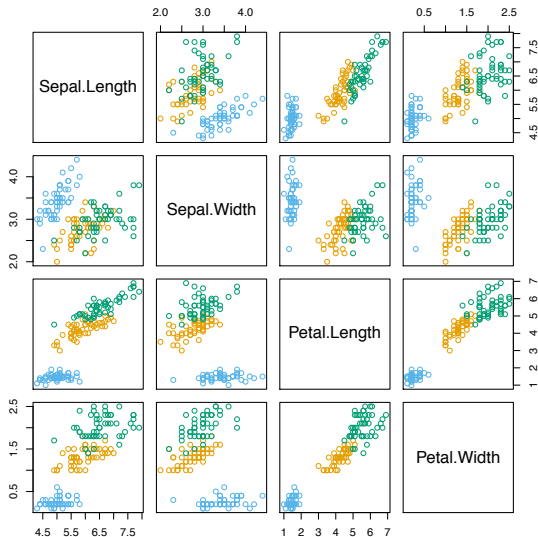


Fisher's Iris Data

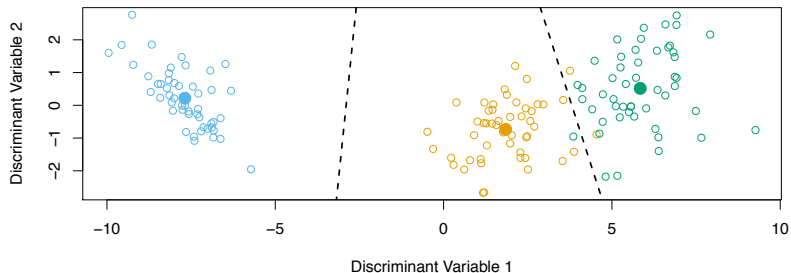
4 variables
3 species
50 samples/class

- Setosa
- Versicolor
- Virginica

LDA classifies all
but 3 of the 150
training samples
correctly.



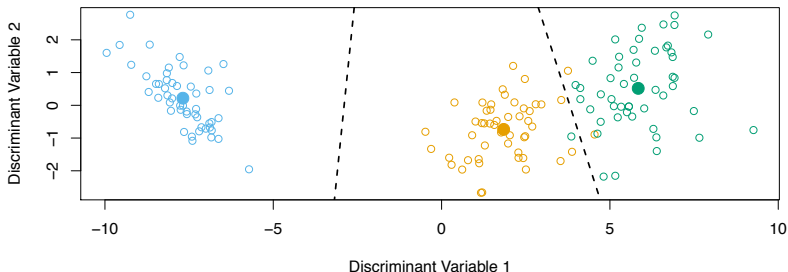
Fisher's Discriminant Plot



When there are K classes, linear discriminant analysis can be viewed exactly in a $K - 1$ dimensional plot.

Why?

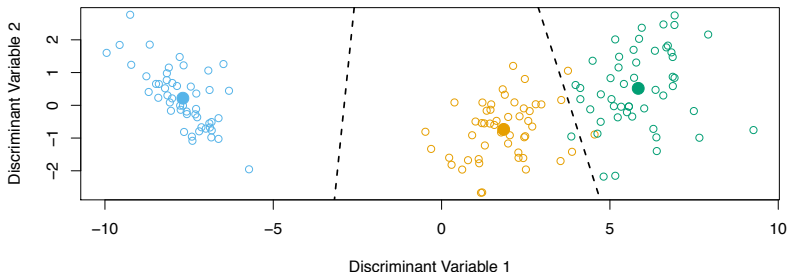
Fisher's Discriminant Plot



When there are K classes, linear discriminant analysis can be viewed exactly in a $K - 1$ dimensional plot.

Why? Because it essentially classifies to the closest centroid, and they span a $K - 1$ dimensional plane.

Fisher's Discriminant Plot



When there are K classes, linear discriminant analysis can be viewed exactly in a $K - 1$ dimensional plot.

Why? Because it essentially classifies to the closest centroid, and they span a $K - 1$ dimensional plane.

Even when $K > 3$, we can find the “best” 2-dimensional plane for visualizing the discriminant rule.

Other forms of Discriminant Analysis

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)}$$

When $f_k(x)$ are Gaussian densities, with the same covariance matrix Σ in each class, this leads to linear discriminant analysis. By altering the forms for $f_k(x)$, we get different classifiers.

- With Gaussians but different Σ_k in each class, we get *quadratic discriminant analysis*.

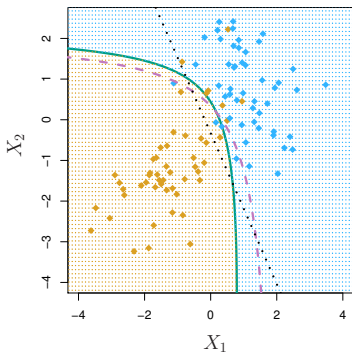
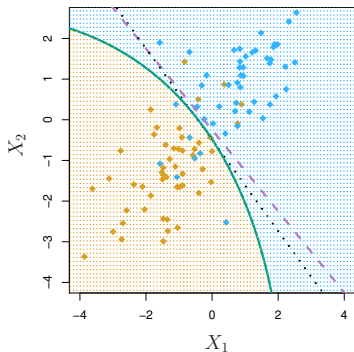
Other forms of Discriminant Analysis

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)}$$

When $f_k(x)$ are Gaussian densities, with the same covariance matrix Σ in each class, this leads to linear discriminant analysis. By altering the forms for $f_k(x)$, we get different classifiers.

- With Gaussians but different Σ_k in each class, we get *quadratic discriminant analysis*.
- With $f_k(x) = \prod_{j=1}^p f_{jk}(x_j)$ (conditional independence model) in each class we get *naive Bayes*. For Gaussian this means the Σ_k are diagonal.

Quadratic Discriminant Analysis



$$\delta_k(x) = -\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) + \log \pi_k$$

Because the Σ_k are different, the quadratic terms matter.

Naive Bayes

Assumes features are independent in each class.

Useful when p is large, and so multivariate methods like QDA and even LDA break down.

- Gaussian naive Bayes assumes each Σ_k is diagonal:

$$\delta_k(x) \propto \log \left[\pi_k \prod_{j=1}^p f_{kj}(x_j) \right] = -\frac{1}{2} \sum_{j=1}^p \frac{(x_j - \mu_{kj})^2}{\sigma_{kj}^2} + \log \pi_k$$

- can use for *mixed* feature vectors (qualitative and quantitative). If X_j is qualitative, replace $f_{kj}(x_j)$ with probability mass function (histogram) over discrete categories.

Despite strong assumptions, naive Bayes often produces good classification results.