Critérios para Discriminação de Duas Classes

2. **Critério de Fisher**

- Este critério procura a maior separação possível entre as classes através de uma combinação linear das variáveis.
- Considere ω_1 com n_1 pontos e ω_2 com n_2 pontos. Os vetores de médias de cada classe são

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{n \in \omega_i} \mathbf{x}_i \qquad (n = n_1 + n_2)$$

- A forma mais simples de separar as classes é escolher um vetor de projeção $\mathbf{w}$ que maximiza a distância entre as médias $\mathbf{m}_1$ e $\mathbf{m}_2$, i.e., $\max\ \mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_1)$.
- Para restringir a magnitude de $\mathbf{w}$, fazemos $\mathbf{w}$ com comprimento unitário, i.e., $\sum_i w_i^2 = 1$.

- O critério de Fisher maximiza a razão das variâncias entre e dentro das classses, obtida pela projeção $\mathbf{w}$, i.e.,

$$\mathrm{J}_F = \frac{||\mathbf{w}^\top(\mathbf{m}_2 - \mathbf{m}_1)||^2}{\mathbf{w}^\top \mathbf{S}_w \mathbf{w}} = \frac{\mathbf{w}^\top \mathbf{S}_B \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_w \mathbf{w}}$$

  em que

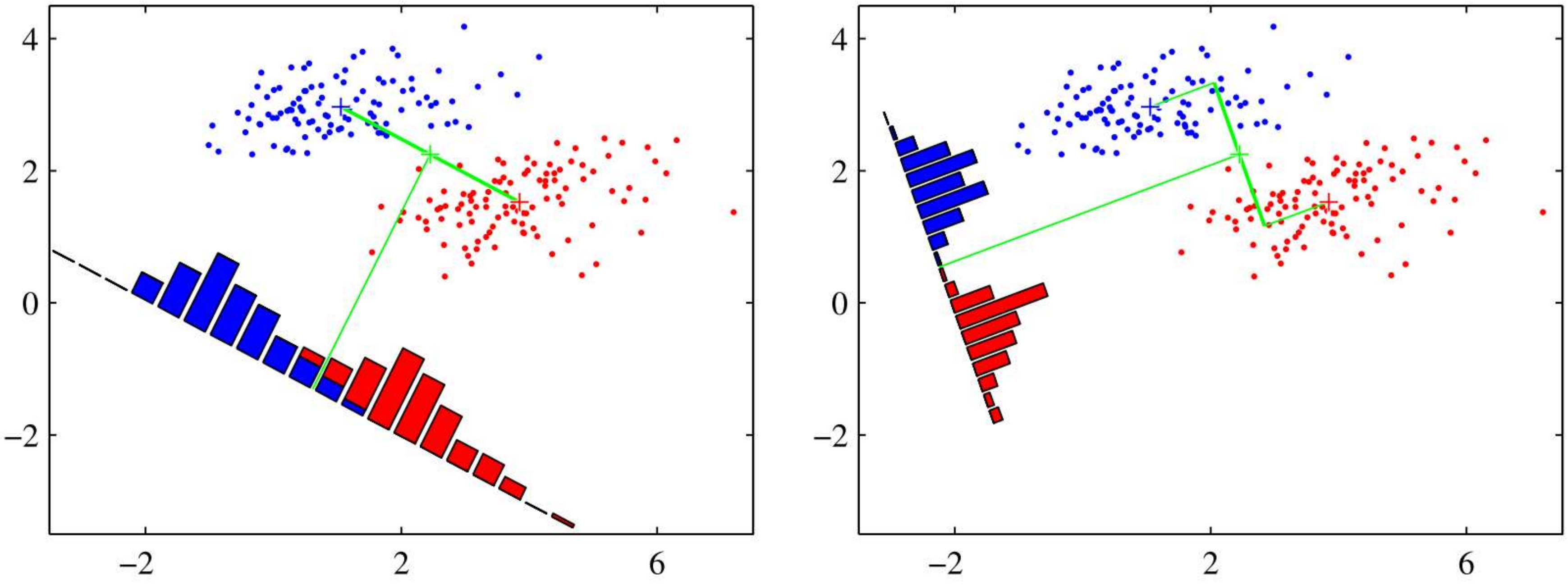
$$\mathbf{S}_w = \frac{1}{n-2}\left(n_1 \hat{\boldsymbol{\Sigma}}_1 + n_2 \hat{\boldsymbol{\Sigma}}_2\right)$$

  Nota: $\hat{\boldsymbol{\Sigma}}_i$ é a estimativa de máxima verossimilhança da matriz de covariância da classe ω_i.

