Clustering

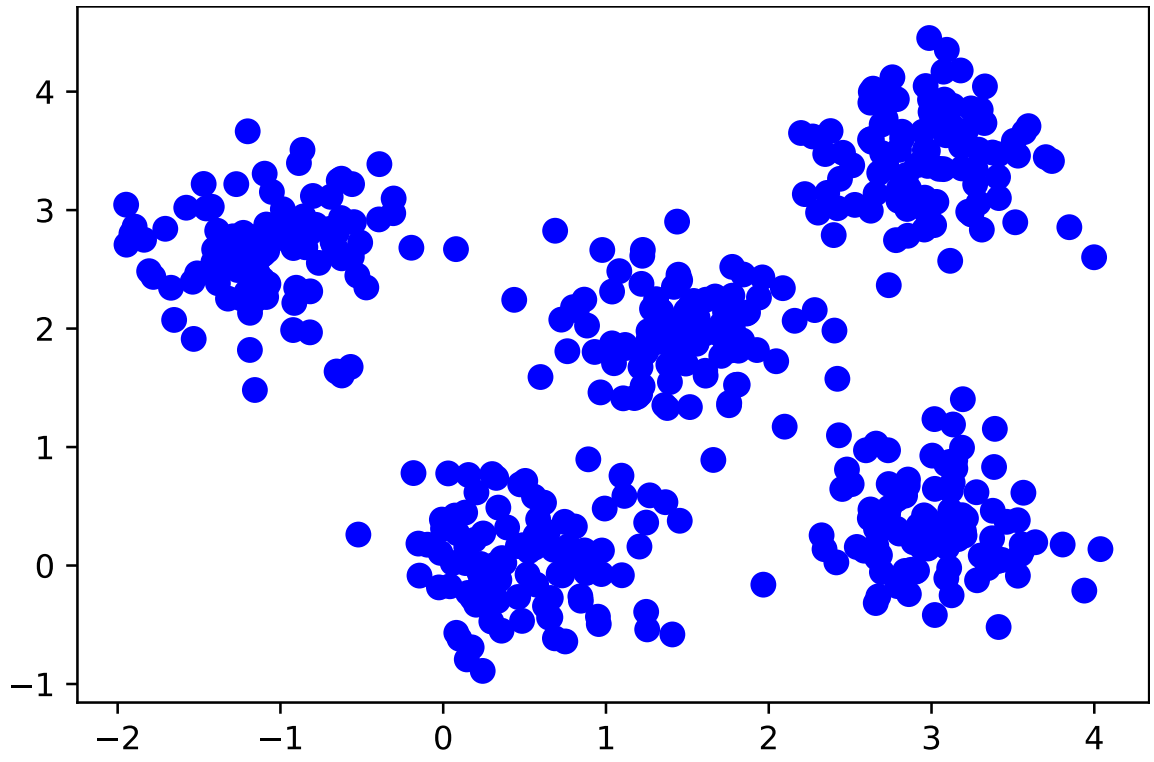
Prof. Danilo Silva

EEL7514/EEL7513 - Tópico Avançado em Processamento de Sinais
EEL410250 - Aprendizado de Máquina

EEL / CTC / UFSC

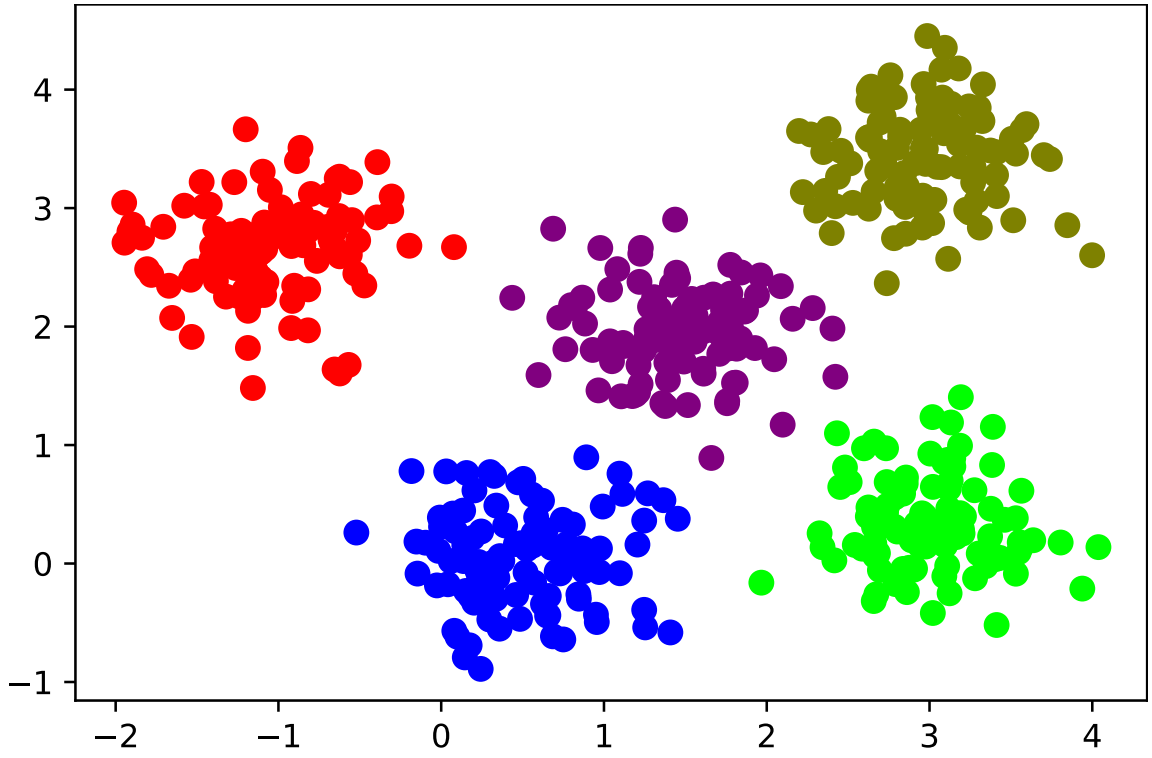
Introdução

Clustering



- Conjunto de dados não-rotulados: $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)} \in \mathbb{R}^n$
- Problema: separar os dados em K grupos (clusters) de amostras "similares", i.e., que estejam mais "próximas" das amostras do mesmo grupo do que das de outros grupos

Clustering



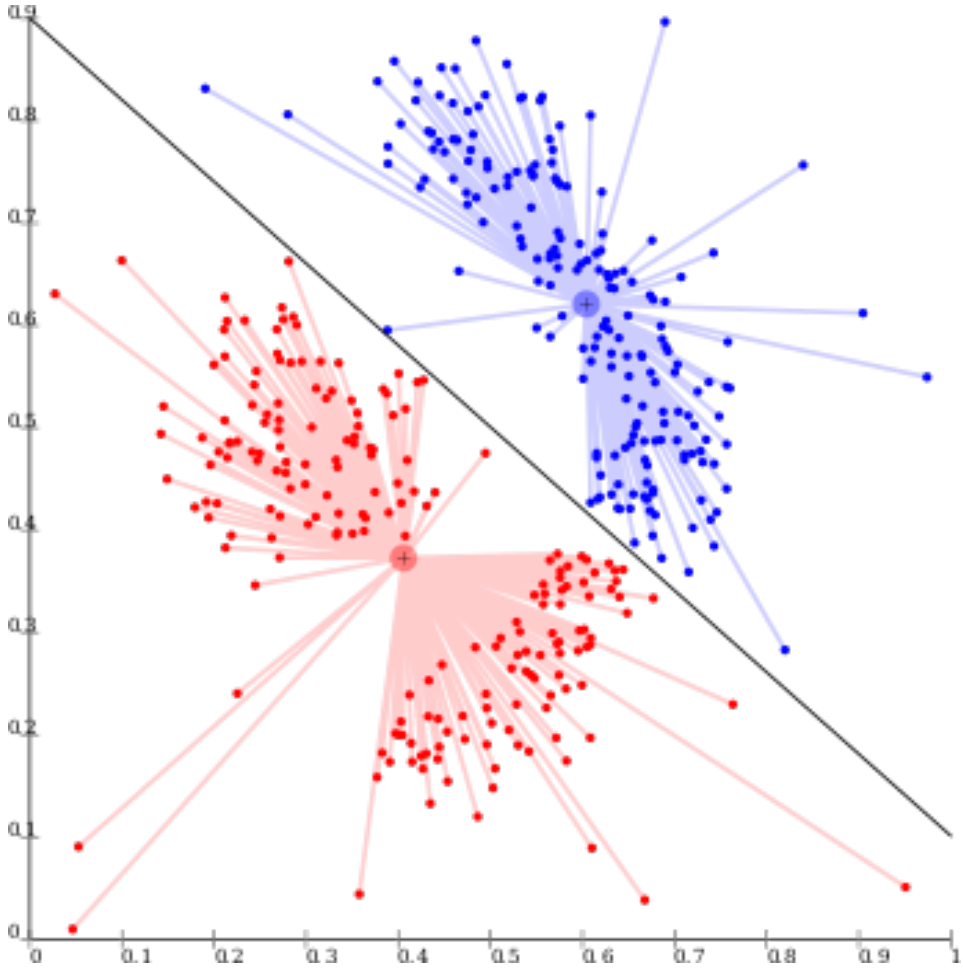
- Conjunto de dados não-rotulados: $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)} \in \mathbb{R}^n$
- Problema: separar os dados em K grupos (clusters) de amostras "similares", i.e., que estejam mais "próximas" das amostras do mesmo grupo do que das de outros grupos

Clustering

- Aplicações:
  - Segmentação de mercado / posicionamento de produto
  - Agrupamento de resultados de busca
  - Identificação de famílias de genes
  - Segmentação de imagens
- Abordagens:
  - Clustering hierárquico/aglomerativo
  - Clustering baseado em centróides (ex: k-means)
  - Clustering baseado em distribuição de probabilidade (ex: GMM)
  - Clustering baseado em densidade de pontos (ex: DBSCAN)
  - Clustering baseado em grafos
  - ...

