Bayes Classifiers and Discriminant Analysis

Donald E. Brown

School of Data Science
University of Virginia
Charlottesville, VA 22904

Discriminant Analysis

- Let $G \in \mathcal{G}$ be the response with $|\mathcal{G}| = K$.
- We have an $N \times p$ data matrix, X.
- From statistical decision theory, we minimize the loss (Bayes risk) by selecting class g_i, $i \in \{1, \ldots, K\}$ if

$$Pr[G = g_i|X] > Pr[G = g_j|X], j \neq i \tag{1}$$

  and BTA
- From Bayes' Rule we can rewrite inequality 1 as

$$f(X|G = g_i)Pr[G = g_i] > f(X|G = g_j)Pr[G = g_j], j \neq i \tag{2}$$

Likelihood Ratio

- We typically write inequality (2) as a ratio:

$$\frac{f(X|G=g_i)}{f(X|G=g_j)} > \frac{Pr[G=g_j]}{Pr[G=g_i]}, j \neq i$$

  again BTA.
- Often the prior comes from the data, and we can get the likelihood from
  - An assumed form (e.g., Gaussian);
  - The data (e.g., distributional fit);
  - Combination or hybrid of the previous two.

Gaussian Likelihood

- Assume a multivariate Gaussian distribution for the likelihood in the Bayes optimal decision rule. Thus,

$$f(X|G = g_j) = \frac{1}{(2\pi)^{k/2}|\Sigma_j|^{1/2}} exp[-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)]$$

- where Σ_j is the variance-covariance matrix for $f(X|G = g_j), j = 1, \ldots, K$ and
- μ_j is the mean vector for $f(X|G = g_j), j = 1, \ldots, K$.

Quadratic Discriminant Analysis

- From this assumed distribution and our statistical decision theory results, we can formulate the Bayes optimal decision rule.
- For $i, j \in \{1, \ldots, K\}$ choose g_i if

$$\frac{Pr[G = g_j]}{Pr[G = g_i]} < \frac{\frac{1}{(2\pi)^{k/2}|\Sigma_i|^{1/2}} exp[-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)]}{\frac{1}{(2\pi)^{k/2}|\Sigma_j|^{1/2}} exp[-\frac{1}{2}(x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j)]}$$

$$\begin{aligned} log\left(\frac{Pr[G = g_j]}{Pr[G = g_i]}\right) &< \frac{1}{2}[log|\Sigma_j| - log|\Sigma_i|] \\ &+ \frac{1}{2}(x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j) \\ &- \frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \end{aligned}$$

Examples

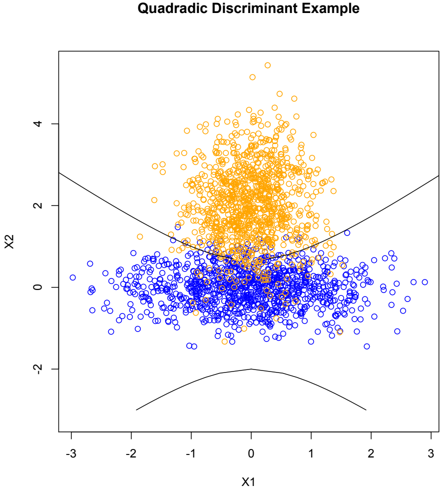
- Let $\mu_1^t = (0, 0)$ and $\mu_2^t = (0, 2)$.
- Let

$$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0.25 \end{bmatrix}$$

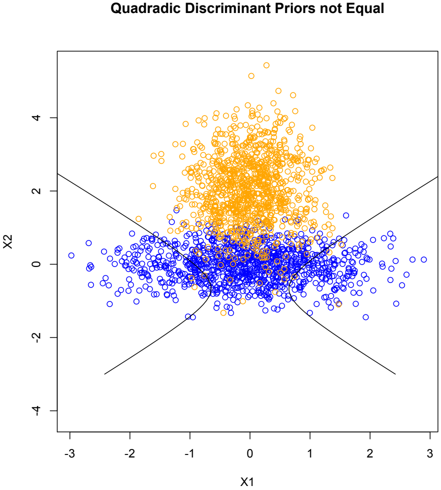
$$\Sigma_2 = \begin{bmatrix} 0.25 & 0 \\ 0 & 1 \end{bmatrix}$$

