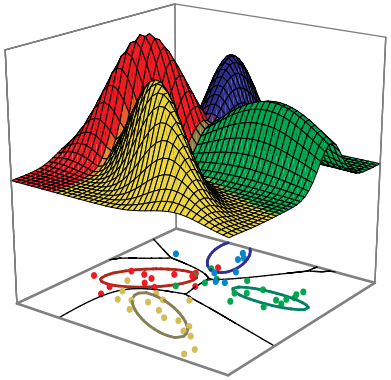
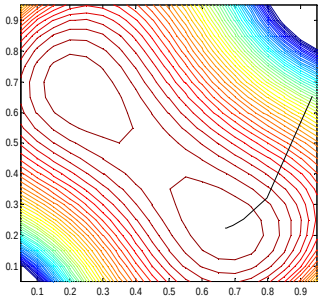
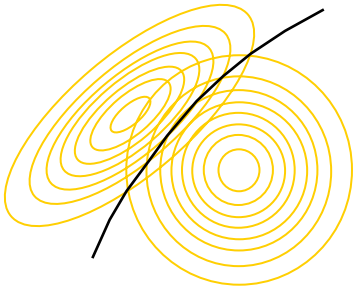
Reconhecimento de Padrões







Escola Superior de Tecnologia – Engenharia Informática
Reconhecimento de Padrões
Prof. João Ascenso

Sumário: Aprendizagem não supervisionada

- Introdução à aprendizagem não supervisionada
- Métodos paramétricos e não paramétricos
- Algoritmo de k-médias
- Algoritmo de LBG
- Algoritmo de k-médias difuso
- Algoritmo ISODATA
- Agrupamento hierárquico (clustering)
- Agrupamento divisivo (clustering)

Aprendizagem não supervisionada

- Os sistemas de reconhecimento de padrões estudados até agora, assumem que:
  - Um padrão é uma par de variáveis {x,w} onde x é um conjunto de características e w é a classe a que pertence.
  - Os sistemas deste tipo são **supervisionados**, uma vez que é dado ao sistema o vector de características e a sua rótulo (resposta correcta)
- Na aprendizagem não supervisionada os algoritmos assumem sempre que não se conhece a que classe pertence a colecção de dados $X=\{x_1,x_2, ..., x_n\}$.
  - A aprendizagem não supervisionada possui capacidades limitadas, no entanto é bastante útil em várias áreas.

Aprendizagem não supervisionada

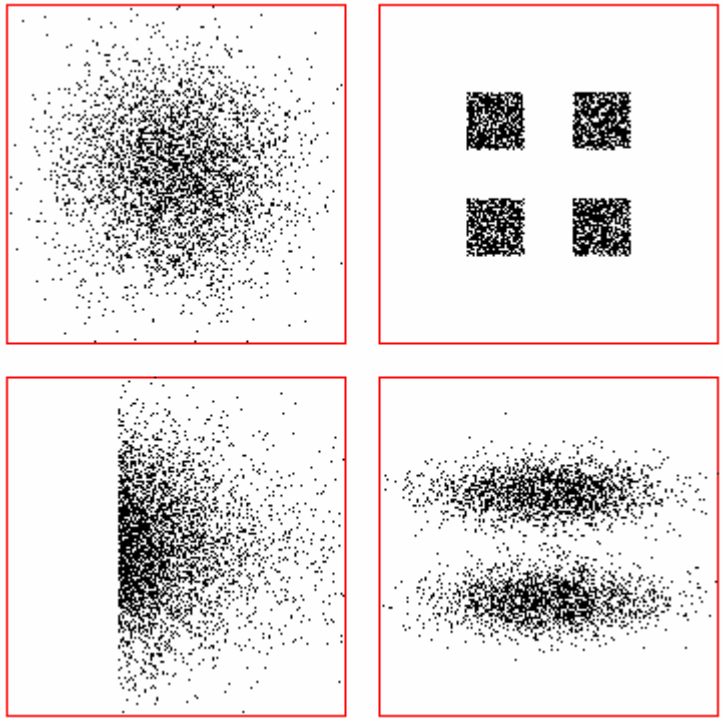
- Aplicações típicas:
  - Classificação extensos conjuntos de dados, p.e. reconhecimento de voz.
  - Não se conhece a que classe pertencem os dados antes da classificação, p.e. data mining.
  - Representação de um conjunto grande de dados por um pequeno conjunto de protótipos, p.e. como pré processamento.
- Existem duas abordagens principais na aprendizagem não supervisionada:
  - Paramétrica: Assume-se a forma da distribuição dos dados, o objectivo principal é estimar os parâmetros.
  - Não paramétrica: Não se assume nenhuma forma de distribuição, o objectivo principal é obter uma partição do espaço de características.