Algoritmo K-means

K-means



- Objetivo: Determinar K representantes dos clusters (centróides) e alocar amostras em clusters de forma a minimizar a soma das distâncias quadráticas de cada amostra ao representante do seu cluster

K-means

- Notação:
  - $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)} \in \mathbb{R}^n$: amostras/pontos
  - K: número de clusters (escolhido a priori)
  - $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K \in \mathbb{R}^n$: médias/centróides (representantes dos clusters)
  - $c^{(i)} \in \{1, \ldots, K\}$: índice do cluster ao qual a amostra $\mathbf{x}^{(i)}$ está atribuída
  - $\mathcal{S}_k = \{\mathbf{x}^{(i)} : c^{(i)} = k\}$: k-ésimo cluster
- Função custo:

$$J(c^{(1)}, \ldots, c^{(m)}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K) = \sum_{i=1}^{m} \|\mathbf{x}^{(i)} - \boldsymbol{\mu}_{c^{(i)}}\|^2$$

$$= \sum_{k=1}^{K} \sum_{\mathbf{x} \in \mathcal{S}_k} \|\mathbf{x} - \boldsymbol{\mu}_k\|^2$$

Otimização Alternada

- Se os centróides $\boldsymbol{\mu}_k$ estão fixos, a solução ótima da atribuição é

$$c^{(i)} = \underset{k}{\operatorname{argmin}} \|\mathbf{x}^{(i)} - \boldsymbol{\mu}_k\|^2$$

  isto é, atribui-se $\mathbf{x}^{(i)}$ ao cluster cujo centróide esteja mais próximo

- Se as atribuições $c^{(i)}$ estão fixas, a solução ótima para $\boldsymbol{\mu}_k$ é

$$\boldsymbol{\mu}_k = \frac{1}{|\mathcal{S}_k|} \sum_{\mathbf{x} \in \mathcal{S}_k} \mathbf{x}$$

  isto é, escolhe-se $\boldsymbol{\mu}_k$ como sendo a média (centróide) das amostras pertencentes ao cluster k

- Alternar estas otimizações nunca pode aumentar o custo

Algoritmo

- Inicialize aleatoriamente $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K \in \mathbb{R}^n$
- Repita até a convergência:
  - Para $i = 1, \ldots, m$:

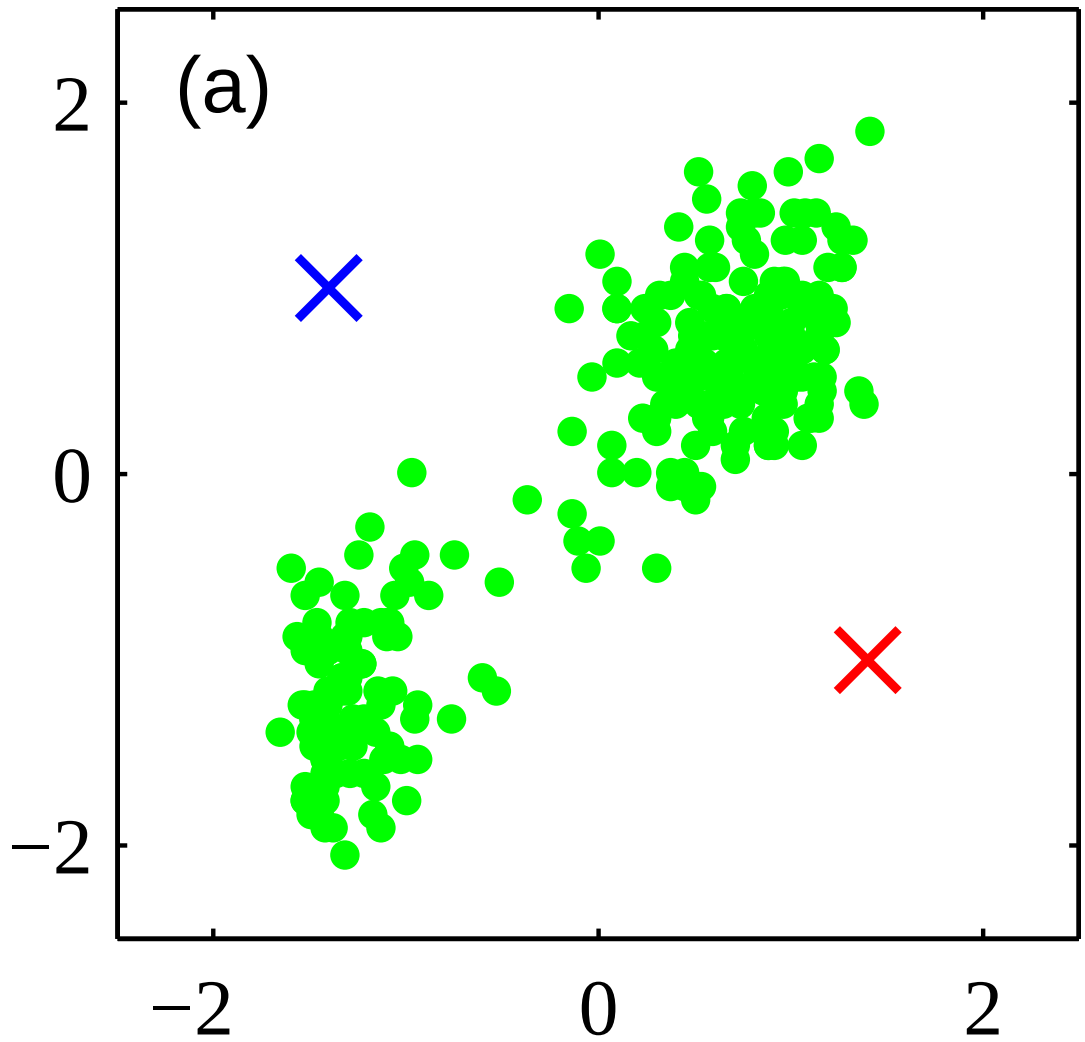
$$c^{(i)} = \operatorname*{argmin}_k \|\mathbf{x}^{(i)} - \boldsymbol{\mu}_k\|^2$$

  - Para $k = 1, \ldots, K$:

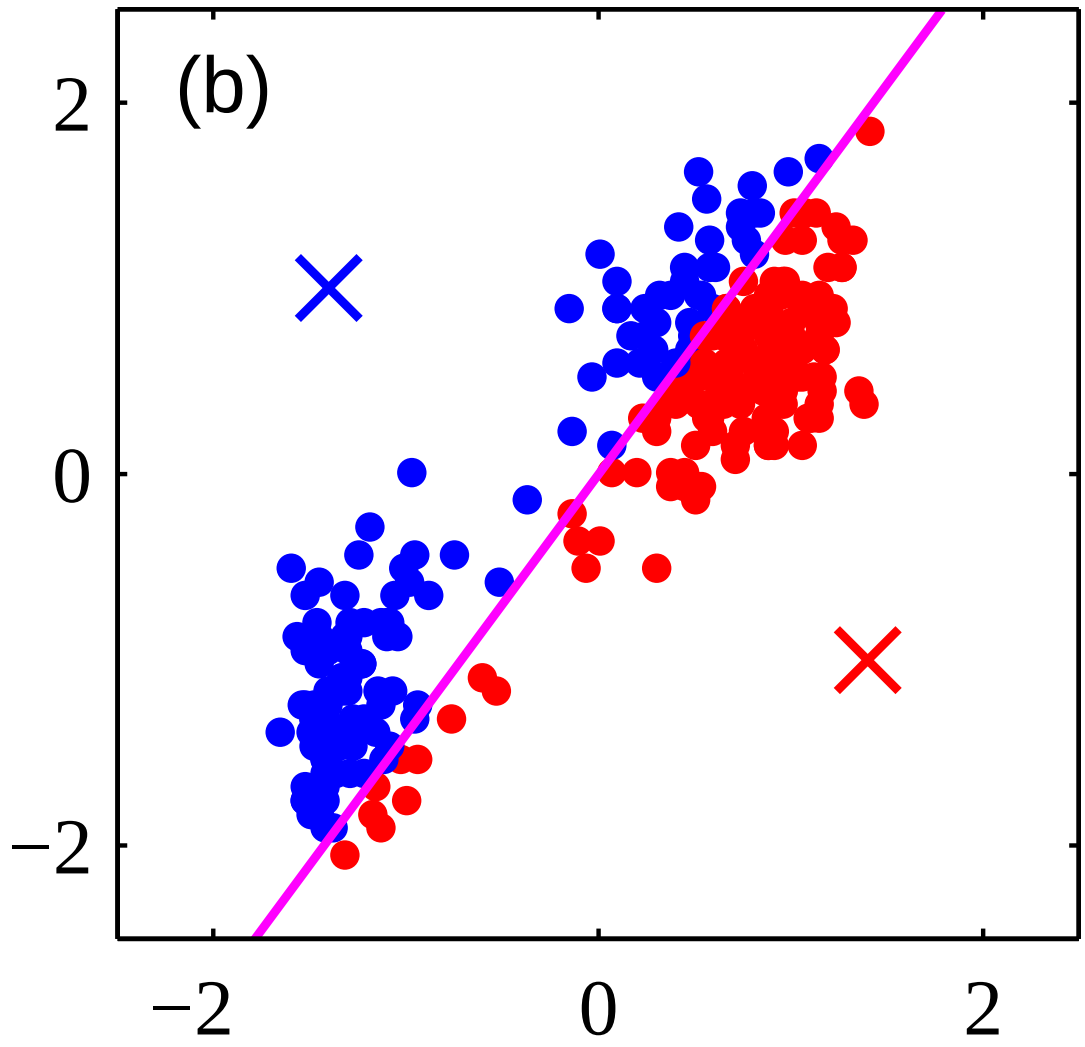
$$\boldsymbol{\mu}_k = \frac{1}{|\{i : c^{(i)} = k\}|} \sum_{i:c^{(i)}=k} \mathbf{x}^{(i)}$$