- A solução para $\mathbf{w}$ que maximiza J_F é obtida através da diferenciação de J_F em relação a $\mathbf{w}$ e igualando a zero, resultando em

$$\frac{2\mathbf{w}^\top(\mathbf{m}_2 - \mathbf{m}_1)}{\mathbf{w}^\top \mathbf{S}_w \mathbf{w}} \left\{ (\mathbf{m}_2 - \mathbf{m}_1) - \frac{\mathbf{w}^\top(\mathbf{m}_2 - \mathbf{m}_1)}{\mathbf{w}^\top \mathbf{S}_w \mathbf{w}} \mathbf{S}_w \mathbf{w} \right\} = 0.$$

- O critério de Fisher não fornece uma regra de alocação, mas apenas uma projeção em que a discriminação das classes é mais fácil de ser realizada.



Esquerda: projeção com superposição das classes. Direita: projeção segundo critério de Fisher. (Bishop, 2006)

- O critério de Fisher não fornece uma regra de alocação, mas apenas uma projeção em que a discriminação das classes é mais fácil de ser realizada.
- Para obter um critério de alocação precisamos especificar um limite (*threshold*) w_0 e alocar $\mathbf{x}$ to ω_1 se

$$\mathbf{w}^T\mathbf{x} + w_0 > 0$$

Exemplo de critério alocação

Considere que os dados são normalmente distribuídos e com $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$.

Alocar $\mathbf{x}$ a ω_1 se $\mathbf{w}^T\mathbf{x} + w_0 > 0$ tal que

$\mathbf{w} = \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$ e

$w_0 = -\frac{1}{2}(\mathbf{m}_1 + \mathbf{m}_2)^T\mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2) - \log\left(\frac{p(\omega_2)}{p(\omega_1)}\right)$

Observação: utilizamos normalidade para especificar w_0 e obter uma solução ótima. No caso de não normalidade, a solução não é necessariamente ótima.

