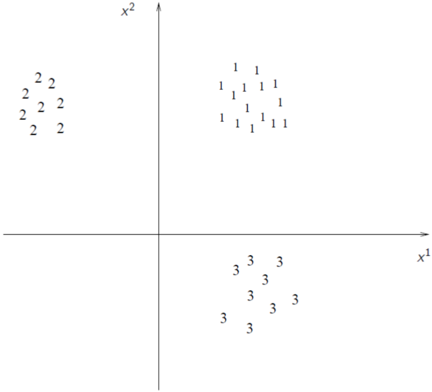
Data Science
Static data analysis

Stéphane Marchand-Maillet

Master en Sciences Informatiques - Semestre d'Automne

Linear Discriminant Analysis (LDA)

We look at the issue of modeling multivariate data (p quantitatives components and one categorical variable). Every data is described by $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i = 1, \ldots, q$.



Linear Discriminant Analysis

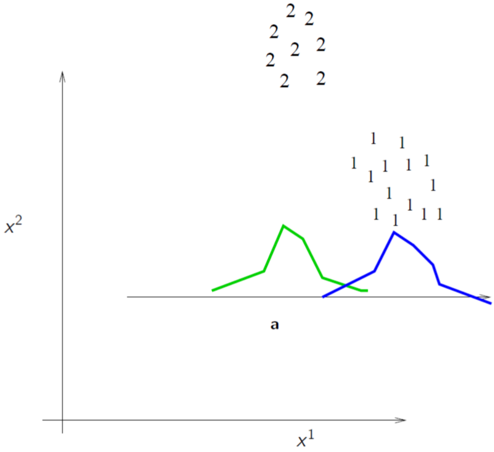
Class definition

Alternative interpretation : The categorical variable y_i describes the class to which data i belongs, characterised by variables $\mathbf{x}_i$.
Hence, point i belongs to class C_k iff $y_i = k$.

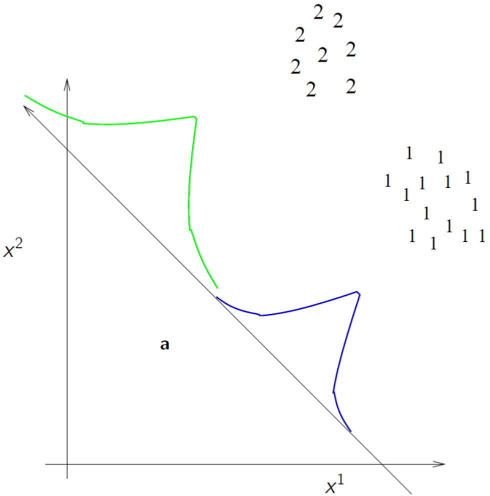
Discriminant Analysis

Can the q classes be discriminated over the space of variables $\mathbf{x}$?
Is there a linear transform of $\mathbf{x}$ such that the q classes are better separated?
$\Rightarrow$ basis for supervised learning

2 Classes - Discriminant Axis



2 Classes - Discriminant Axis



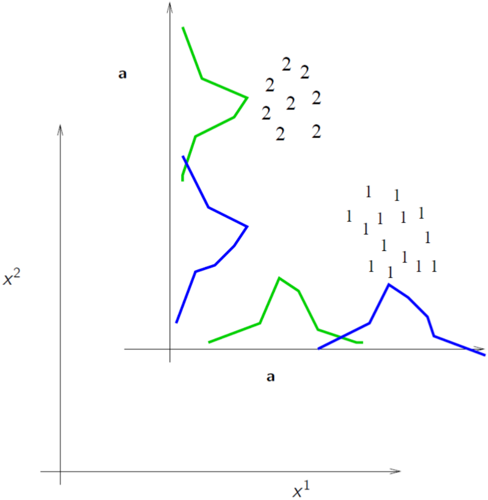
Inter-classs discrimination criterion

- ⋆ Search for direction $\mathbf{a}$ where the *inter-class* discrimination is maximum.
- ⋆ Clearly, $\mathbf{a}$ must be parallel to the line $(\mathbf{g}_1, \mathbf{g}_2)$ across the centers of mass of the classes, since:

$$(\hat{\mathbf{g_1}} - \hat{\mathbf{g_2}})^2 = \left(\frac{\mathbf{a}^\mathsf{T}\mathbf{g_1}}{||\mathbf{a}||} - \frac{\mathbf{a}^\mathsf{T}\mathbf{g_2}}{||\mathbf{a}||}\right)^2 = \left(\frac{\mathbf{a}^\mathsf{T}}{||\mathbf{a}||}(\mathbf{g_1} - \mathbf{g_2})\right)^2$$

  is maximum when $\mathbf{a} \propto \mathbf{g_1} - \mathbf{g_2}$

Intra-class discrimination criterion



Intra-class discrimination criterion

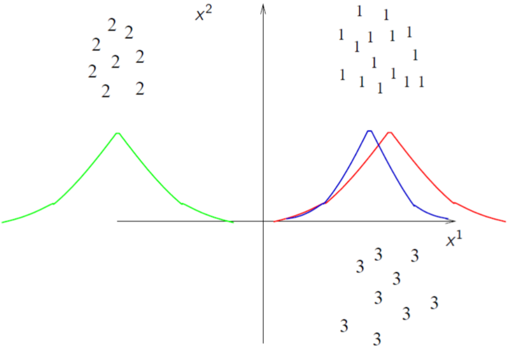
- We must account for the *intra*-class variance of the projected data
- Fisher criterion maximises

$$\max_{\mathbf{a}} \frac{(\hat{\mathbf{g}_1} - \hat{\mathbf{g}_2})^2}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}$$

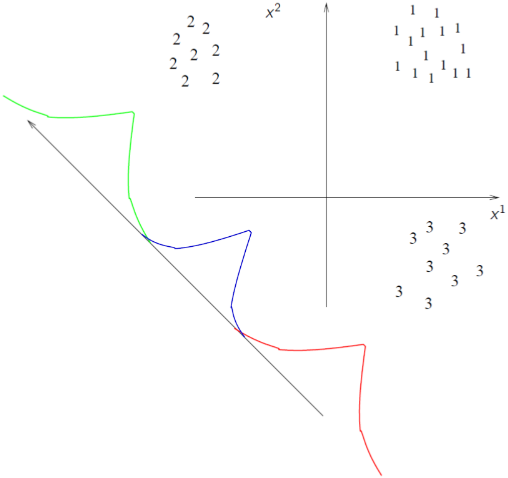
where $\hat{\sigma_k}$ is the normalised variance of the projection of class k

$$\hat{\sigma}_k^2 = \sum_{\mathbf{x} \in \mathbf{C_k}} (\hat{\mathbf{x}} - \hat{\mathbf{g_k}})^\mathsf{T} (\hat{\mathbf{x}} - \hat{\mathbf{g_k}})$$

Discriminant axes



Discriminant axes



Generalisation: intra-classe criteria

- Let $A_k = [\mathbf{x}_1 - \mathbf{g}_k, \ldots, \mathbf{x}_{n_k} - \mathbf{g}_k], x_i \in C_k$ be the matrix of centered data
- $\frac{1}{n_k} A_k A_k^\mathsf{T}$ is the *intra*-class covariance matrix
- $S_w = \sum_k \frac{1}{n_k} A_k A_k^\mathsf{T}$ is the sum of *intra*-class covariance matrices
- We minimise

$$\sum_k \sum_{x_i \in c_k} (\hat{\mathbf{x}}_\mathbf{i} - \hat{\mathbf{g}}_k)^\mathsf{T} (\hat{\mathbf{x}}_\mathbf{i} - \hat{\mathbf{g}}_k) = \sum_k \sum_{x_i \in c_k} \frac{(\mathbf{a}^\mathsf{T}(\mathbf{x}_\mathbf{i} - \mathbf{g}_k))^\mathsf{T} \mathbf{a}^\mathsf{T}(\mathbf{x}_\mathbf{i} - \mathbf{g}_k)}{||\mathbf{a}||^2}$$