Algoritmos de aprendizagem não supervisionada

- Apenas vamos aprender algoritmos não paramétricos de aprendizagem não supervisionada.
- O principal objectivo destes algoritmos é encontrar agrupamentos “naturais” de dados e é constituída por três passos:
  - Definição de uma medida de semelhança entre amostras
  - Definição de uma função de custo para agrupar os dados
  - Definição de um algoritmo para minimizar (ou maximizar) a função de custo.
- Medida de semelhança: O cálculo da semelhança entre agrupamento de dados pode ser expresso por uma única fórmula que dá a similaridade ou distância entre *clusters*.

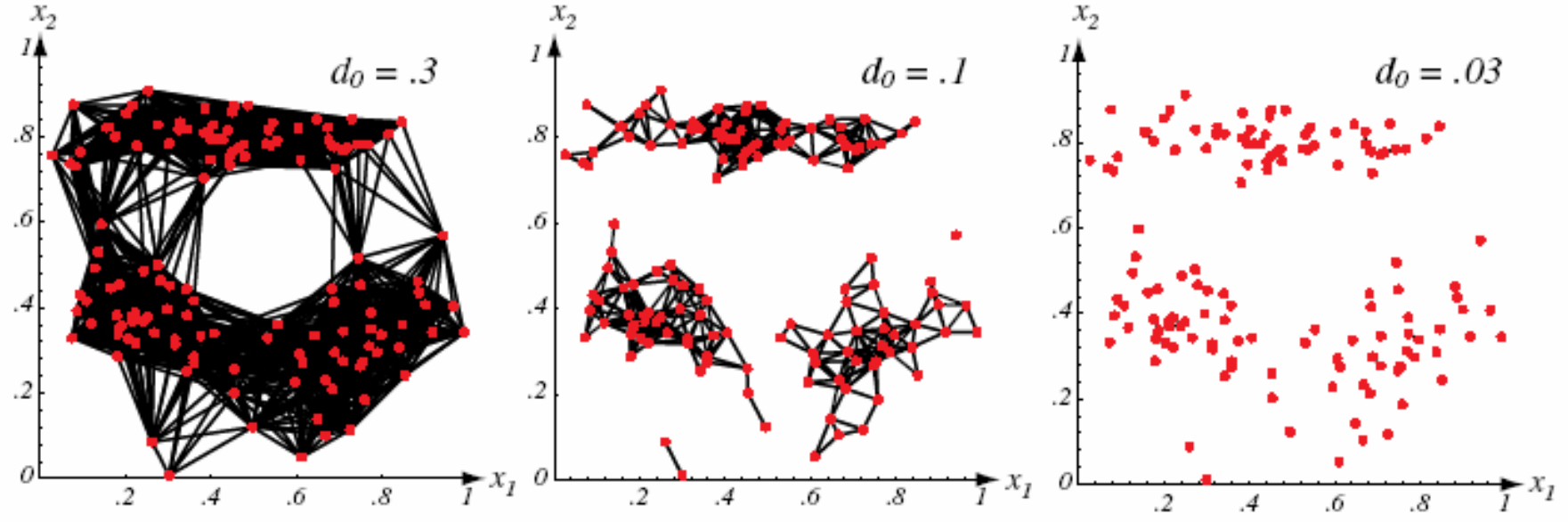
Algoritmos de aprendizagem não supervisionada

- Factores importantes:
  - Estrutura dos padrões no espaço de características.
  - De que forma devemos medir a semelhança entre amostras ?