- Obs: o algoritmo sempre converge, mas não necessariamente para o ótimo global
- Normalmente utilizado com múltiplas reinicializações
  - Escolhe-se a melhor de N tentativas (ex: $N = 100$)

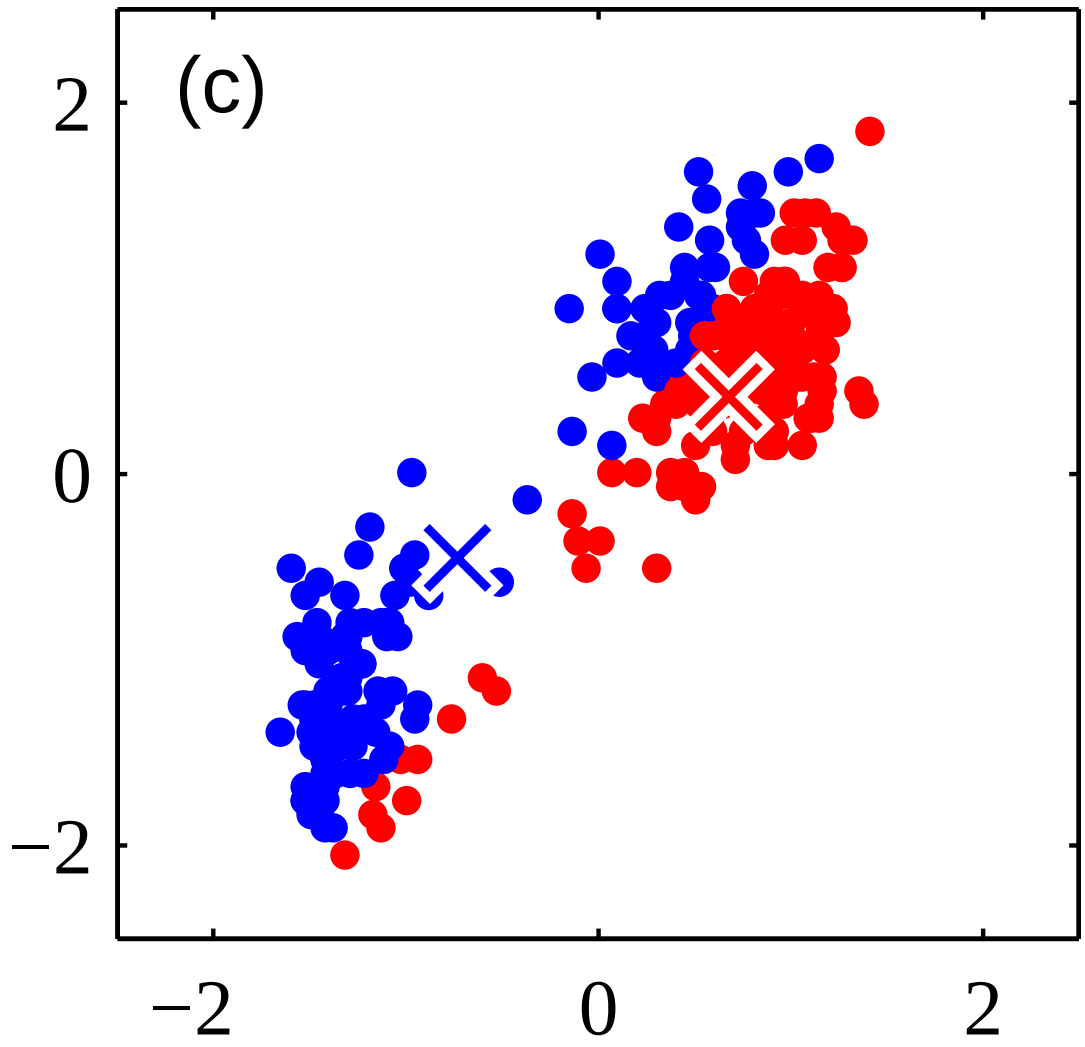
Exemplo



