Linear methods for classification: Linear Discriminant Analysis

Matthew S. Shotwell, Ph.D.

Department of Biostatistics
Vanderbilt University School of Medicine
Nashville, TN, USA

January 31, 2020

Linear methods for classification

- G has K classes in $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_2\}$
- e.g., $\mathcal{G} = \{\text{blue}, \text{orange}\}$
- Y is a target varible:

$$Y = [Y_1, \ldots, Y_K]^T$$

$$Y_k = \begin{cases} 1 & G = \mathcal{G}_k \\ 0 & G \neq \mathcal{G}_k \end{cases}$$

- sample of y and x:

$$y = [y_1, \ldots, y_n]^T \qquad (n \times K)$$

$$x = [x_1, \ldots, x_n]^T \qquad (n \times p)$$

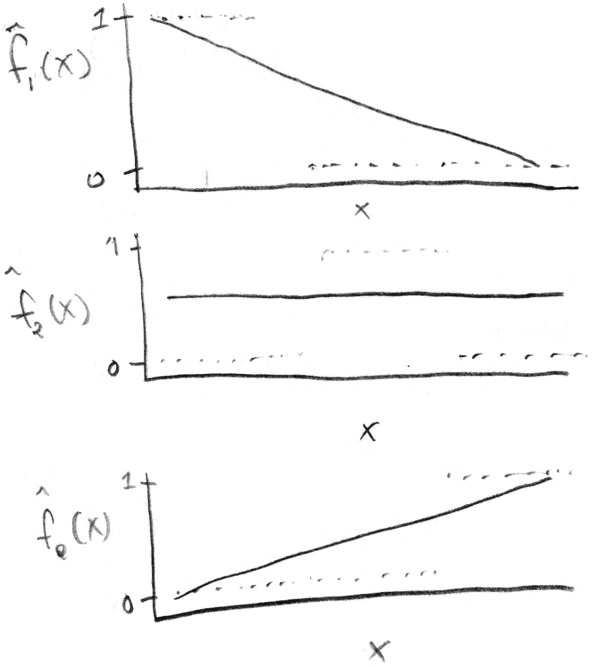
Linear regression of indicator variables

- this is not LDA; this method doesn't work very well
- $\hat{y}_k = \hat{f}_k(x) = \hat{\beta}_{0k} + x\hat{\beta}_k$ where $\hat{\beta}_k$ is p vector$k$
- $\hat{f}_k(x)$ is kind of like $Pr(G = \mathcal{G}_k | X = x)$
- $\hat{G}(x) = \mathcal{G}_k$ with largest $\hat{f}_k(x)$
- $\hat{f}_k(x) = \delta_k(x)$ is "discriminant function"

3 class problem

Histogram of x by 3 classes

Linear fits to indicators



$\hat{f}_2$ is never larger than $\hat{f}_1$ or $\hat{f}_3$

$\Rightarrow$ class 2 always misclassified

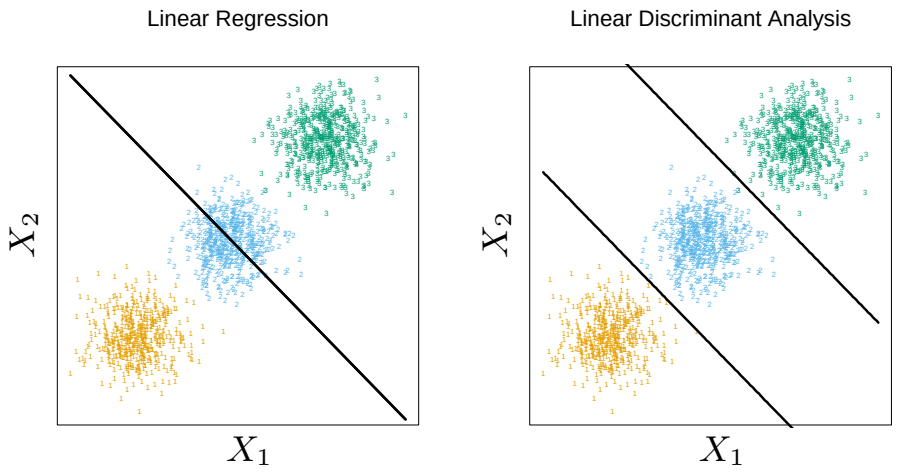


FIGURE 4.2. *The data come from three classes in* $\mathbb{R}^2$ *and are easily separated by linear decision boundaries. The right plot shows the boundaries found by linear discriminant analysis. The left plot shows the boundaries found by linear regression of the indicator response variables. The middle class is completely masked (never dominates).*

Think PCA would help?