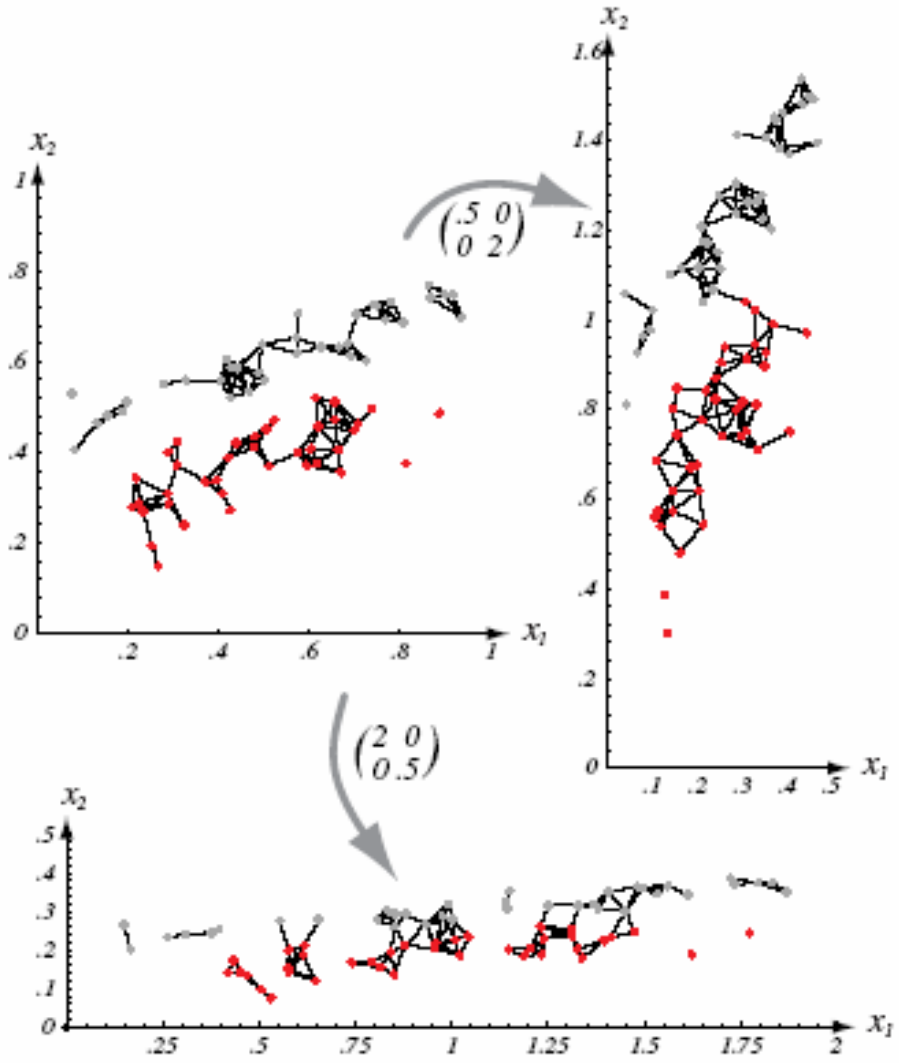
Medida de semelhança

- Se usarmos a distância euclidiana, e estabelecermos um limite d_0 para a distância entre duas amostras obtemos:
  - Diferentes agrupamentos consoante o valor de d_0
  - Dependência da medida de semelhança !



Medida de semelhança

- Sensível a transformações lineares dos dados !



Medida de semelhança

- Medidas de semelhança:
  - Distância euclidiana, city block ou mahalanobis
  - A distância euclidiana e city block são sensíveis a normalizações dos eixos.
  - De uma forma geral podemos definir uma medida de semelhança s(x,y) cujo valor irá ser grande quando x e y são “semelhantes”.
  - Produto interno como função de semelhança:

$$s(x,y) = \frac{x^t y}{\|x\|\|y\|}$$

  - Função de Tanimoto para valores binários:

