Discriminant Analysis

Discriminant Analysis versus Logistic Regression

- Logistic regression directly estimates $\Pr(Y_i = k|X_i)$ while discriminant analysis estimates the likelihood $f(X_i|Y_i = k)$ and then applies the Bayesian theorem to obtain $\Pr(Y_i = k|X_i)$.
- For well separated classes logistic regression can be numerically unstable (remember the problem of high discrimination).
- For small sample sizes and approximately normal distributed $N(X_i|Y_i = k)$ classes discriminant analysis is superior.
- Due to the distributional assumption, discriminant analysis performs best for metrically scaled variables
- It is easier to handle more than $K > 2$ classes in discriminant analysis.
- Both methods do not require that the features are normalized or standardized.

Use of Bayesian Classification

- Let $\pi_k \; with \; k \in \{1,2, \dots, K\}$ be the prior probability of class k, i.e., an observation Y_i coming from class k prior to any additional information.
- Combined with the likelihood $f(X_i|Y_i = k)$ we get the Bayesian posteriori probability

$$\Pr(Y_i = k|X_i) = \frac{\pi_k \cdot f(X_i|Y_i = k)}{\sum_{l=1}^{K} \pi_l \cdot f(X_i|Y_i = l)}$$

- An object i is assigned to class k for which $\max_k \Pr(Y_i = k|X_i)$ is the largest.
- In order to obtain the posterior probabilities, the prior probabilities and the likelihood need to be estimated from the data.

  The exact functional form of the likelihood needs to be fully specified. In discriminant analysis this is the normal distribution.

Example with $K = 2$ and $p = 1$

- For $K = 2$ and $p = 1$ the two normal distributions have parameters
  - $N(x_i|\mu_1, \sigma_1^2)$ with a mixing proportion π_1 and
  - $N(x_i|\mu_2, \sigma_2^2)$ with a mixing proportion π_2.

- For $\sigma_1^2 = \sigma_2^2$ the Bayesian decision function separating both classes is

$$\delta_k(x_i) = x_i \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2 \cdot \sigma^2} + \log \pi_k$$

and x_i is assigned to class k for which $\delta_k(x_i) > \delta_l(x_i)$ with $k \neq l$.

- Theoretical and empirical mixture distributions:

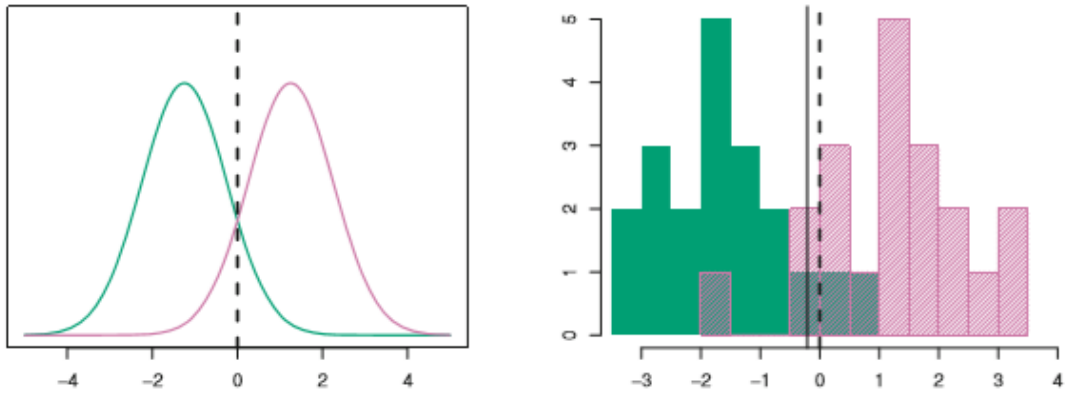

FIGURE 4.4. Left: *Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary.* Right: *20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The solid vertical line represents the LDA decision boundary estimated from the training data.*

- The estimated parameters are:

- $\hat{\pi}_k = \frac{n_k}{\sum_{l=1}^{K} n_l}$
- $\hat{\mu}_k = \frac{1}{n_k} \cdot \sum_{i \in \{y_i = k\}} x_i$
- $\hat{\sigma}^2 = \frac{1}{n-K} \cdot \sum_{k=1}^{K} \sum_{i \in \{y_i = k\}} (x_i - \hat{\mu}_k)^2$

Linear Discriminant functions for $K \geq 2$, $p \geq 2$ and $\boldsymbol{\Sigma_k = \Sigma_l}$ $\forall \boldsymbol{k, l}$

- For two and more classes and more than one feature but identical group covariance matrices the Bayesian classification functions are:

$$\delta_k(\mathbf{x}_i) = \mathbf{x}_i^T \cdot \boldsymbol{\Sigma}^{-1} \cdot \boldsymbol{\mu}_k - \frac{1}{2} \cdot \boldsymbol{\mu}_k^T \cdot \boldsymbol{\Sigma}^{-1} \cdot \boldsymbol{\mu}_k + \log \pi_k$$

  This is a linear function in $\mathbf{x}_i$

- Again an object i is assigned to the group for which $\delta_k(\mathbf{x}_i)$ is the largest.