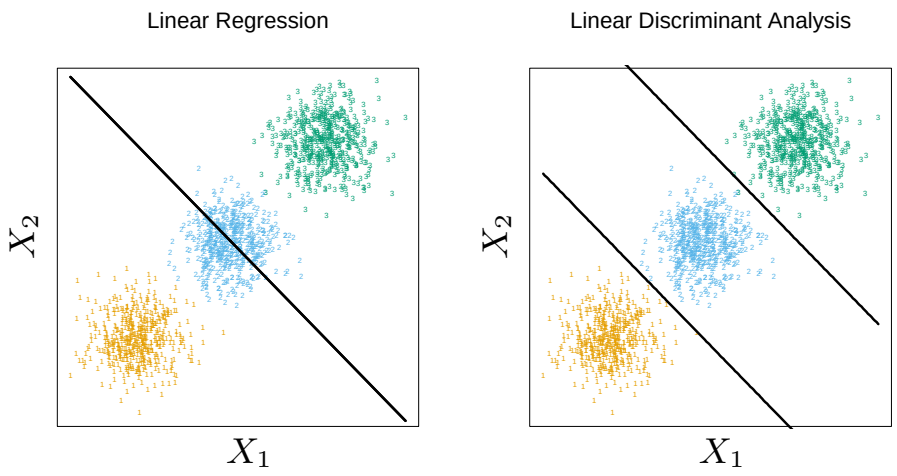


FIGURE 4.2. *The data come from three classes in* $\mathbb{R}^2$ *and are easily separated by linear decision boundaries. The right plot shows the boundaries found by linear discriminant analysis. The left plot shows the boundaries found by linear regression of the indicator response variables. The middle class is completely masked (never dominates).*

Linear regression of indicator variables

- suppose we did PCA and only use first PC in regression
- $\hat{y}_k = \hat{f}_k(x) = \hat{\theta}_{0k} + z_1\hat{\theta}_{1k}$
- now suppose we added a quadratic term in z_1
- $\hat{y}_k = \hat{f}_k(x) = \hat{\theta}_{0k} + z_1\hat{\theta}_{1k} + z_1^2\hat{\theta}_{2k}$
- again, $\hat{f}_k(x) = \delta_k(x)$ is "discriminant function"
- again, $\hat{G}(x) = \mathcal{G}_k$ with largest $\delta_k(x)$

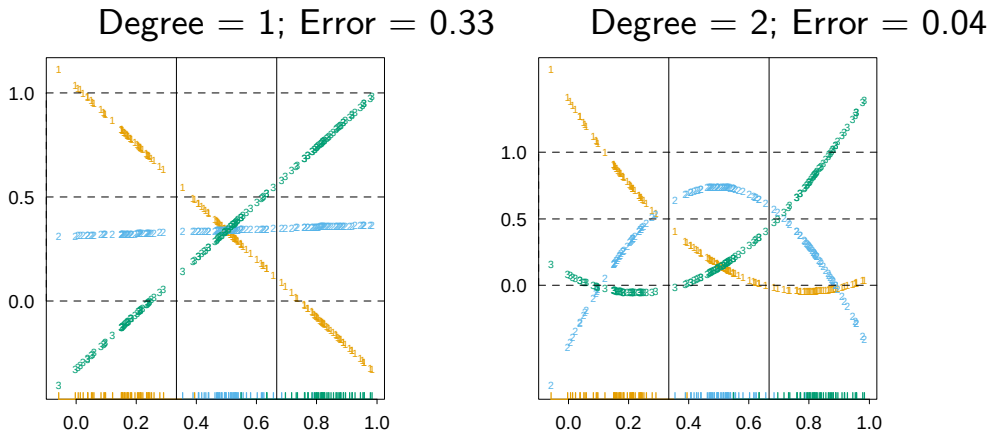


FIGURE 4.3. *The effects of masking on linear regression in* $\mathbb{R}$ *for a three-class problem. The* rug plot *at the base indicates the positions and class membership of each observation. The three curves in each panel are the fitted regressions to the three-class indicator variables; for example, for the blue class,* y_{blue} *is* 1 *for the blue observations, and* 0 *for the green and orange. The fits are linear and quadratic polynomials. Above each plot is the training error rate. The Bayes error rate is* 0.025 *for this problem, as is the LDA error rate.*