$$s(x,y) \frac{x^t y}{x^t x + y^t y - x^t y}$$

Funções de custo

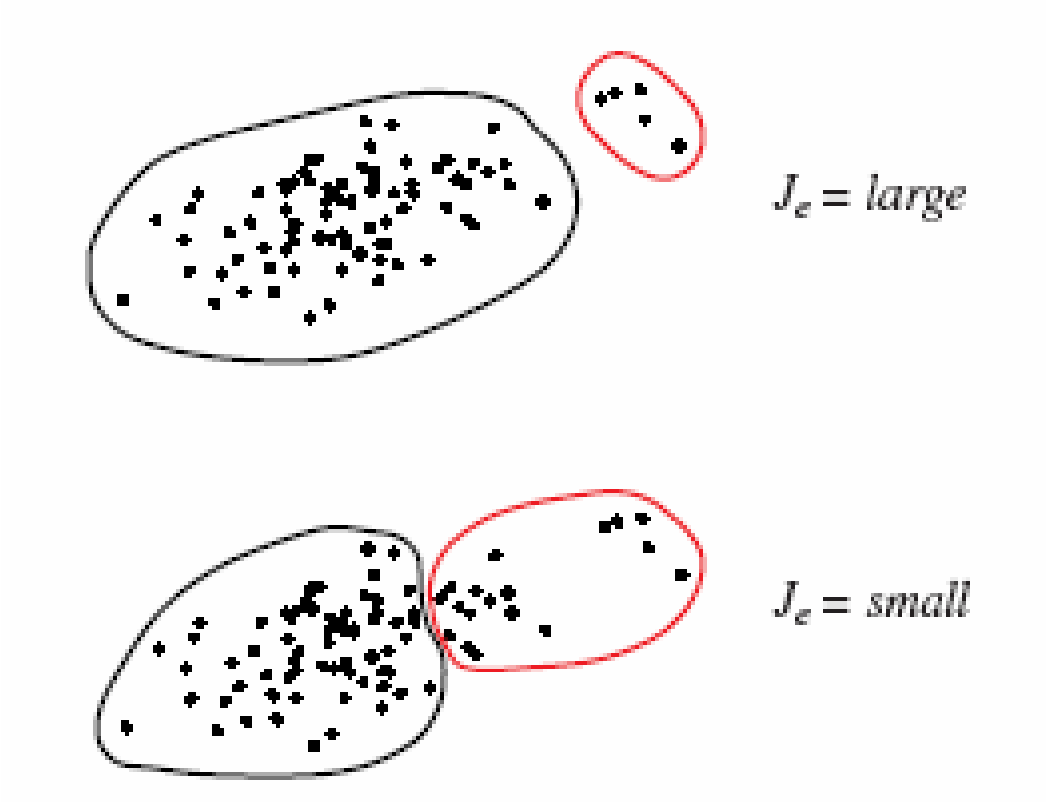
- Uma vez que a medida de semelhança foi determinada, é necessário definir uma função de custo para ser optimizada.
- Mede a "qualidade" de qualquer partição dos dados.
- A função de custo mais utilizada é o erro quadrático médio.
  - Seja ni o número de amostras de Di e mi a média dessas amostras, então o erro quadrático médio defini-se como:

$$J_e = \sum_{i=1}^{c} \sum_{x \in Di} \left\| x - m_i \right\|^2$$

  - Interpretação: Para um dado cluster Di, o vector mi é a melhor representante das amostras em Di de forma a minimizar a soma dos comprimentos quadrados de cada x em relação a mi.

Funções de custo

- Je mede o erro quadrático total em representar as amostras $x_1, x_2, \ldots, x_n$ por c agrupamentos com o centro em $m_1, m_2, \ldots, m_n$



Funções de custo

- Mais funções de custo:
  - Média
  - Mediana
  - Distância máxima
- A expressão generalizada fica:

$$s_i = \frac{1}{n^2} \sum_{x \in Di} \sum_{x' \in Di} s(x, x')$$

- ou

$$s_i = \min_{x, x' \in Di} s(x, s')$$

- Outros critérios existem baseados em matrizes esparsas:
  - O critério do traço da matriz: soma de todos os elementos da diagonal.
  - O critério do determinante da matriz.
  - O critério invariante, baseado nos valores próprios.

