UNIVERSITY *of* DELAWARE

Modern Machine Learning
Linear Methods for Classification — Linear Discriminate Analysis

Kenneth E. Barner
Department of Electrical and Computer Engineering
University of Delaware

Linear Methods for Classification

Problem: Classification is the prediction of qualitative responses

Consider the binary classification problem

Examples:

- Email: Spam/Not Spam
- Tumor: Benign/Malignant

Output $Y \in \{0,1\}$ $\begin{cases} 0 \coloneqq \text{Negative class (e.g. Not Spam)} \\ 1 \coloneqq \text{Positive class (e.g. Spam)} \end{cases}$

Objective: Divide the input space into a collection of regions labeled according to classification. Linear methods for classification result in linear decision boundaries

Linear Methods for Classification

Assumption: Suppose there are K classes and the fitted linear model for the k^{th} indicator response variable is

$$\hat{f}_k(\boldsymbol{x}) = \hat{\beta}_{k0} + \widehat{\boldsymbol{\beta}}_k^T \boldsymbol{x} \quad k = 1,2,\dots,K$$

The decision boundary between classes k and l is the set of points for which $\hat{f}_k(\boldsymbol{x}) = \hat{f}_l(\boldsymbol{x})$

$$\{\boldsymbol{x}: (\hat{\beta}_{k0} - \hat{\beta}_{l0}) + (\widehat{\boldsymbol{\beta}}_k^T - \widehat{\boldsymbol{\beta}}_l^T)\boldsymbol{x} = 0\}$$

Note: The decision boundaries are a set of hyper planes $\Rightarrow$the input space is divided into regions with piecewise hyper plane decision boundaries

- This regression approach is within the class of methods that model discrimination functions for each class $\delta_k(\boldsymbol{x})$
  - $\boldsymbol{x}$ is classified to the class with the largest value for its discriminant function
  - Posterior probability models $\Pr(G = k|\boldsymbol{X} = \boldsymbol{x})$ are in this class
  - If $\delta_k(\boldsymbol{x})$ or $\Pr(G = k|\boldsymbol{X} = \boldsymbol{x})$ are linear in $\boldsymbol{x}$, then the decision boundaries are linear

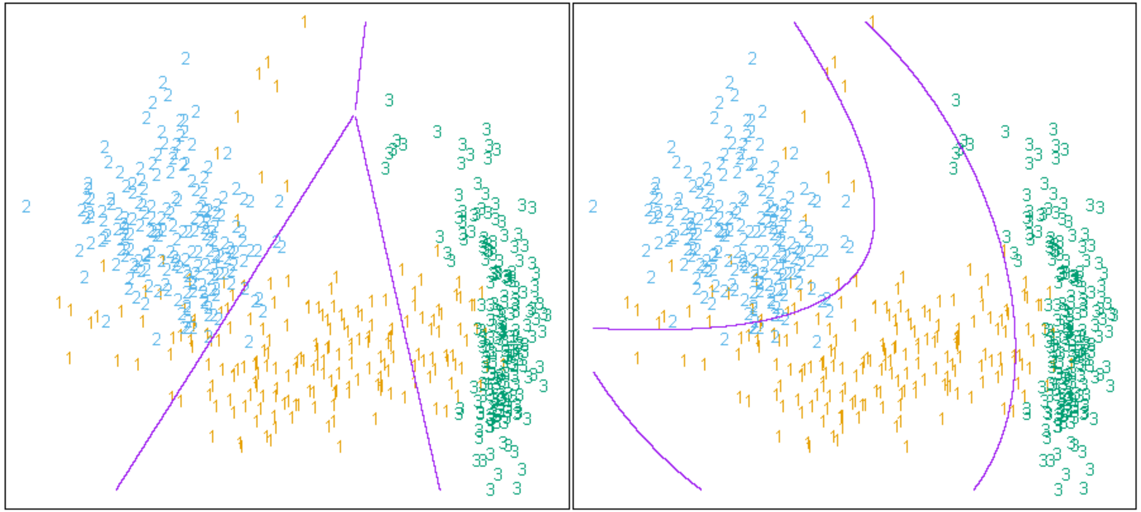
Notation: $G(\boldsymbol{x})$ is the class predictor

Example: A three class example is shown.

- Left Plot: linear decision boundaries obtained through linear discriminant analysis (LDA)
- Right Plot: quadratic decision boundaries obtained by expanding the observation space to five dimensions:

$$X_1, X_2, X_1X_2, X_1^2, X_2^2$$

Note: Linear boundaries in the five-dimension space are quadratic boundaries in the original space



Linear Discriminant Analysis (LDA): Methodology utilizes the class posteriors $\Pr(G|\boldsymbol{X})$ for optimal classification and a Gaussian model.

