boldsymbol{x}, \boldsymbol{\mu}_k) \\
&= \arg\min_k \quad -\log\widehat{\pi}_k + \frac{1}{2}D^2_{\widehat{\boldsymbol{\Sigma}}^{-1}}(\boldsymbol{x}, \boldsymbol{\mu}_k).
\end{aligned}$$

Therefore, if we ignore the $\log\widehat{\pi}_k$ terms, LDA can be viewed as assigning a test point $\boldsymbol{x}$ to the class whose mean $\boldsymbol{\mu}_k$ is closest to $\boldsymbol{x}$ according to the Mahalanobis distance associated to $\widehat{\boldsymbol{\Sigma}}^{-1}$. See Figure 1.

Figure 1: This figure assumes $\widehat{\pi}_1 = \widehat{\pi}_2$. Points on one side of the line are closer to one class's mean in the Mahalanobis distance, while points on the other side are closer to the other mean. The ellipses are contours of the estimated class-conditional distributions.



5 Exercises

1. (★) Prove that if f is a linear classifier according to (2) with $K = 2$, then f is linear according to (1) with an appropriate reassignment of labels. Express $\boldsymbol{w}$ and b in terms of $\boldsymbol{w}_1, \boldsymbol{w}_2, b_1$, and b_2. Verify your result in the special case of LDA by referencing the derivations above.

2. (★) Is LDA generative or discriminative? Parametric or nonparametric?

3. (★★) Show that if $n < d$, then $\widehat{\boldsymbol{\Sigma}}$ is not invertible, meaning LDA cannot be applied.

4. (☆☆) Show that the Mahalanobis distance $D_{\boldsymbol{A}}$ is the metric induced by a certain inner product. This inner product reduces to the dot product when $\boldsymbol{A}$ is the identity matrix. Then show that with respect to this inner product (with $\boldsymbol{A} = \boldsymbol{\Sigma}^{-1}$), the LDA ($K = 2$) decision boundary is orthogonal to the perpendicular bisector of the line segment joining the class sample means.

5. (★★) *Quadratic discriminant analysis* (QDA) is like LDA except each class is allowed to have its own covariance matrix $\boldsymbol{\Sigma}_k$, which is estimated by the sample covariance matrix for that class, that is,

$$\widehat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k} \sum_{i:y_i=k} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)^T.$$

   In the two class case with labels -1 and 1, determine $\boldsymbol{A}$, $\boldsymbol{b}$, and c such that the QDA classifier an be expressed

$$\widehat{f}(\boldsymbol{x}) = \operatorname{sign}\Big(\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + c\Big).$$

6. (☆☆☆) Consider binary classification where the class labels are -1 and 1. Let the training data be $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$. Let $\boldsymbol{\mu}_{-1}$ and $\boldsymbol{\mu}_1$ denote the sample means of the two classes. Define the "between class scatter matrix"

$$S_b = (\boldsymbol{\mu}_{-1} - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_{-1} - \boldsymbol{\mu}_1)^T,$$

and the "within class scatter matrix"

$$S_w = \sum_{k \in \{\pm 1\}} \sum_{i: y_i = k} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T.$$

Fisher's Linear Discriminant (FLD) seeks a linear classifier with normal vector $\boldsymbol{w}$ obtained by solving

$$\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^T S_b \boldsymbol{w}}{\boldsymbol{w}^T S_w \boldsymbol{w}}.$$

The offset b of the linear classifier is determined by some other criterion after $\boldsymbol{w}$ has been determined. FLD has nothing to say about b.

(a) Determine the optimal $\boldsymbol{w}$ according to the FLD criterion, and argue that it coincides with the LDA normal vector.

(b) Justify FLD by interpreting the numerator and denominator of the objective function. Be as quantitative as possible. *Hint:* Think of FLD as a supervised version of PCA, but with just one principal component.