Algoritmos iterativos

- Uma vez que a função de custo já foi definida devemos encontrar uma partição dos dados que minimize o critério.
  - O método exaustivo, consiste em definir todas as partições possíveis, e garante uma solução óptima
  - No entanto não é possível ser realizado, por exemplo para 5 clusters (agrupamentos de dados) e 100 amostras existem 10^{67} partições possíveis.
- A abordagem mais comum é definir um processo iterativo
  - Encontrar uma partição inicial dos dados razoável.
  - Mover as amostras de uma partição para outra de forma a minimizar a função de custo.
  - Solução sub óptima mas computacionalmente fazível

Métodos iterativos

- Considerar dois grupos de métodos iterativos
  - Algoritmos de agrupamento planos:
  - Produzem um conjunto de agrupamentos disjuntos
  - Algoritmos mais conhecidos:
    - Algoritmo k-means
    - Algoritmo LBG
    - Algoritmo k-means difuso
    - ISODATA
  - Algoritmos de agrupamento hierárquicos:
  - O resultado é uma hierarquia de agrupamentos
  - Podem ser divididos em duas classes: aglomerativos e hierárquicos.

Algoritmo de k-médias

- O algoritmo de k-médias é um algoritmo iterativo que tenta minimizar o erro quadrático médio (função de custo Je)
- O algoritmo de k-médias é bastante usado nas áreas de processamento de sinal, telecomunicações e data mining.
  - Relação íntima com a quantificação vectorial:
  - Valores unidimensionais são habitualmente quantificados num número finito de níveis (tipicamente uma potência de 2 para serem transmitidos ou guardados em binário)
  - O conjunto dos agrupamento é referido como um “codebook”, e o algoritmo de k-médias permite encontrar o “codebook” que minimiza o erro quadrático médio
  - Pode ser estendido a múltiplos canais.

Algoritmo de k-médias

1. Defini-se o número de clusters (agrupamentos)
2. Incializa-se os clusters através de:
   - Atribuição arbitrária das amostras aos clusters ou
   - um conjunto de centróides em posições arbitrárias coincidentes com amostras dos dados.
3. Iteração nas amostras; para cada amostra:
   - Procura-se o centróide mais próximo.
   - Atribui-se a amostra ao cluster correspondente.
   - Recalcula-se o centróide para esse cluster.
4. Volta-se ao passo 3, até um critério de convergência ser cumprido.