- Let the set of classes be $G \in \{1, \dots, K\}$
- Define $f_k(\boldsymbol{x})$ as the class-conditional density of $\boldsymbol{X}$ in class $G = k$
  $\implies f_k(\boldsymbol{x}) = P(\boldsymbol{X} = \boldsymbol{x}|G = k)$, where $k \in \{1, \dots, K\}$
- Let π_k be the prior probability of class k
  $\implies \pi_k = P(G = k)$, where $k \in \{1, \dots, K\}$

Recall **Bayes theorem**

$$P(G = k|\boldsymbol{X} = \boldsymbol{x}) = \frac{P(\boldsymbol{X} = \boldsymbol{x}|G = k)P(G = k)}{P(\boldsymbol{X} = \boldsymbol{x})}$$

$$= \frac{P(\boldsymbol{X} = \boldsymbol{x}|G = k)P(G = k)}{\sum_{\ell=1}^{K} P(\boldsymbol{X} = \mathbf{x}|G = \ell)\, P(G = \ell)}$$

Using the previous definitions

$$P(G = k|\boldsymbol{X} = \boldsymbol{x}) = \frac{f_k(\boldsymbol{x})\pi_k}{\sum_{\ell}^{K} f_\ell(\boldsymbol{x})\pi_\ell} \qquad (*)$$

Observation: The denominator in $(*)$ is not a function of k. Thus to determine the class, we need only be concerned with maximizing the numerator.

- Need to determine an appropriate distribution model, e.g., Gaussian, mixtures of Gaussians, nonparametric or other densities

Assumption: Choose a Gaussian mixture model for $f_k(\boldsymbol{x})$

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k{}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)}$$

Also assume that the classes have the same covariance matrix $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}\ \forall k$

In practice we do not know the Gaussian distribution parameter values

$\Longrightarrow$ estimate them using the training data

$\hat{\pi}_k = N_k/N$, where N_k and N are the number of class-k observation and the total number of observations, respectively

$\hat{\mu}_k = \sum_{g_i=k} \mathbf{x}_i / N_k$ (average of $\boldsymbol{x}$ over each category)

$\hat{\Sigma} = \sum_{k=1}^{K} \sum_{g_i=k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T / (N - K)$

We can show that LDA produces linear decision boundaries by taking the log-odds for classes k and ℓ

$$\log\frac{P(G=k|\boldsymbol{X}=\mathbf{x})}{P(G=\ell|\boldsymbol{X}=\mathbf{x})}=\log\frac{f_k(x)}{f_l(x)}+\log\frac{\pi_k}{\pi_\ell}$$
$$=\log\frac{\pi_k}{\pi_\ell}-(\boldsymbol{\mu}_k+\boldsymbol{\mu}_l)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_l)$$
$$+\mathbf{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_l)$$

Let $\beta_0=\log\frac{\pi_k}{\pi_\ell}-(\boldsymbol{\mu}_k+\boldsymbol{\mu}_l)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_l)$ and $\boldsymbol{\beta}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_l)$, then

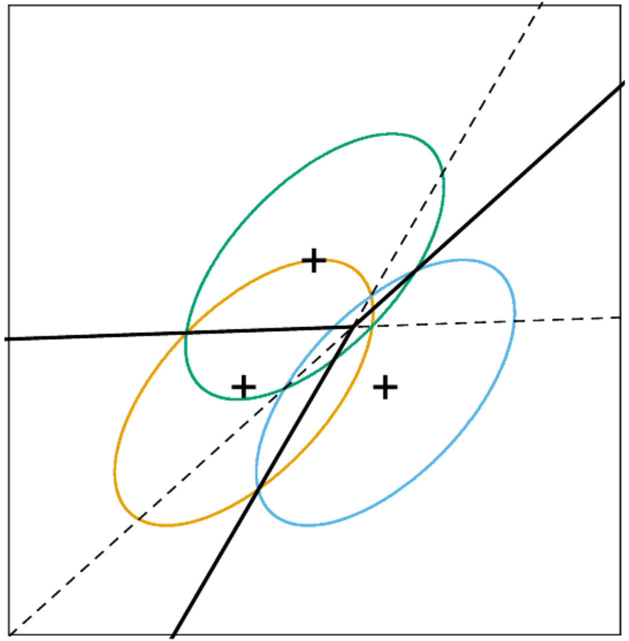
$$\log\frac{P(G=k|\boldsymbol{X}=\mathbf{x})}{P(G=\ell|\boldsymbol{X}=\mathbf{x})}=\beta_0+\mathbf{x}^T\boldsymbol{\beta},$$