Critérios para Discriminação de Duas Classes

3. **Critério de Mínimos Quadrados**

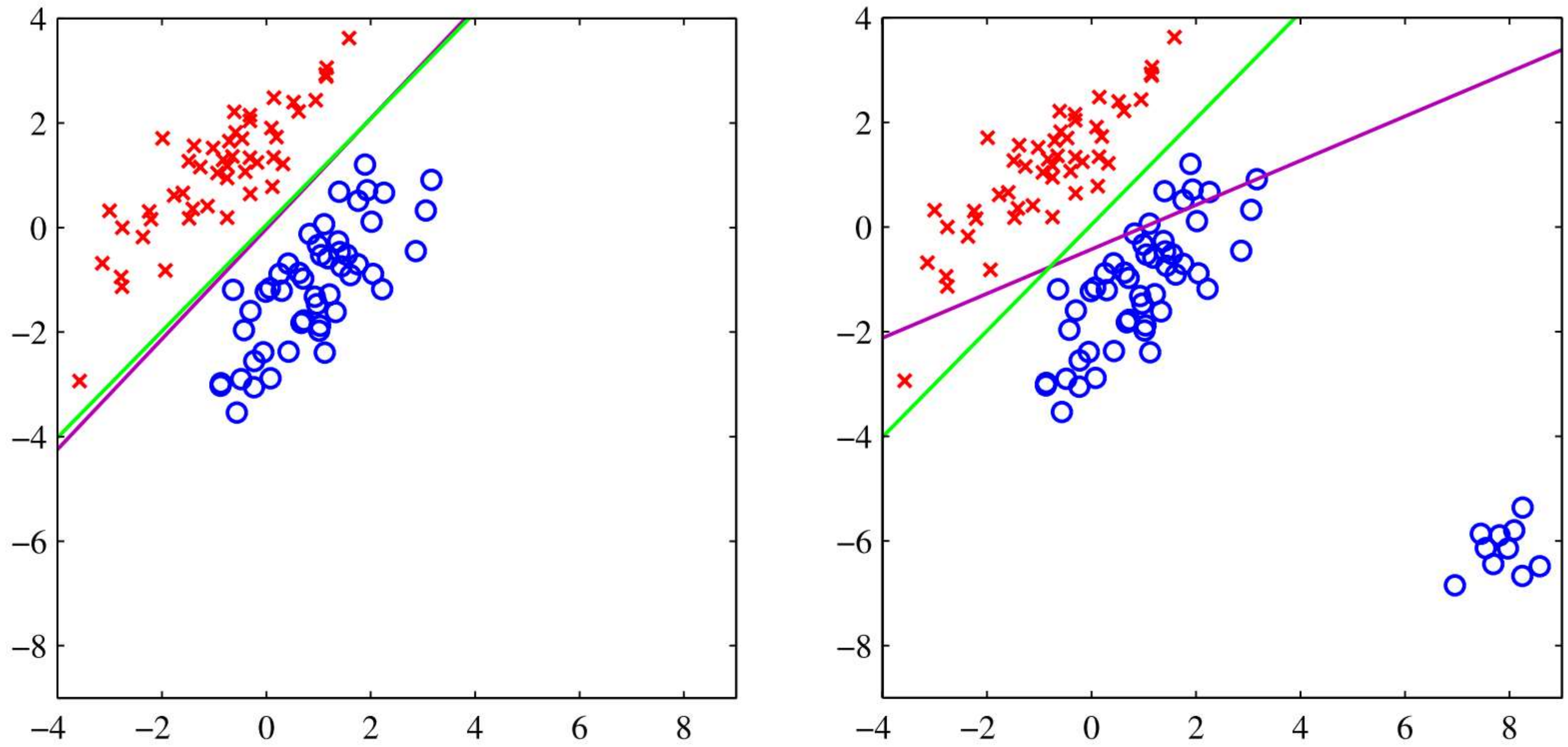
- Este critério utiliza todos os dados da amostra para encontrar uma solução para a qual
$$\boldsymbol{\nu}^{\mathsf{T}}\mathbf{y}_i = \mathbf{t}_i$$
é satisfeita para constantes positivas $\mathbf{t}_i$.
- Lembre que $\mathbf{y}_i^{\mathsf{T}} = (1, \mathbf{x}_i^{\mathsf{T}})$ para $\mathbf{x}_i \in \omega_1$ e $\mathbf{y}_i^{\mathsf{T}} = (-1, -\mathbf{x}_i^{\mathsf{T}})$ para $\mathbf{x}_i \in \omega_2$.
- A solução é obtida por mínimos quadrados, i.e., minimizando
$$\mathrm{J}_S = ||\mathbf{Y}\boldsymbol{\nu} - \mathbf{t}||^2.$$
- De Análise de Regressão, se $\mathbf{Y}^{\mathsf{T}}\mathbf{Y}$ é não singular,
$$\boldsymbol{\nu} = (\mathbf{Y}^{\mathsf{T}}\mathbf{Y})^{-1}\mathbf{Y}^{\mathsf{T}}\mathbf{t}.$$
- Outra solução para $\boldsymbol{\nu}$, minimizando J_S é
$$\boldsymbol{\nu} = \mathbf{Y}^{-}\mathbf{t},$$
em que $\mathbf{Y}^{-}$ é a pseudo-inversa (ou inversa generalizada) de $\mathbf{Y}$.

- A inversa generalizada $\mathbf{A}^-$ pode ser obtida da decomposição em valores singulares de $\mathbf{A}$.

Decomposição em Valores Singulares (SVD)

- Uma matriz $\mathbf{A}_{n\times d}$ (valores reais) pode ser sempre escrita como $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i \mathbf{v}_i^\top$ em que
  - r é o posto da matriz $\mathbf{A}$;
  - $\mathbf{U}$ é uma matriz de colunas ortonormais contendo n autovetores da matriz $\mathbf{A}\mathbf{A}^\top$ ("left singular vectors");
  - $\mathbf{V}$ é uma matriz de colunas ortonormais contendo d autovetores da matriz $\mathbf{A}^\top\mathbf{A}$ ("right singular vectors");
  - $\mathbf{D}$ é uma matriz $n \times d$ diagonal com $\sqrt{\lambda_i}$ em que $\lambda_i's$ ($\min\{n, d\}$) são os maiores autovalores de $\mathbf{A}\mathbf{A}^\top$ ou $\mathbf{A}^\top\mathbf{A}$, chamados de valores singulares.