- Example with $K = 3$ and $p = 2$:

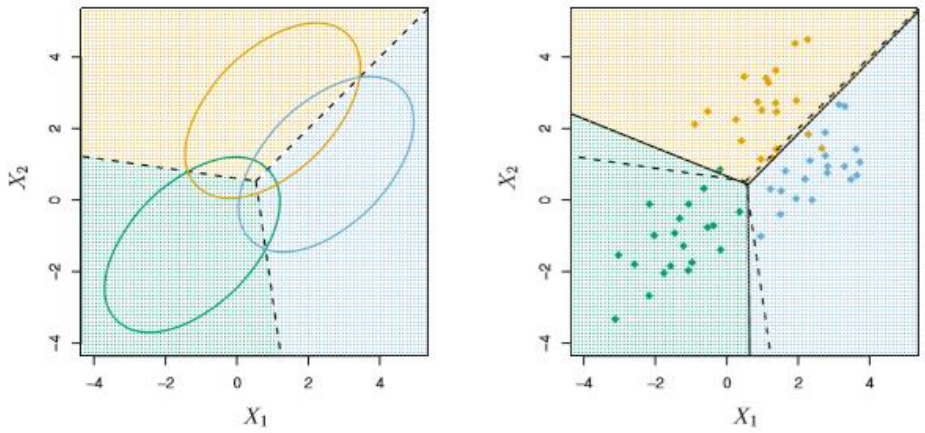

FIGURE 4.6. *An example with three classes. The observations from each class are drawn from a multivariate Gaussian distribution with $p = 2$, with a class-specific mean vector and a common covariance matrix.* Left: *Ellipses that contain 95 % of the probability for each of the three classes are shown. The dashed lines are the Bayes decision boundaries.* Right: 20 *observations were generated from each class, and the corresponding LDA decision boundaries are indicated using solid black lines. The Bayes decision boundaries are once again shown as dashed lines.*

- In linear discriminant analysis the function $\delta_k(\mathbf{x}_i)$ is of dimensionality $p - 1$.

Quadratic Discriminant Analysis with $\boldsymbol{\Sigma_k \neq \Sigma_l}$

- In contrast to linear discriminant analysis in quadratic discriminant analysis each group is normally distributed with a unique covariance matrix $\mathbf{x}_i \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

- The Bayesian classifier becomes

$$\delta_k(\mathbf{x}_i) = -\frac{1}{2} \cdot (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \cdot \boldsymbol{\Sigma}_k^{-1} \cdot (\mathbf{x}_i - \boldsymbol{\mu}_k) - \frac{1}{2} \cdot \log|\boldsymbol{\Sigma}_k| + \log \pi_k$$

- Here the Bayesian classifier $\delta_k(\mathbf{x}_i)$ is a quadratic rather than a linear function. It therefore is more flexible in separating the domains of classes in the $\mathbf{x}_i$ space.

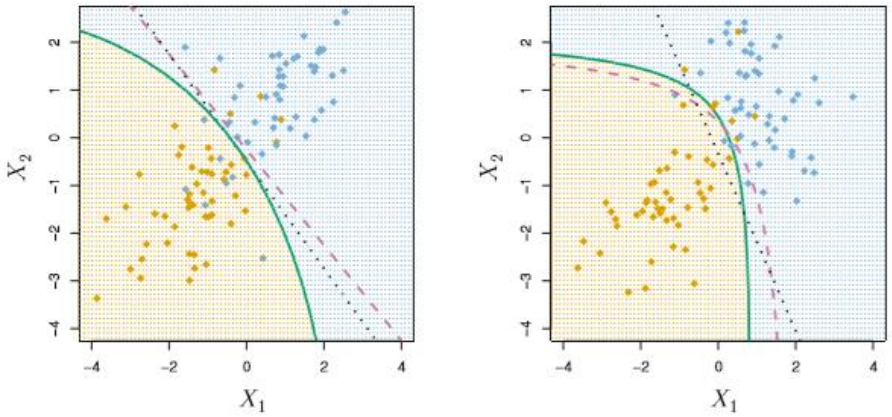

FIGURE 4.9. Left: *The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with* $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. *The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA.* Right: *Details are as given in the left-hand panel, except that* $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. *Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.*

- This flexibility, however, comes at the expense of estimating substantially more parameters for the K covariance matrices.

- Trade-off between linear and quadratic discriminant analysis:
    - Linear discriminant has a substantial bias (mis-specified functional form) if the covariance matrices are substantially different
    - For a low number of training observations linear discriminant analysis may have a lower variance (less variation of the estimated function from sample to sample).

Comparison of the different classification methods

- Structurally linear discriminant analysis and logistic regression are similar because in both cases the log-odds $\log(\Pr(Y_i = 1|X_i)/1 - (\Pr(Y_i = k|X_i)))$ are expressed as linear functions in X_i.
- However, discriminant analysis makes the stronger assumption of normal distributions. If the Gaussian assumption is not satisfied then logistic regression may outperform linear discriminant analysis.
- Discriminant analysis relies on substantially more estimated parameters (the group probabilities, vectors of group means and covariance matrices).