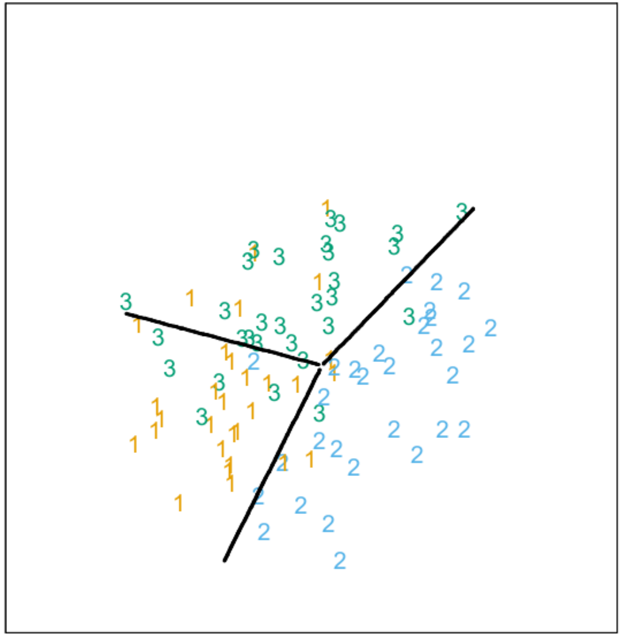
This is a linear equation, $\beta_0+\mathbf{x}^T\boldsymbol{\beta}=0$, that defines the LDA decision boundary between classes k and l

Example:

Left —Three Gaussian distributions with the same covariance and different means. Contours enclosing 95% of the density are shown. Broken lines show class pair decision boundaries. Solid lines show boundaries separating all three classes.

Right — 30 samples from each class and the fitted LDA boundaries.





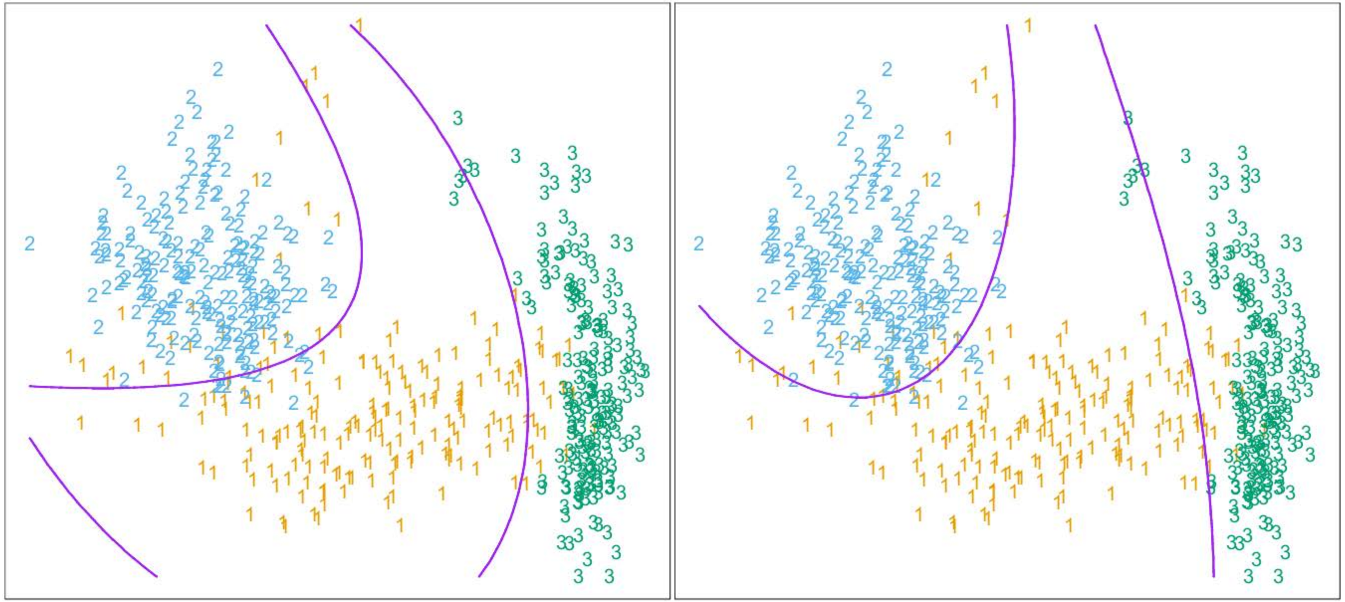
Quadratic Discriminant Analysis (QDA): The covariance matrices $\mathbf{\Sigma}_k$ are taken to be distinct (no equality simplifying assumption)