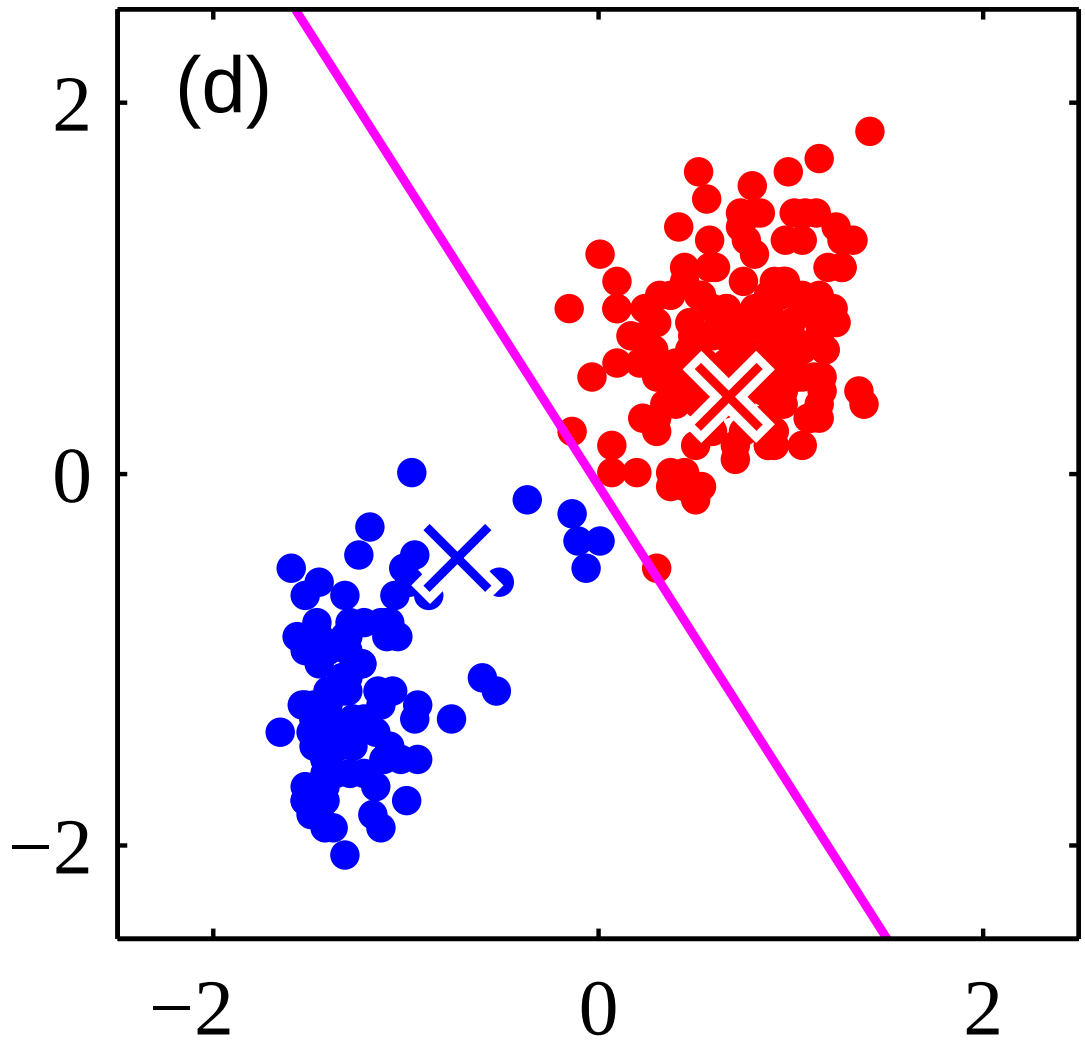
Exemplo



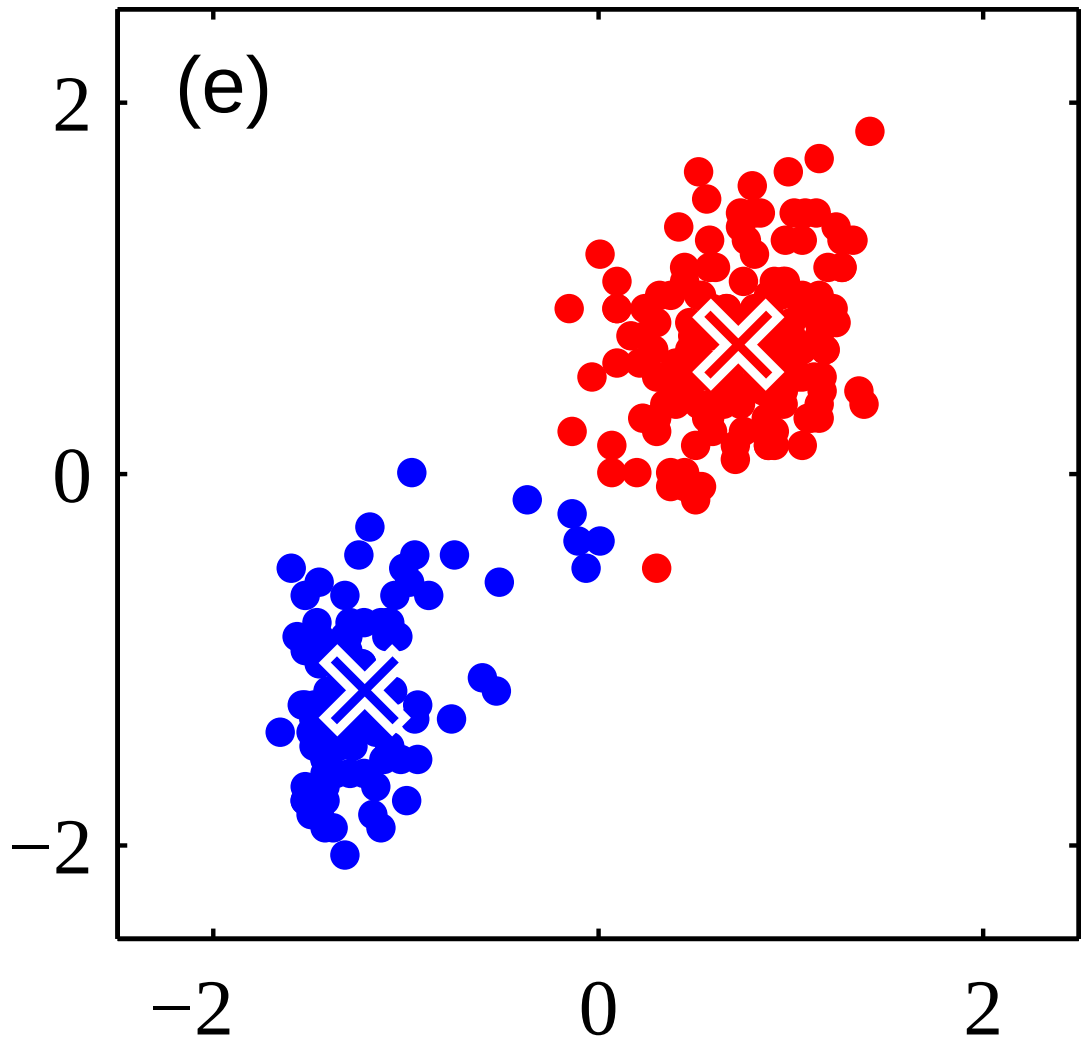
Exemplo



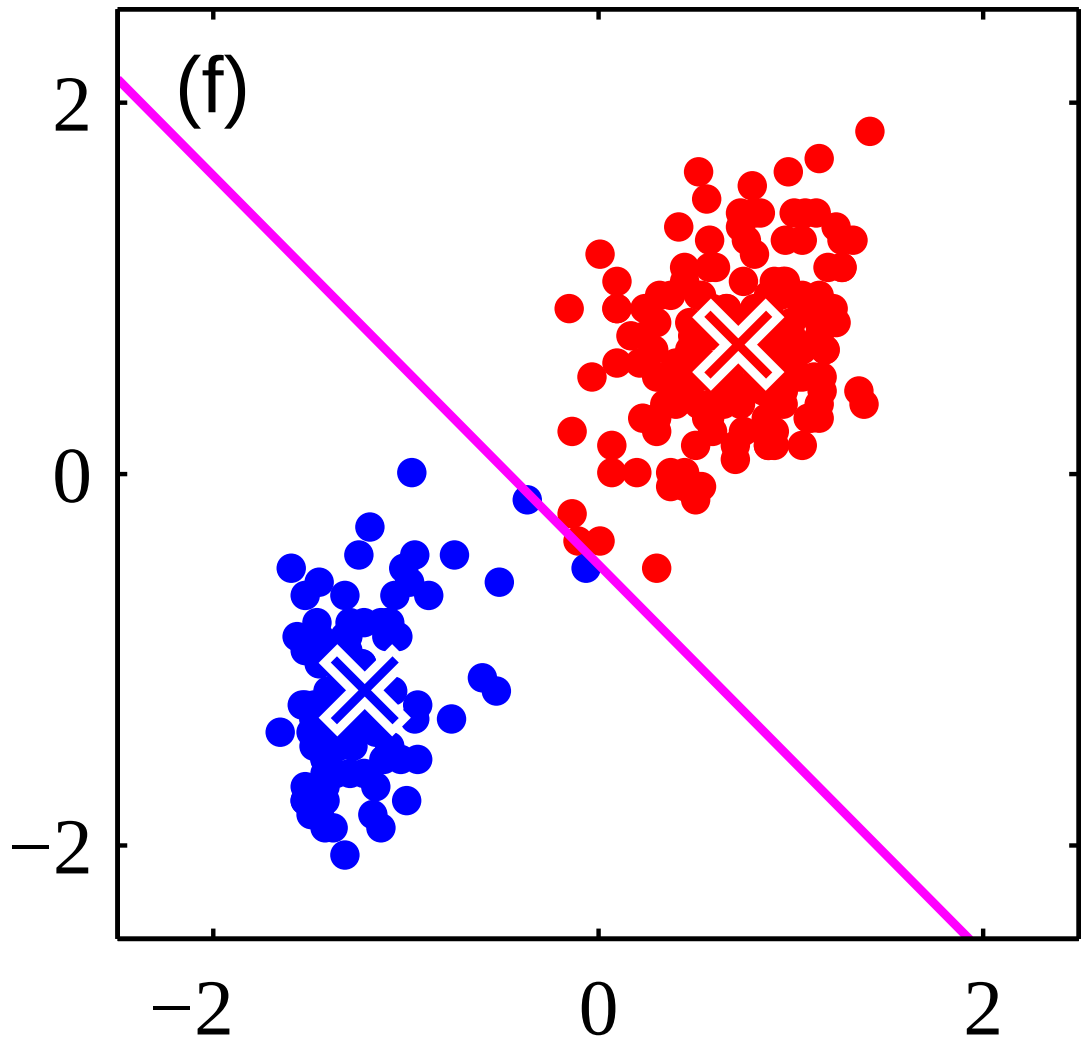
Exemplo



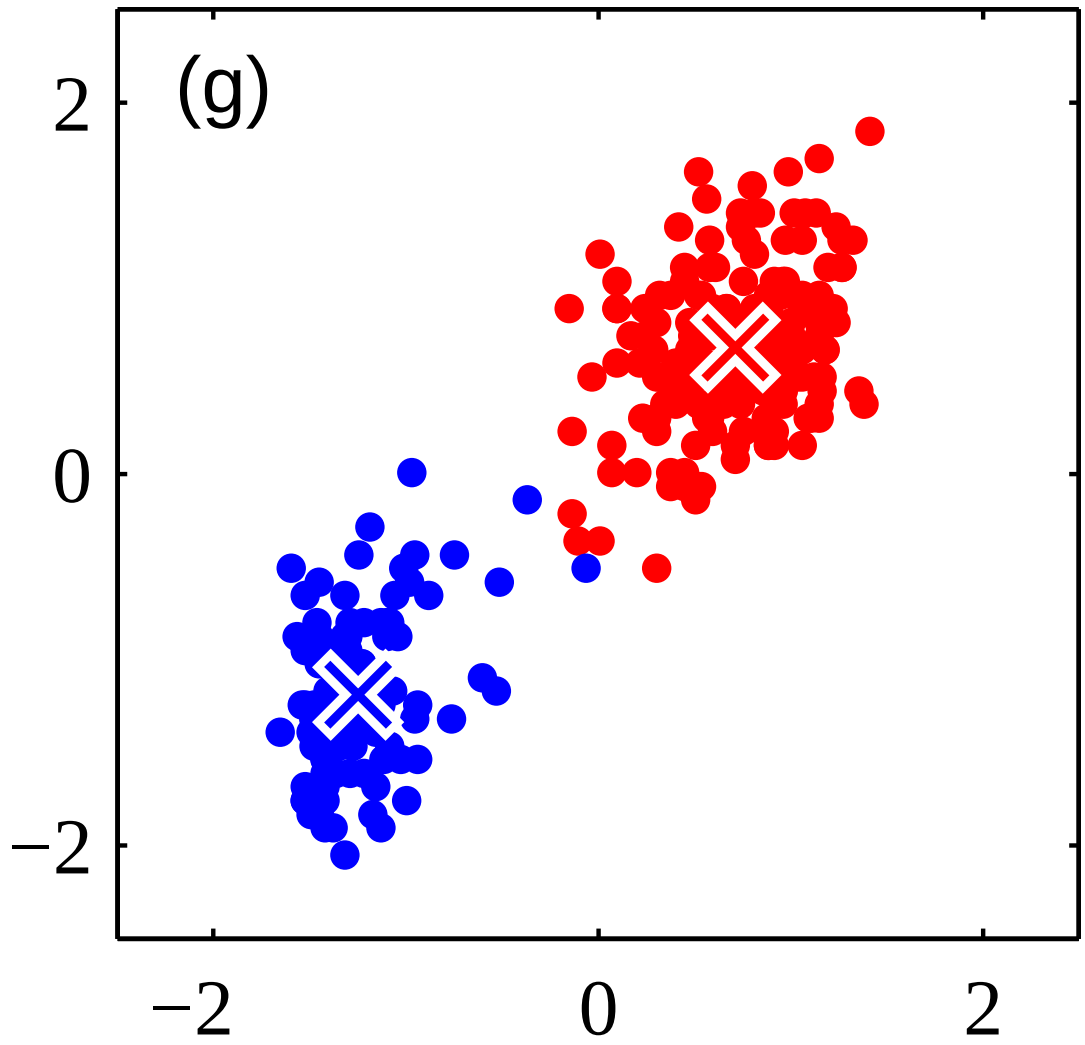
Exemplo



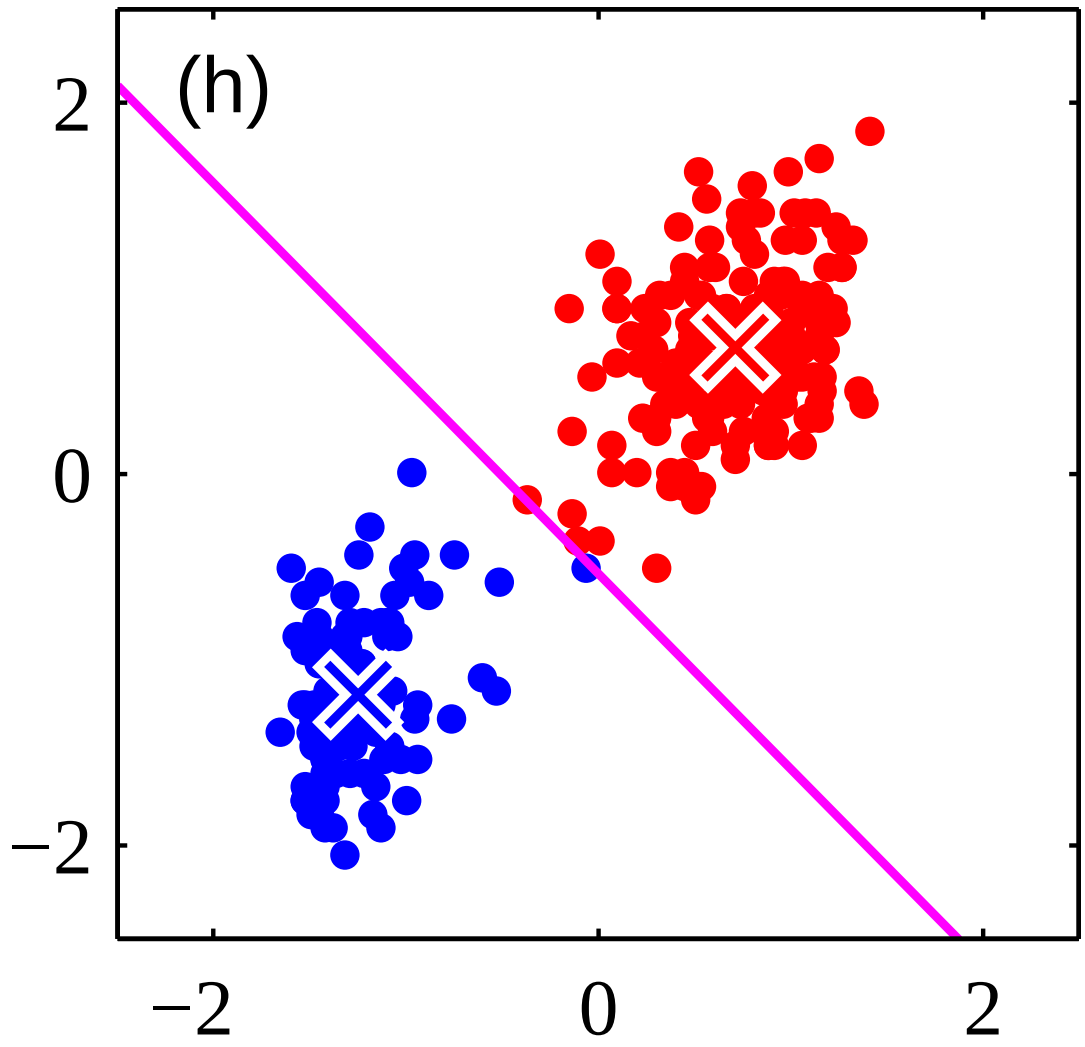