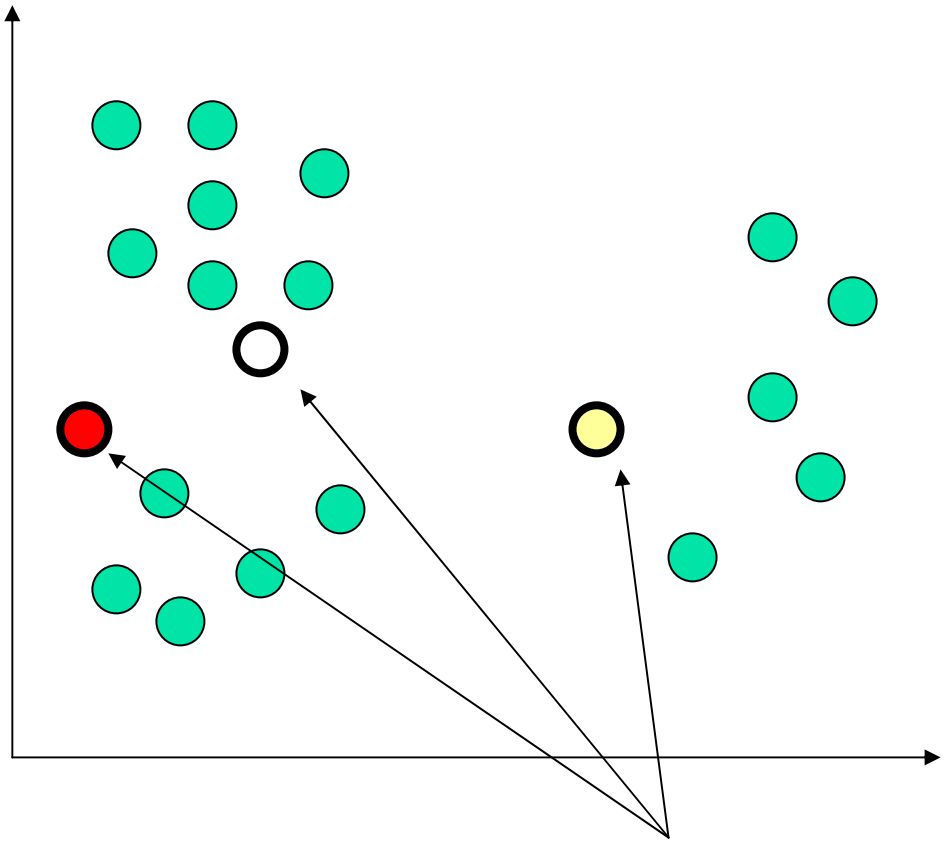
Algoritmo de k-médias

- O centróide corresponde á média de todas as amostras que pertence a esse centróide: $\hat{x}_i = \frac{1}{Ni} \sum_{x \in X^i} x$
- Classificação:

$$x \in X^k : k = \arg\min_i \left\| x - \hat{x}_i \right\|$$

- Critérios de convergência:
  - Até que nenhuma amostra mude de agrupamento.
  - Até que os centróides não sejam alterados.
  - Até que o valor da função de custo se mantenha constante, ou menor que um limite.
  - Até alcançar um limite para o número de iterações.