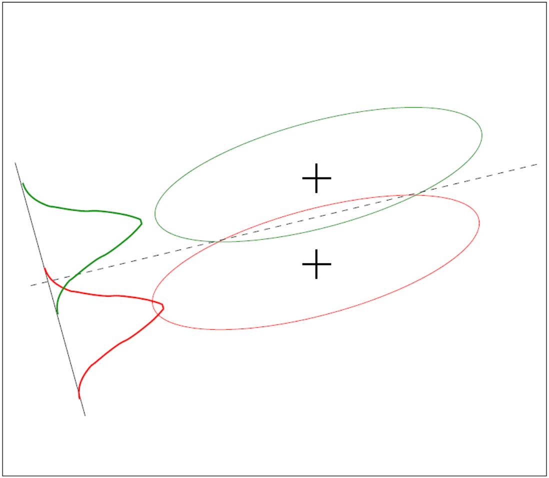
$$= \sum_k \frac{1}{||\mathbf{a}||^2} \mathbf{a}^\mathsf{T} A_k A_k^\mathsf{T} \mathbf{a} = \frac{1}{||\mathbf{a}||^2} \mathbf{a}^\mathsf{T} S_w \mathbf{a}$$

Generalisation: inter-class criteria

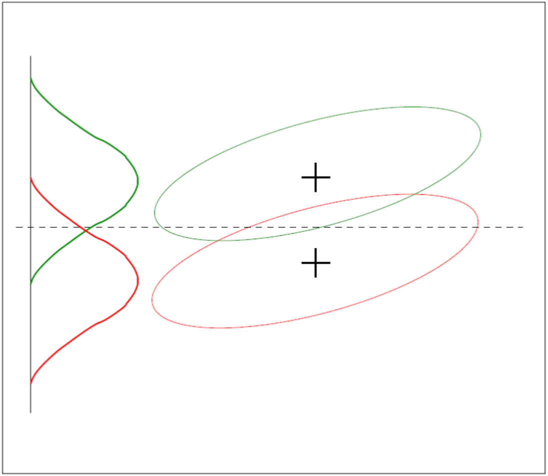
- Let $B = [\mathbf{g}_1 - \mathbf{g}, \ldots, \mathbf{g}_q - \mathbf{g}]$ be the matrix of centered data centers ($\mathbf{g} = \frac{1}{N} \sum_N \mathbf{x_i}$ and $N = \sum_k n_k$)
- $S_b = \frac{1}{q} BB^\mathsf{T}$ is the covariance matrix of class centers
- We maximise

$$\sum_k (\hat{\mathbf{g}}_k - \hat{\mathbf{g}})^\mathsf{T} (\hat{\mathbf{g}}_k - \hat{\mathbf{g}}) = \frac{1}{||\mathbf{a}||^2} \mathbf{a}^\mathsf{T} S_b \mathbf{a}$$

Mixing both criteria



Mixing both criteria



Fisher discrimination criteria: Raleigh coefficient

- ⋆ Combining both

$$max_{\mathbf{a}} J_{\mathbf{a}} = max_{\mathbf{a}} \frac{\mathbf{a}^{\mathsf{T}} S_b \mathbf{a}}{\mathbf{a}^{\mathsf{T}} S_w \mathbf{a}}$$

- ⋆ which is found if:

$$\frac{\partial J_{\mathbf{a}}}{\partial \mathbf{a}} = \frac{S_b \mathbf{a}(\mathbf{a}^{\mathsf{T}} S_w \mathbf{a}) - S_w \mathbf{a}(\mathbf{a}^{\mathsf{T}} S_b \mathbf{a})}{(\mathbf{a}^{\mathsf{T}} S_w \mathbf{a})^2} = 0$$

$\Rightarrow$ **a** is solution of the generalised eigen system: $S_b \mathbf{a} = J_{\mathbf{a}} S_w \mathbf{a}$

- ⋆ Hence, **a** is the first e.v of $S_w^{-1} S_b$

Discriminant subspaces

- ⋆ eigenvectors corresponding to the largest eigenvalues λ_i are the most discriminative dimensions

$$\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_p \text{ avec } \lambda_1 > \lambda_2 > \ldots \lambda_p$$