Exemplo



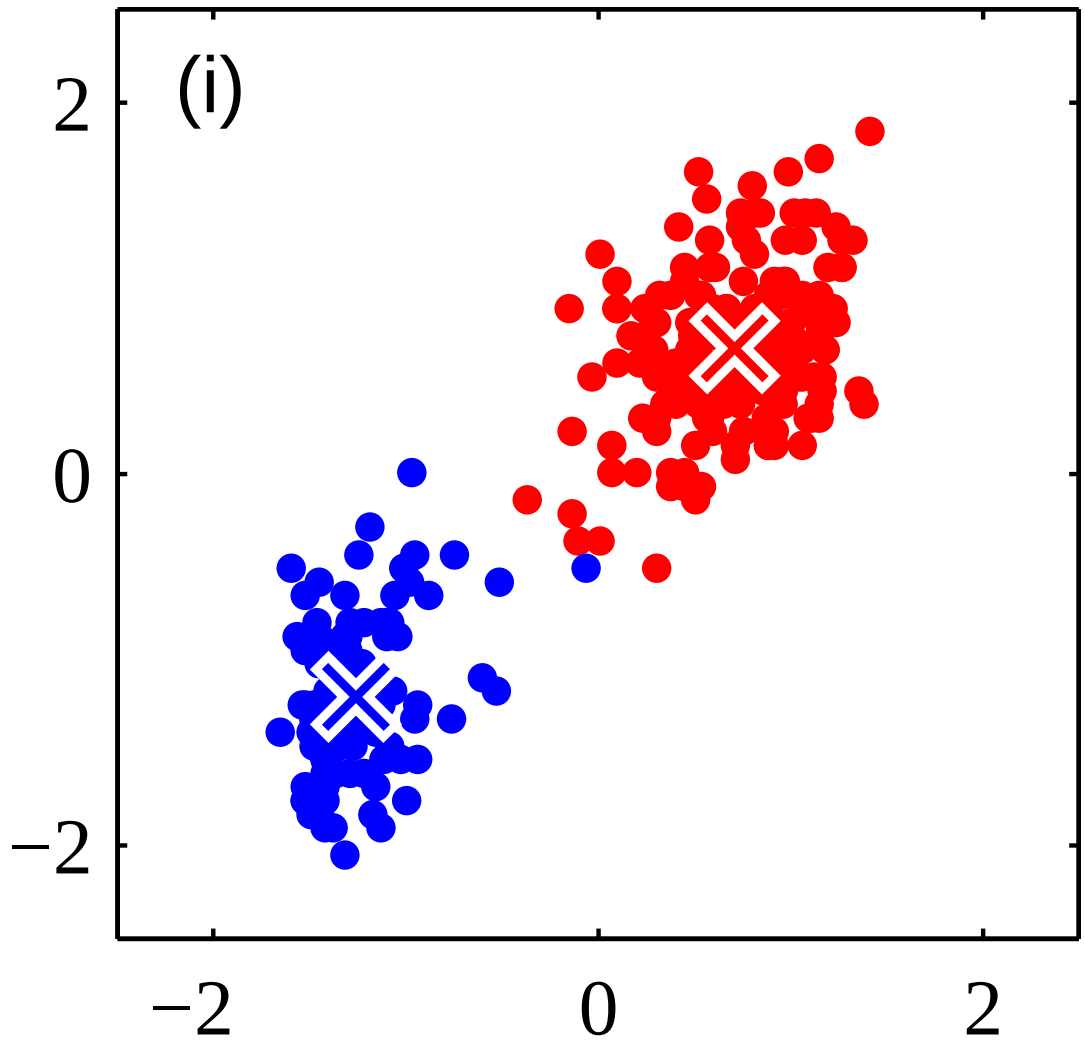
Exemplo



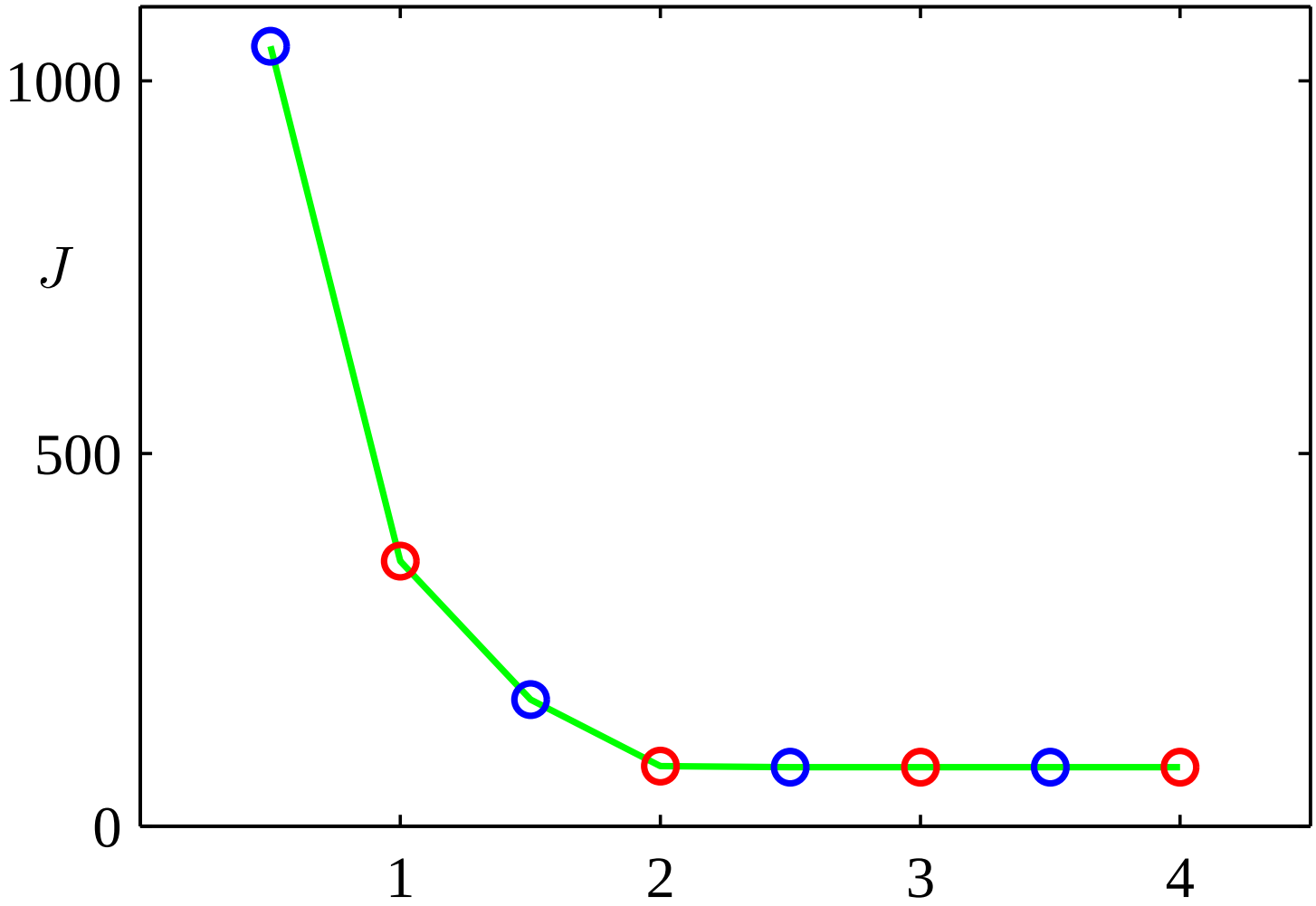
Exemplo



Exemplo



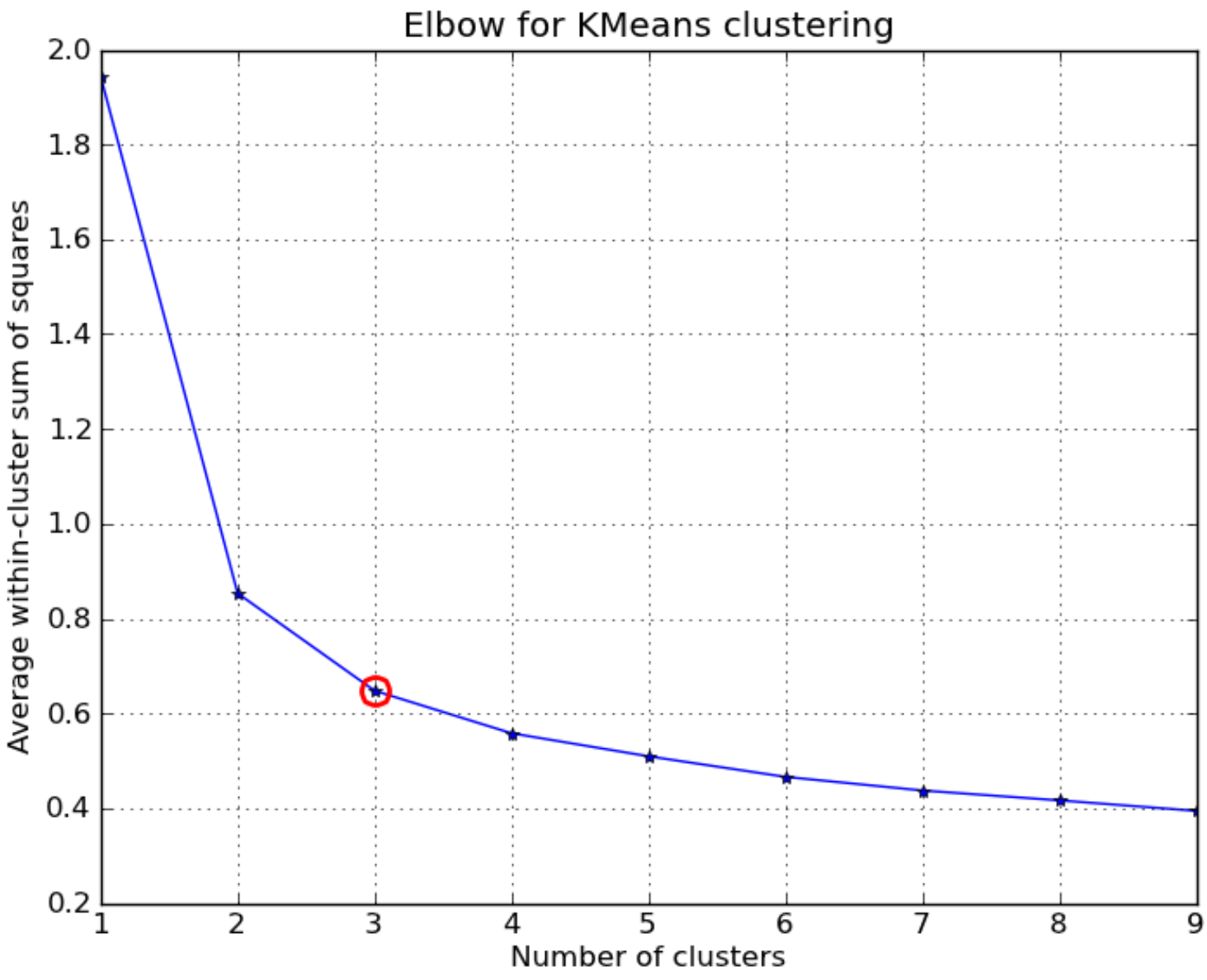
Exemplo: Função Custo



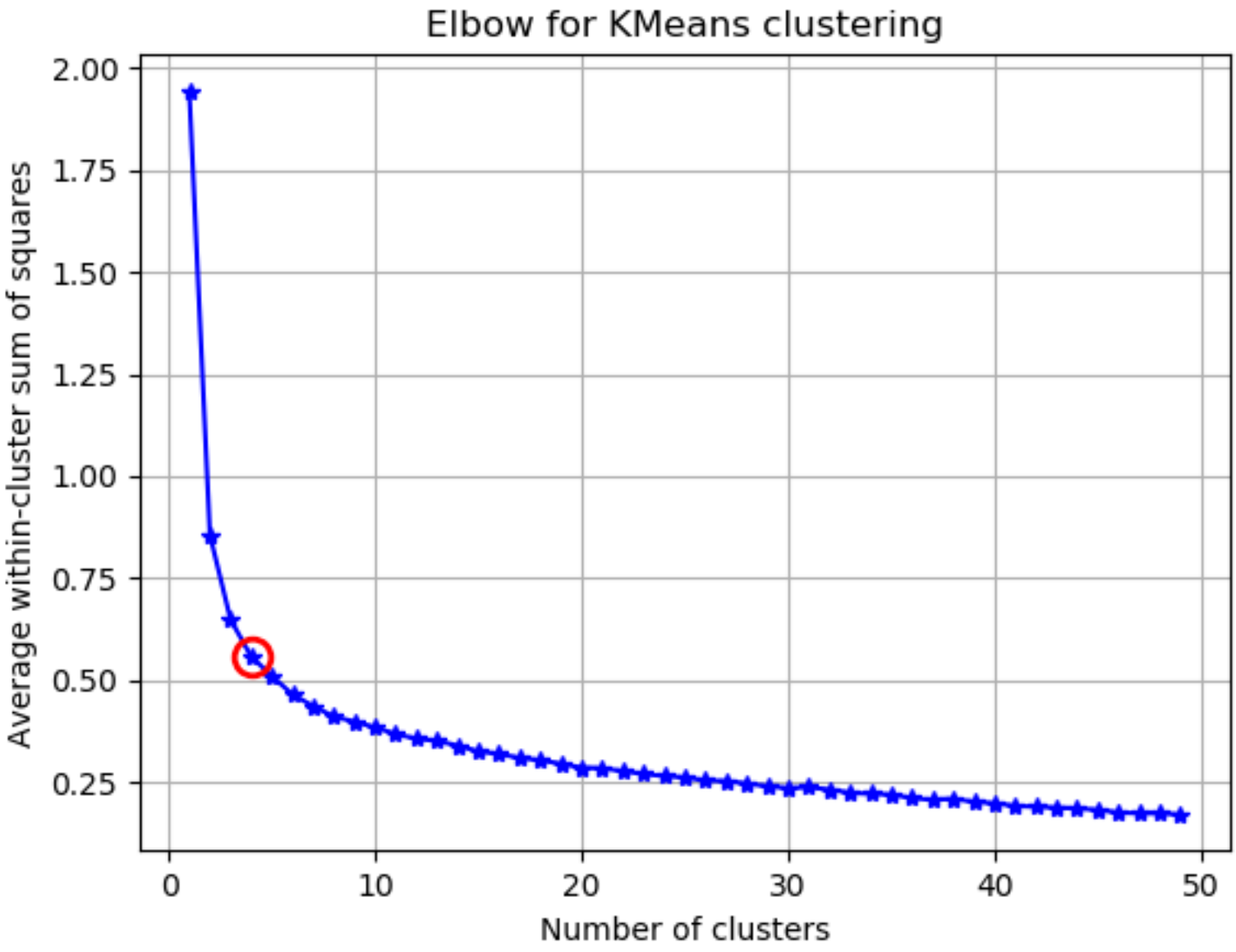
Escolha de K