- First consider equal priors and then $P(G = g_1) = 0.1$

Quadratic Discriminant with Equal Priors



Quadratic Discriminant with Unequal Priors



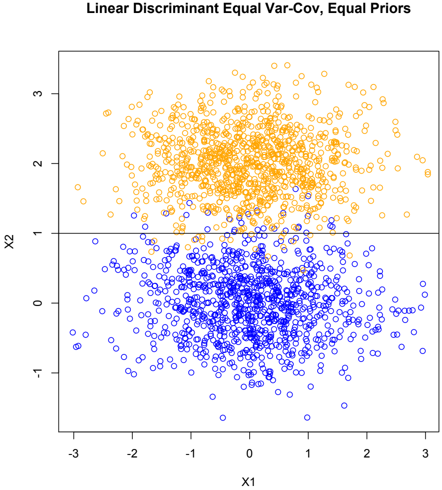
Linear Discriminant Analysis

- Also assume a multivariate Gaussian distribution for the likelihood but with $\Sigma_i = \Sigma_j \forall i, j \in \{1, \ldots, K\}$.
- For $i, j \in \{1, \ldots, K\}$ choose g_i if

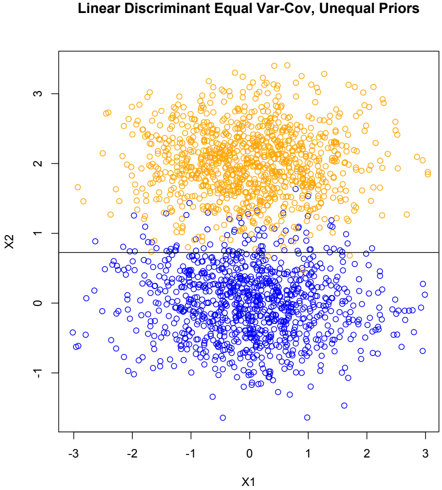
$$\frac{Pr[G = g_j]}{Pr[G = g_i]} < \frac{\frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} exp[-\frac{1}{2}(x - \mu_i)^T \Sigma^{-1}(x - \mu_i)]}{\frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} exp[-\frac{1}{2}(x - \mu_j)^T \Sigma^{-1}(x - \mu_j)]}$$

$$\begin{aligned} log\left(\frac{Pr[G = g_j]}{Pr[G = g_i]}\right) &< (\mu_i - \mu_j)^T \Sigma^{-1} x \\ &+ \frac{1}{2}\mu_j^T \Sigma^{-1} \mu_j \\ &- \frac{1}{2}\mu_i^T \Sigma^{-1} \mu_i \end{aligned}$$

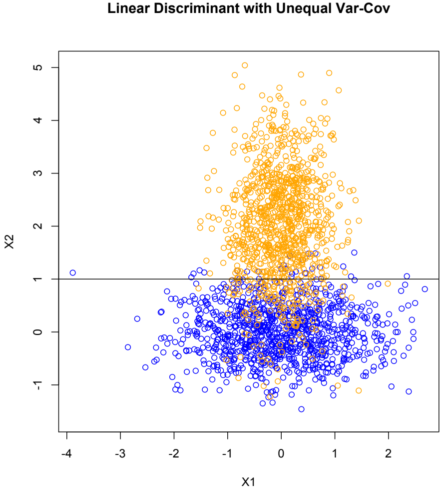
Linear Discriminant with Equal Priors



Linear Discriminant with Unequal Priors



Linear Discriminant with Unequal Var-Cov



Discriminant Analysis with Cost of Misclassification

- Let $c_{i,k}$ be the cost of choosing class i when the true class is k.
- Then rule that minimizes Bayes' risk is: Choose g_i if

$$\sum_{k=1}^{p} c_{i,k} Pr(g_k|x) < \sum_{k=1}^{p} c_{j,k} Pr(g_k|x)$$

- For $i \neq j$, and $i, j \in 1, \ldots, p$. BTA
- Often we assume $c_{i,i} = 0$ for all i. In what applications is this not a good assumption?

Likelihood Ratio with Costs

- Suppose we have only two classes: g_1 and g_2. Also assume $c_{i,i} = 0$. With these assumptions we apply Bayes rule to our previous rule to get a likelihood ratio form for the rule.
- Choose g_i if

$$\frac{f(x|g_i)}{f(x|g_j)} > \frac{c_{i,j} Pr(g_j)}{c_{j,i} Pr(g_i)}$$

- For $i \neq j$, and $i, j \in 1, 2$. Break ties arbitrarily.
- For linear discriminants (i.e., equal covariance matrices) the likelihood ratio becomes

$$(\mu_i - \mu_j)^T \Sigma^{-1} x + \frac{1}{2} \mu_j^T \Sigma^{-1} \mu_j - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i$$

$$> ln \left[\frac{c_{i,j} Pr(g_j)}{c_{j,i} Pr(g_i)} \right]$$