- A inversa generalizada será $\mathbf{A}^- = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^\top = \sum_{i=1}^{r} \frac{1}{\lambda_i} \mathbf{v}_i \mathbf{u}_i^\top$.
- Propriedades:
  - $\mathbf{A}\mathbf{A}^-$ e $\mathbf{A}^-\mathbf{A}$ são simétricas;
  - $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$;
  - $\mathbf{A}^-\mathbf{A}\mathbf{A}^- = \mathbf{A}^-$.



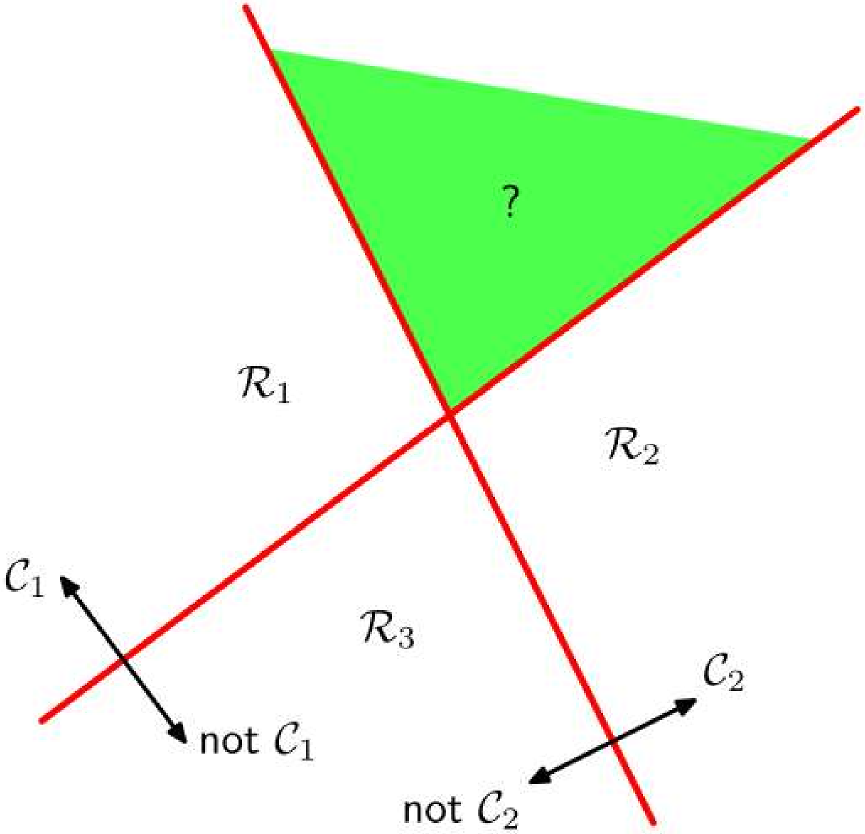
O critério de mínimos quadrados é fortemente influenciado por *outliers*. Linha rosa indica limite obtido por mínimos quadrados e linha verde por regressão logística. (Bishop, 2006)

- Se $t_i = t_1$ (constante) para todo $\mathbf{x}_i \in \omega_1$ e $t_i = t_2$ (constante) para todo $\mathbf{x}_i \in \omega_2$, a solução de mínimos quadrados corresponde ao discriminante de Fisher.
- Observação: usual fazer $t_1 + t_2 \neq 0$.

Algoritmos para 3 ou mais Classes

Ideias Gerais: considere $c > 2$ classes para classificar.