- ⋆ q classes may be discriminated in a (at most) $(q-1)$-dimensional subspace

$\Rightarrow$ only $q-1$ non-zero eigenvalues

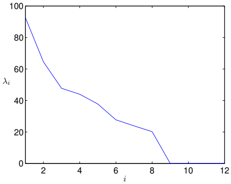
Particular case: 2 classes

- $BB^{\mathsf{T}} = (\mathbf{g}_1 - \mathbf{g})(\mathbf{g}_1 - \mathbf{g})^{\mathsf{T}} + (\mathbf{g}_2 - \mathbf{g})(\mathbf{g}_2 - \mathbf{g})^{\mathsf{T}} = (\mathbf{g}_1 - \mathbf{g}_2)(\mathbf{g}_1 - \mathbf{g}_2)^{\mathsf{T}}$
- hence $BB^{\mathsf{T}}\mathbf{a}$ is a vector along direction $(\mathbf{g}_1 - \mathbf{g}_2)$
- hence $\mathbf{a} \simeq S_w^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$

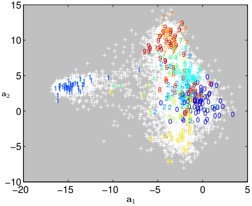
Illustrations : character recognition

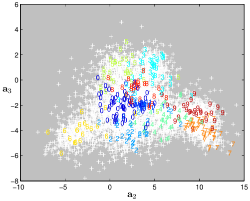
7291 images 16×16 (8 bits) numbers from 0 to 9

$\Rightarrow \{\mathbf{x}_i, y_i\}$ avec $\mathbf{x}_i \in \mathbb{R}^{256}$ et $y_i = 1, \ldots, 10$, $i = 1 \ldots 7291$



Projection





⇒ LDA finds the optimal subspace to (linearly) separate data along labels y_i.

LDA as a support for decision making

- ⋆ New data $j \rightarrow \mathbf{x}_j$ known, y_j unknown
- ⋆ To which class C_k point j belongs? (classification)

$\Rightarrow$ Predict $P(C_k|\mathbf{x}_j)$ (Bayes rule):

$$P(C_k|\mathbf{x}_j) = \frac{P(\mathbf{x}_j|C_k)P(C_k)}{P(\mathbf{x}_j)}$$

Gaussian approximation

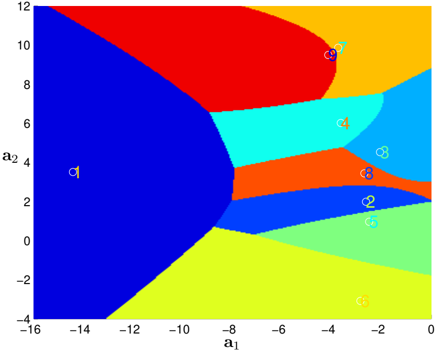
- ⋆ Each class is modeled by $\mathcal{N}(\mu_k, W_k)$
- ⋆ Prior: $P(C_k) = 1/q$
- ⋆ evidence $P(\mathbf{x}_j)$ is ignored

$\Rightarrow$ Maximun likelihood

$$p(\mathbf{x}|C_k) \approx \exp\left(-(\mathbf{x}-\mu_k)^\mathsf{T} W_k (\mathbf{x}-\mu_k)\right)$$

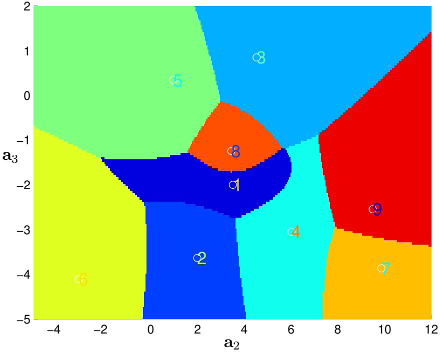
Decision (classification)

$$\delta(\mathbf{x}) = \arg\max_k P(\mathbf{x}|C_k)$$



Decision (classification)

$$\delta(\mathbf{x}) = \arg\max_k P(\mathbf{x}|C_k)$$



Optimality

- ⋆ LDA is optimal when the q classes are each Gaussian distributed

$\Rightarrow$ because of the discrimination criteria based on covariance matrices S_w et S_b

- ⋆ Linear discriminant Analysis $\rightarrow$ does not account for non-linear relationships between variables