Linear discriminant functions

- if there exists a monotome transformation of a descriminant function $\delta_k(x_0)$ that is linear in x_0, then $\delta_k(x_0)$ is a *linear* discriminant function and decision boundary is a hyperplane
- e.g., $\delta_k(x_0) = Pr(G = \mathcal{G}_k | x = x_0) = \mathrm{logit}(x_0\beta)$
- inverse logit is monotone transformation

Linear discriminant analysis (LDA)

- model X given G rather than G given X
- $h_k(x) = Pr(X = x | G = \mathcal{G}_k)$
- $\pi_k = Pr(G = \mathcal{G}_k)$ "prior"
- apply Bayes rule:

$$
\begin{aligned}
f_k(x) &= Pr(G = \mathcal{G}_k | X = x) \\
&= \frac{Pr(X = x | G = \mathcal{G}_k) Pr(G = \mathcal{G}_k)}{Pr(X = x)} \\
&= \frac{h_k(x)\pi_k}{\sum_{l=1}^{K} h_l(x)\pi_l}
\end{aligned}
$$

Linear discriminant analysis (LDA)

- LDA assumes that X given G is multivariate normal
- LDA assumes that $h_k(x) = \phi(x, \mu_k, \Sigma)$
- different mean μ_k for each class
- same variance-covariance $\Sigma_1 = \cdots = \Sigma_K = \Sigma$

Linear discriminant analysis (LDA)

- find linear discriminant function
- $f_k(x)$ is no linear in x
- $\log$ is a monotone transformation
- $\log f_k(x)$ *is* linear in x
- linear discriminant function for LDA is:

$$
\begin{aligned}
\delta_k(x) &= \log f_k(x) \\
&= \log Pr(G = \mathcal{G}_k | X = x) \\
&= \log \pi_k + \log h_k(x) + c \\
&= \log \pi_k + \log \phi(x, \mu_k, \Sigma) + c \\
&= \log \pi_k - \frac{1}{2}(x - \mu_k)^T \Sigma^{-1} (x - \mu_k) + c \\
&= \log \pi_k + x^T \Sigma \mu_k - \frac{1}{2} \mu_k^T \Sigma \mu_k + \frac{1}{2} x^T \Sigma x
\end{aligned}
$$

- c because that part doesn't involve k (doesn't help us discriminate between classes)

LDA decision boundary

- each $\delta_k(x)$ defines a plane
- decision boundary between any two clases k and j occurs where the planes intersect, at $\delta_k(x) = \delta_l(x)$:

$$\delta_k(x) = \delta_l(x)$$

$$0 = \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k - \mu_l)^T \Sigma^{-1} (\mu_k - \mu_l) + x^T \Sigma^{-1} (\mu_k - \mu_l)$$

- place where they intersect is a line

LDA Estimation

- $\hat{\pi}_k = n_k/n$
- $\hat{\mu}_k = 1/n \sum_{g_i=\mathcal{G}_k} x_i$
- $\hat{\Sigma} = 1/(n-K) \sum_k \sum_{g_i=\mathcal{G}_k} (x_i - \hat{\mu}_k)^T (x_i - \hat{\mu}_k)$
- plug-in estimates to compute $\delta_k(x)$
- again, $\hat{G}(x) = \mathcal{G}_k$ with largest $\delta_k(x)$

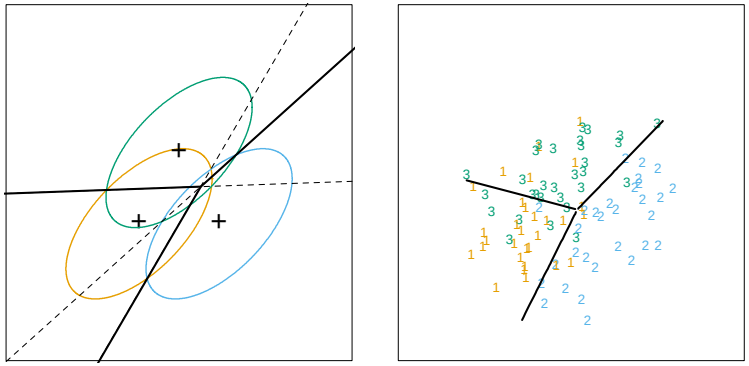


FIGURE 4.5. *The left panel shows three Gaussian distributions, with the same covariance and different means. Included are the contours of constant density enclosing* 95% *of the probability in each case. The Bayes decision boundaries between each pair of classes are shown (broken straight lines), and the Bayes decision boundaries separating all three classes are the thicker solid lines (a subset of the former). On the right we see a sample of* 30 *drawn from each Gaussian distribution, and the fitted LDA decision boundaries.*

Code example

`simple-LDA-3D.R`

LDA tuning parameters?