- Uma limitação do K-means é a necessidade de escolher K a priori
- Diversas métricas propostas na literatura (ex: *Silhouette cofficient*)
- A melhor métrica é o desempenho na tarefa final, se for viável estimar
- Em alguns casos o valor de K pode ser imposto pela aplicação

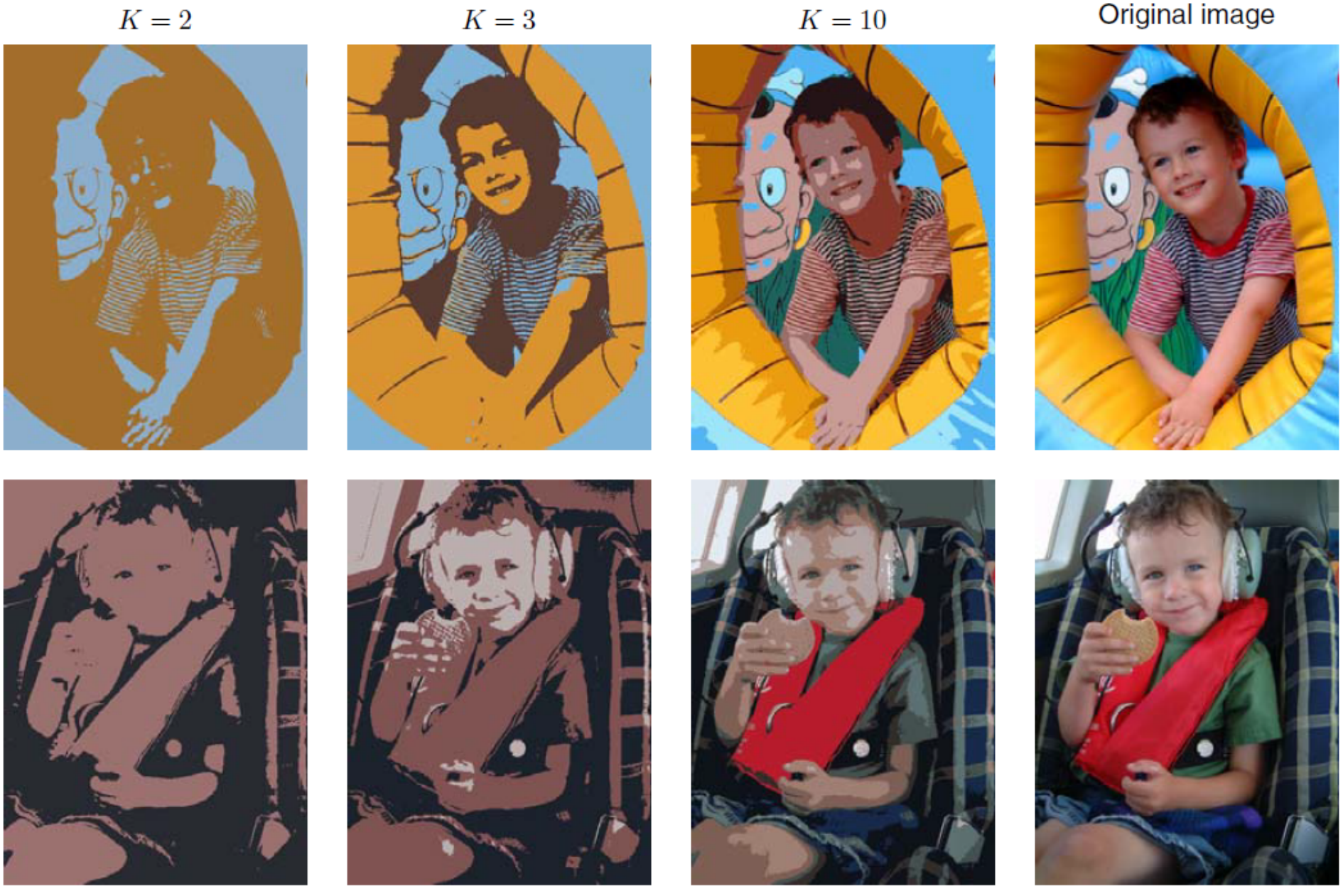
Exemplo: Método do “Joelho” (*Elbow Method*)



Exemplo: Método do “Joelho” (*Elbow Method*)



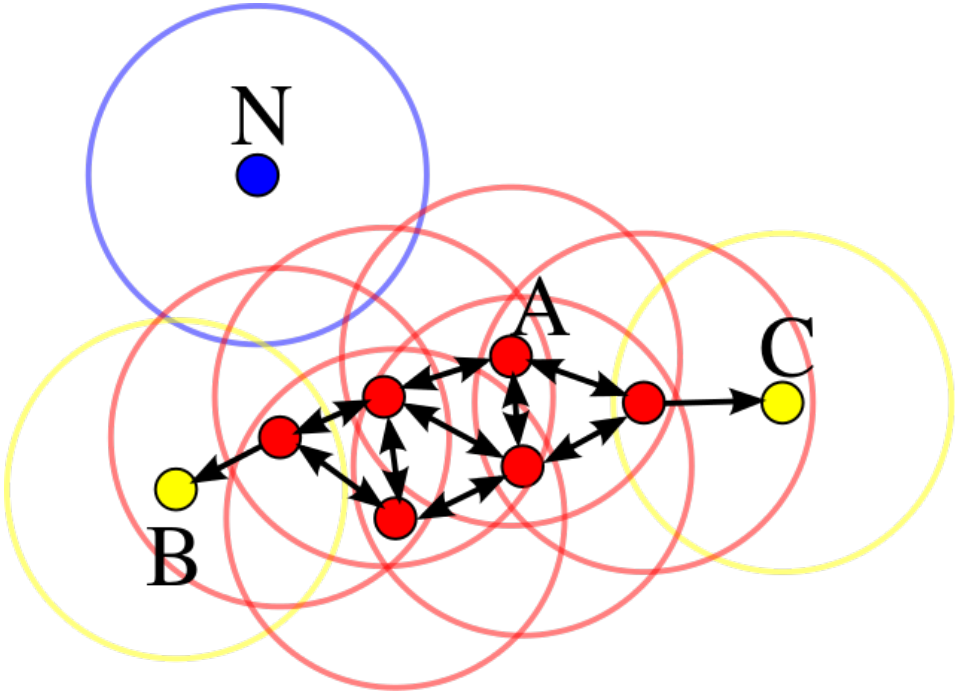
Exemplo: Segmentação/Compressão de Imagens