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k{}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)}$$

Examining the log-odds for classes k and ℓ for QDA

$$\log\frac{P(G=k|\boldsymbol{X}=\mathbf{x})}{P(G=\ell|\boldsymbol{X}=\mathbf{x})} = \log\frac{\pi_k}{\pi_\ell} + \frac{1}{2}\log\frac{\det\mathbf{\Sigma}_\ell}{\det\mathbf{\Sigma}_k} - \frac{1}{2}(\mathbf{x}+\boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)$$
$$-\frac{1}{2}(\mathbf{x}+\boldsymbol{\mu}_\ell)^T\mathbf{\Sigma}_\ell^{-1}(\mathbf{x}-\boldsymbol{\mu}_\ell)$$

- There is no simplification as in LDA
- QDA produces quadratic nonlinear decision boundaries,
- The number of parameters to be estimated is higher in QDA than in LDA



Example: Three classes are utilized. The left plot shows the quadratic decision boundaries obtained using LDA by expanding the observation space to five dimensions: $X_1, X_2, X_1X_2, X_1^2, X_2^2$. The right plot shows the quadratic decision boundaries found by QDA. The differences are small, as is usually the case.

Fisher Approach: Fisher derived LDA without invoking the Gaussian assumption. He posed the problem:

> *Find the linear combination $Z = \boldsymbol{a}^T\boldsymbol{X}$ such that the between-class variance is maximized relative to the within-class variance*

Between-class variance: variance of the class means

Within-class variance: pooled variance about the class means

For $\boldsymbol{X}$, define the within-class covariance, $\boldsymbol{W}$, and between-class covariance, $\boldsymbol{B}$, matrices

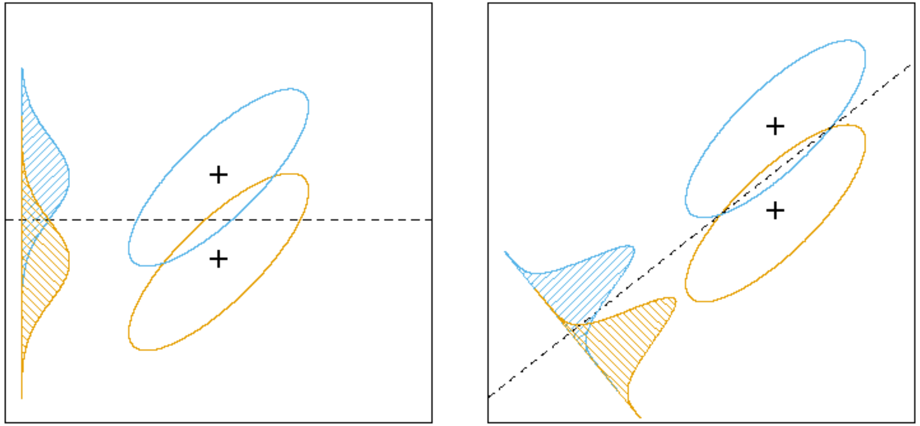
Then the between-class variance of Z is $\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a}$ and the within-class variance is $\boldsymbol{a}^T\boldsymbol{W}\boldsymbol{a}$

Fisher's formulation amounts to maximizing the *Rayleigh quotient*

$$\max_{\boldsymbol{a}} \frac{\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{W}\boldsymbol{a}}$$

Or equivalently $\max_{\boldsymbol{a}} \boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a}$ subject to $\boldsymbol{a}^T\boldsymbol{W}\boldsymbol{a} = 1$

Note: this is a generalized eigenvalue problem, with $\boldsymbol{a}$ given by the largest eigenvalue of $\boldsymbol{W}^{-1}\boldsymbol{B}$



Note:

- The line through the two centroids defines the direction of greatest centroid variance
- The data projected onto that line overlaps
- The Fisher discrimination direction minimizes the overlap of projected data (Gaussian case)

Summary:

- Classification is the prediction of qualitative responses
- Linear methods for classification result in linear decision boundaries
- LDA utilizes the class posteriors $\mathrm{Pr}(G|\boldsymbol{X})$ for optimal classification and a Gaussian distribution model
  – Yields a linear boundary function, $\beta_0 + \mathbf{x}^T\boldsymbol{\beta}$
- Fisher derived the LDA result without invoking the Gaussian assumption
  – Find the linear combination $Z = \boldsymbol{a}^T\boldsymbol{X}$ such that the between-class variance is maximized relative to the within-class variance