- LDA is a bit like linear regression
- no natural tuning parameters
- how do we make model more/less flexible?
- how do we tune LDA?

LDA tuning parameters?

How to allow more/less flexibility in LDA

- use basis functions (e.g., interactions or splines)
- use subset selection on the inputs
- use regularization (work a little differently from ridge/lasso)

Quadratic discriminant analysis (QDA)

- relax assumption: $\Sigma_1 = \cdots = \Sigma_K$
- $\delta_k(x) = \log \pi_k - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)$
- discriminant function is not linear, but quadratic in x
- decision boundary is also quadratic in x
- $\hat{\Sigma}_k = 1/(n_k - 1) \sum_{g_i = \mathcal{G}_k} (x_i - \hat{\mu}_k)^T (x_i - \hat{\mu}_k)$

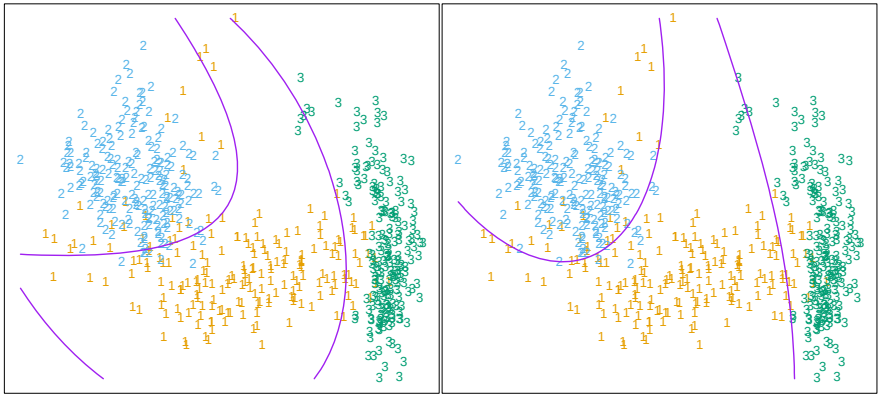


FIGURE 4.6. *Two methods for fitting quadratic boundaries. The left plot shows the quadratic decision boundaries for the data in Figure 4.1 (obtained using LDA in the five-dimensional space* $X_1, X_2, X_1X_2, X_1^2, X_2^2$*). The right plot shows the quadratic decision boundaries found by QDA. The differences are small, as is usually the case.*

Regularized Descriminant Analysis (RDA)

- mix of LDA and QDA
- $\hat{\Sigma}_k^\alpha = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}$
- α makes model more/less flexible (bias/variance)
- α chosen by to minimize good estimate of test error

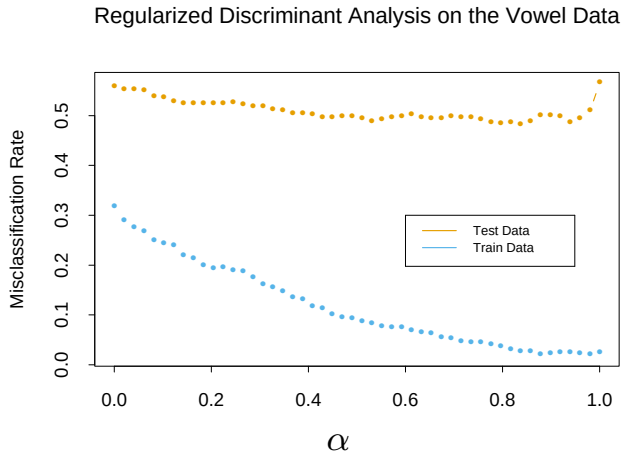


FIGURE 4.7. *Test and training errors for the vowel data, using regularized discriminant analysis with a series of values of* $\alpha \in [0, 1]$. *The optimum for the test data occurs around* $\alpha = 0.9$, *close to quadratic discriminant analysis.*

Where do LDA and QDA not work well?