DBSCAN

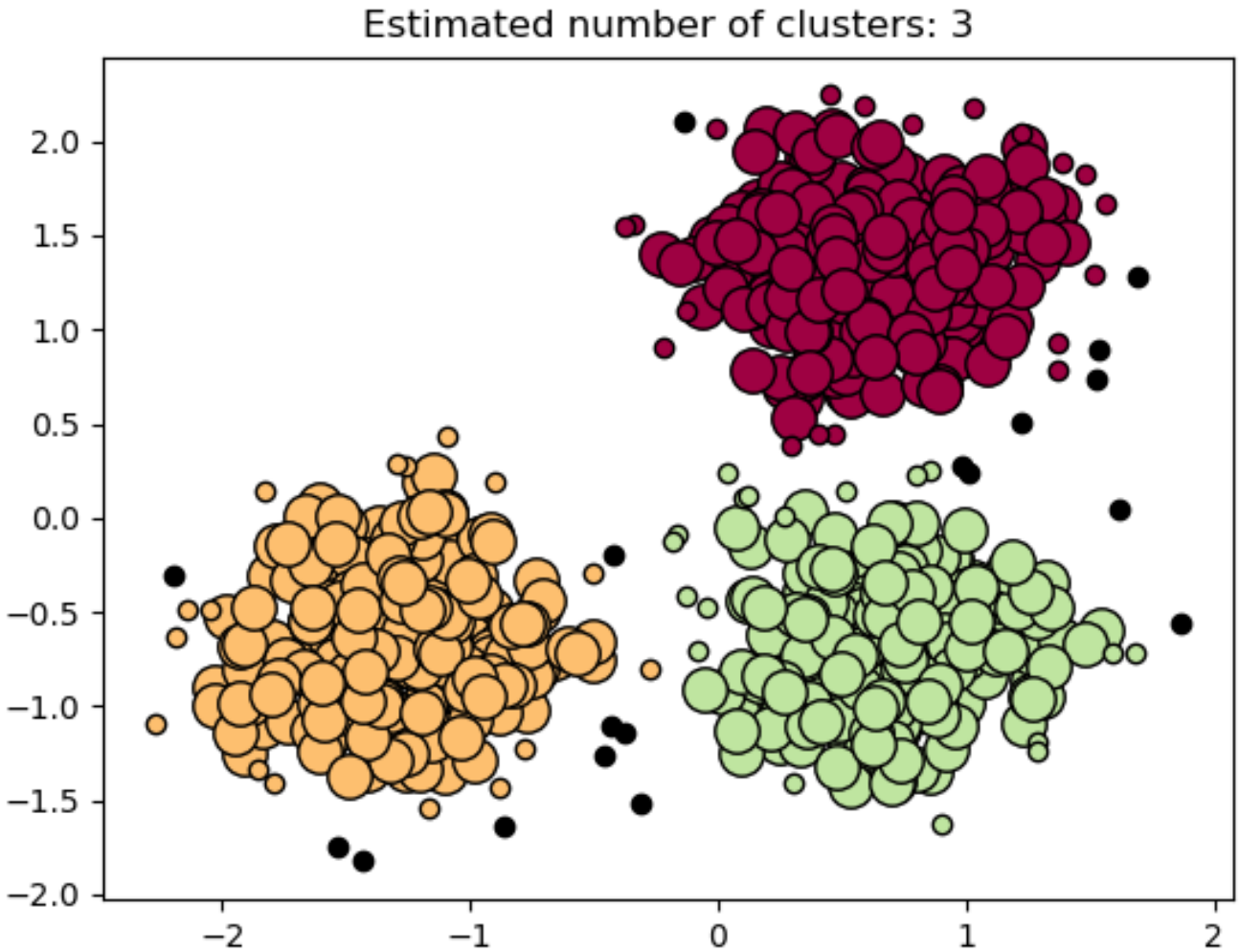
DBSCAN

Density-Based Spatial Clustering of Applications with Noise

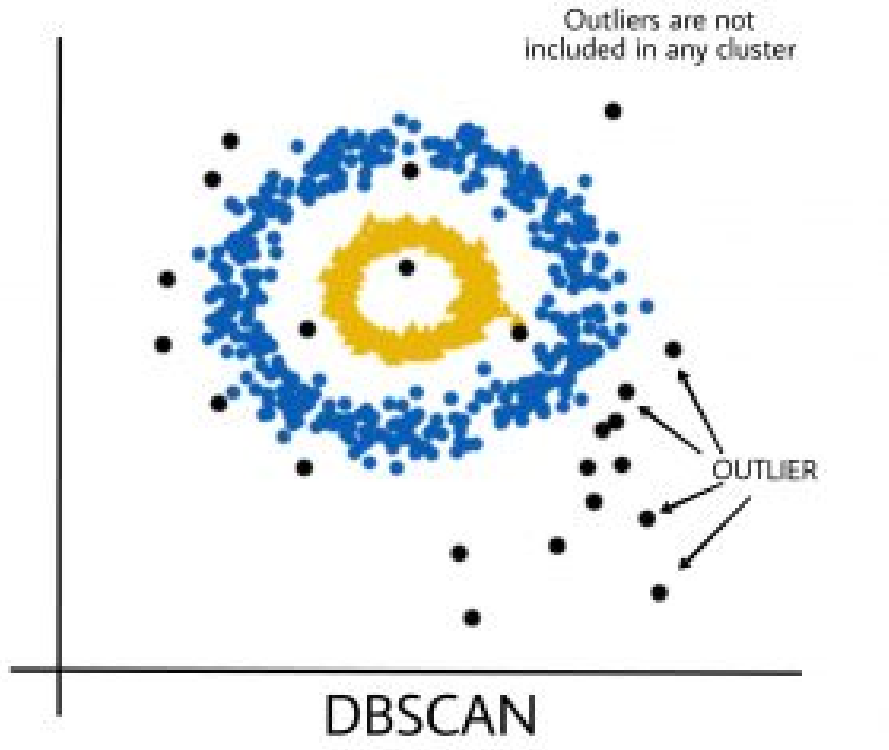


- Vizinhos: pontos com distância ϵ ou menor
- Ponto central: ponto com $\mathrm{minPts} - 1$ ou mais vizinhos
- Ruído/outlier: ponto que não é vizinho de nenhum ponto central
- Todos os vizinhos de um ponto central fazem parte do mesmo cluster

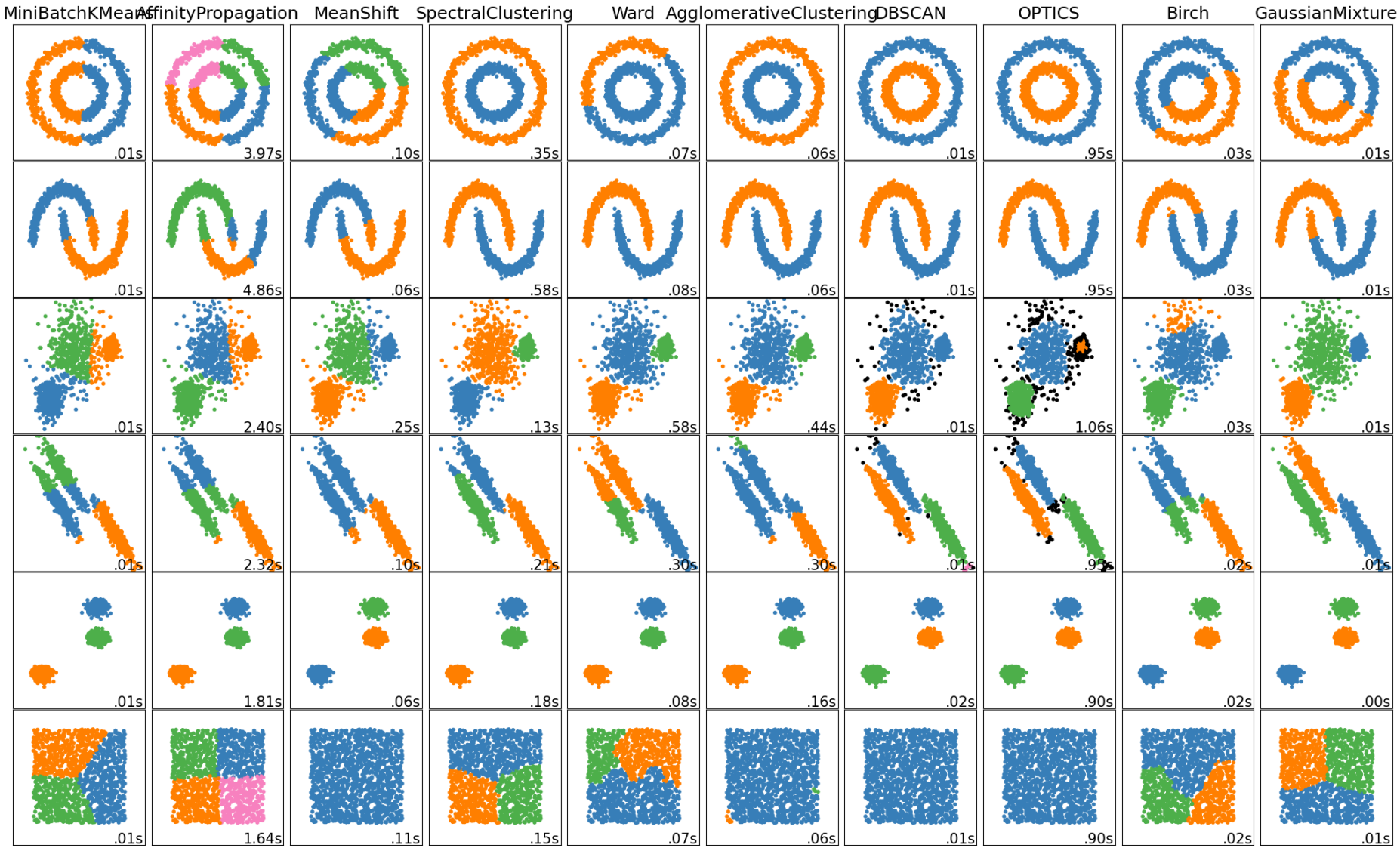
Exemplo



Exemplo



Comparação


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
OPTICS
Birch
GaussianMixture
.01s
3.97s
.10s
.35s
.07s
.06s
.01s
.95s
.03s
.01s
.01s
4.86s
.06s
.58s
.08s
.06s
.01s
.95s
.03s
.01s
.01s
2.40s
.25s
.13s
.58s
.44s
.01s
1.06s
.03s
.01s
.01s
2.32s
.10s
.21s
.30s
.30s
.01s
.92s
.02s
.01s
.01s
1.81s
.06s
.18s
.08s
.16s
.02s
.90s
.02s
.00s
.01s
1.64s
.11s
.15s
.07s
.06s
.01s
.90s
.02s
.01s