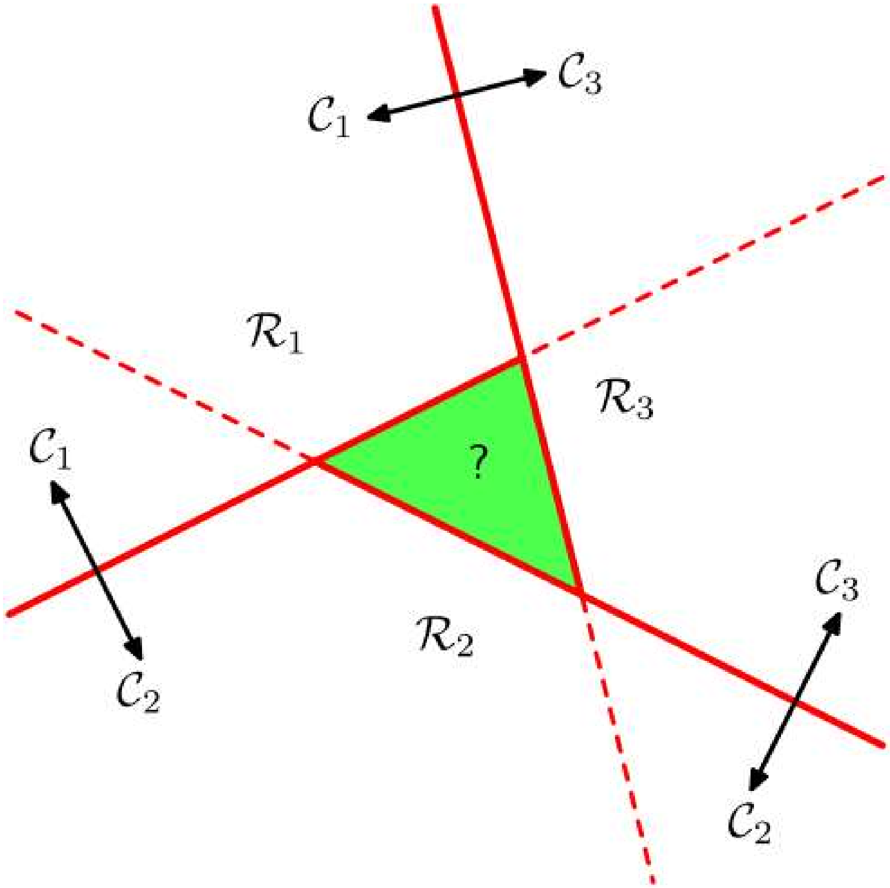
1. **Uma classe contra as demais:** utiliza $c - 1$ classificadores e para cada um deles resolve o problema de separação de pontos de uma classe ω_i $(i = 1, \ldots, c)$ e os pontos que não pertencem a esta classe.
   Problema: pode gerar regiões no espaço com pontos que não possuem uma classe definida.



(Bishop, 2006)

2. **Pares de classes:** utiliza $c(c-1)/2$ discriminantes binários para comparar todos os possíveis pares de classes.

   Problema: número muito grande de comparações quando temos muitas classes. Também pode gerar regiões no espaço com pontos que não possuem uma classe definida.



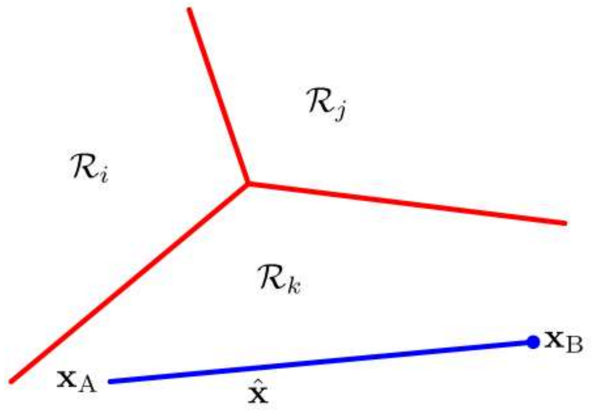
(Bishop, 2006)

3. **Discriminantes para c funções lineares:**

- Função: $y_i(\mathbf{x}) = \mathbf{w}_i^{\mathsf{T}}\mathbf{x} + w_{i0}$
- Alocar o ponto $\mathbf{x}$ a classe ω_i se $y_i(\mathbf{x}) > y_j(\mathbf{x})$ para todo $i \neq j$ em $\{1, \ldots, c\}$.
- O limite (*boundary*) será dado por $y_i(\mathbf{x}) = y_j(\mathbf{x})$ e corresponde ao hiperplano definido por

$$(\mathbf{w}_i - \mathbf{w}_j)^{\mathsf{T}}\mathbf{x} + (w_{i0} - w_{j0}) = 0.$$

- As regiões definidas pelos discriminantes são conectadas e convexas.



Regiões convexas: Seja $0 \leq \lambda \leq 1$. Então $y_k(\hat{\mathbf{x}}) = \lambda y_k(\mathbf{x}_A) + (1-\lambda)y_k(\mathbf{x}_B)$. (Bishop, 2006)

- Webb e Copsey (2011) descrevem os algoritmos de Fisher e Mínimos Quadrados para o caso de $c > 2$ classes.