LDA and QDA don't work well when the assumptions are violated:

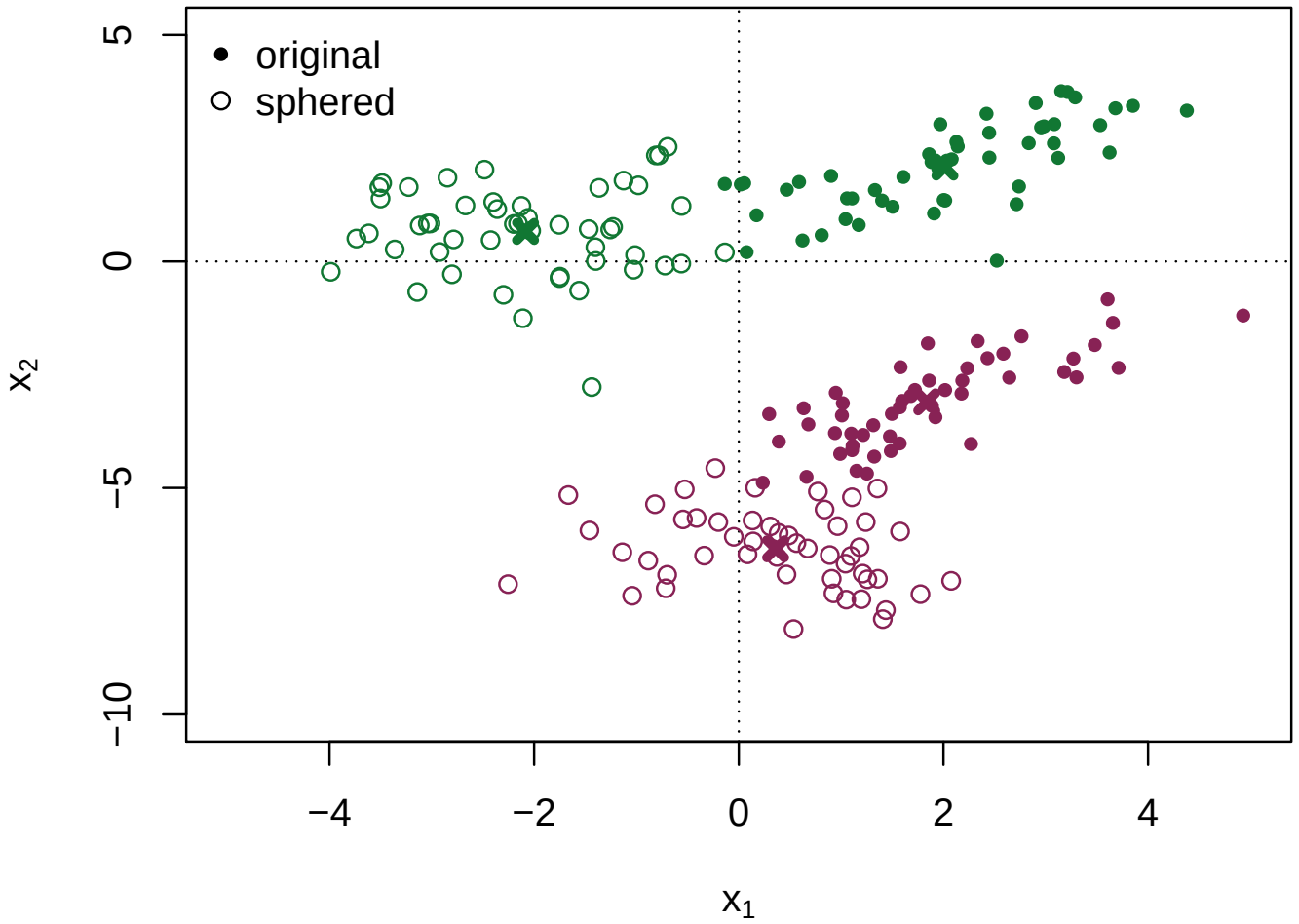
- neither works well when X given G is not multivariate normal
- LDA doesn't work well when variance-covariance is different across the classes (although QDA might work well)

LDA and Eigen-decomposition of Σ

- consider $\Sigma = VDV^T$
- V is $p \times p$ orthonormal
- D is $p \times p$ diagonal, then

$$\begin{aligned}
\delta_k(x) &= \log \pi_k - \frac{1}{2}(x - \mu_k)^T \Sigma^{-1} (x - \mu_k) \\
&= \log \pi_k - \frac{1}{2}(x - \mu_k)^T (V D^{-1/2} D^{-1/2} V^T)(x - \mu_k) \\
&= \log \pi_k - \frac{1}{2}[D^{-1/2} V^T (x - \mu_k)]^T [D^{-1/2} V^T (x - \mu_k)] \\
&= \log \pi_k - \frac{1}{2}[x^* - \mu_k^*]^T [x^* - \mu_k^*]
\end{aligned}$$

- x^* is "sphered" since $\mathrm{cov}(x^*) = I_p$
- if $\pi_1 = \cdots = \pi_K$ then classify by minimizing Euclidean distance from x^* to μ_k^*


original
sphered
5
0
−5
−10
−4
−2
0
2
4
x_1
x_2