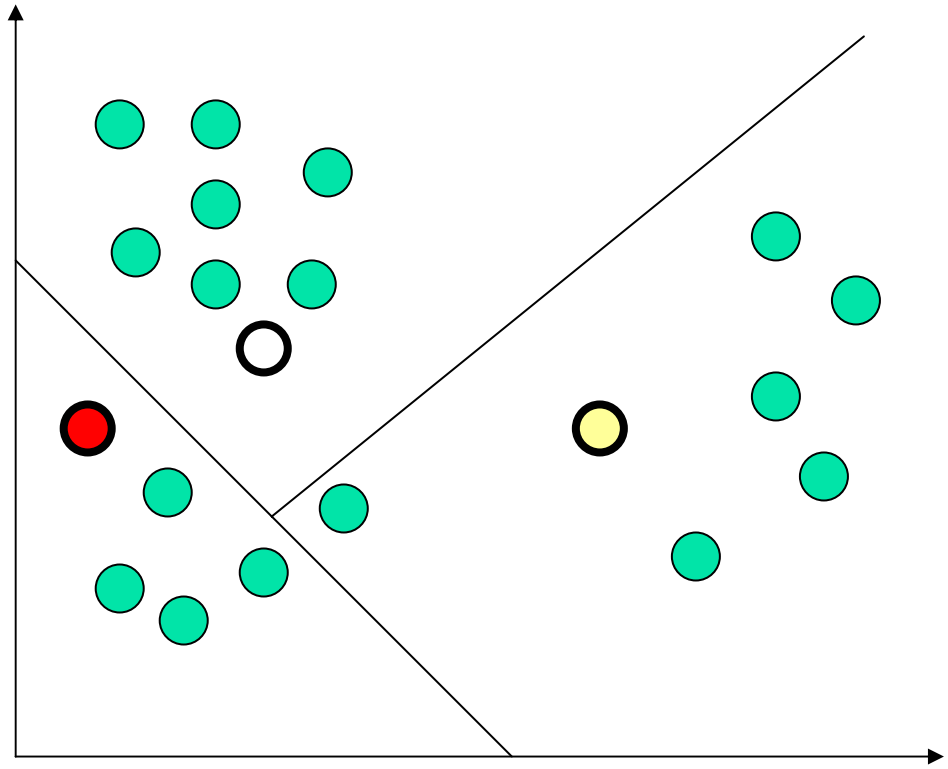
Algoritmo de k-médias



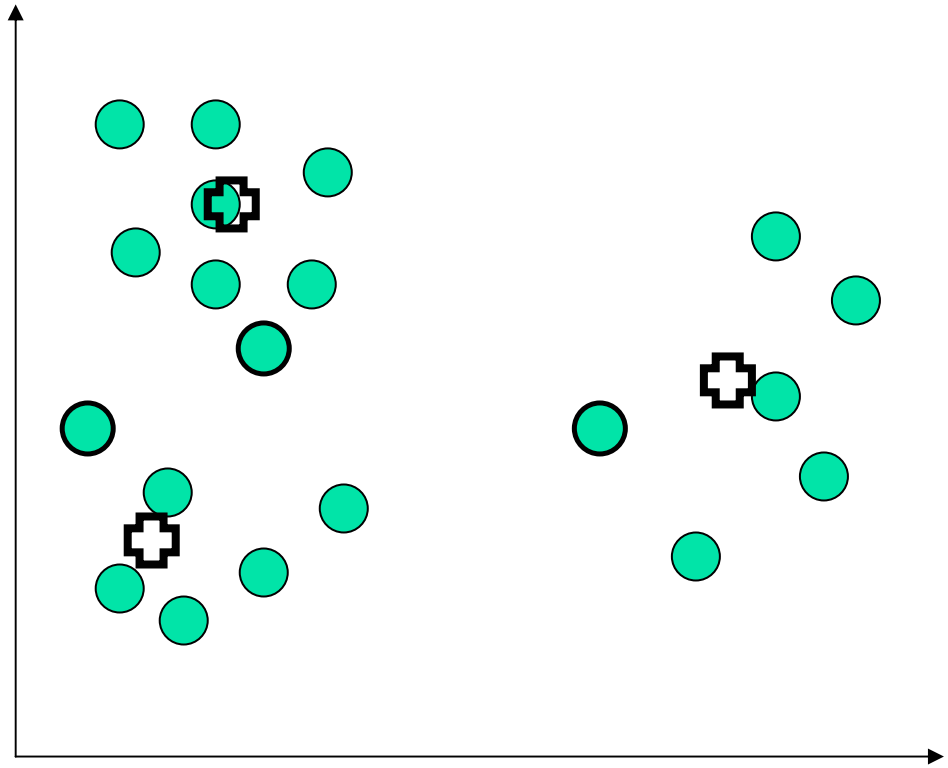
Número
pré-determinado
de clusters

Inicialização dos centróides

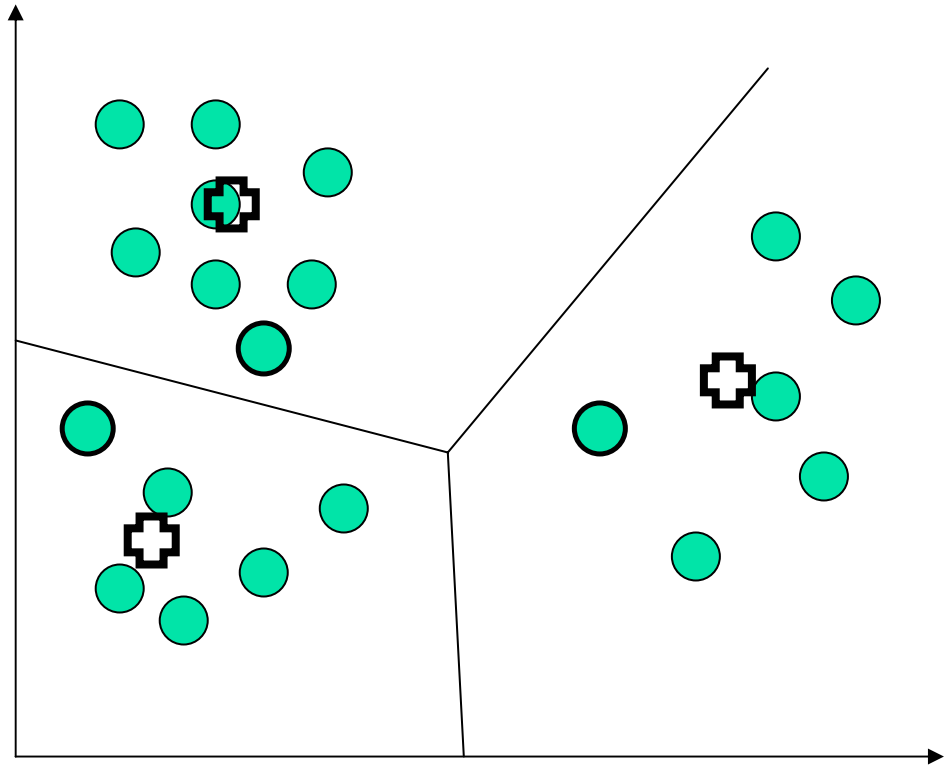
Atribuição das amostras



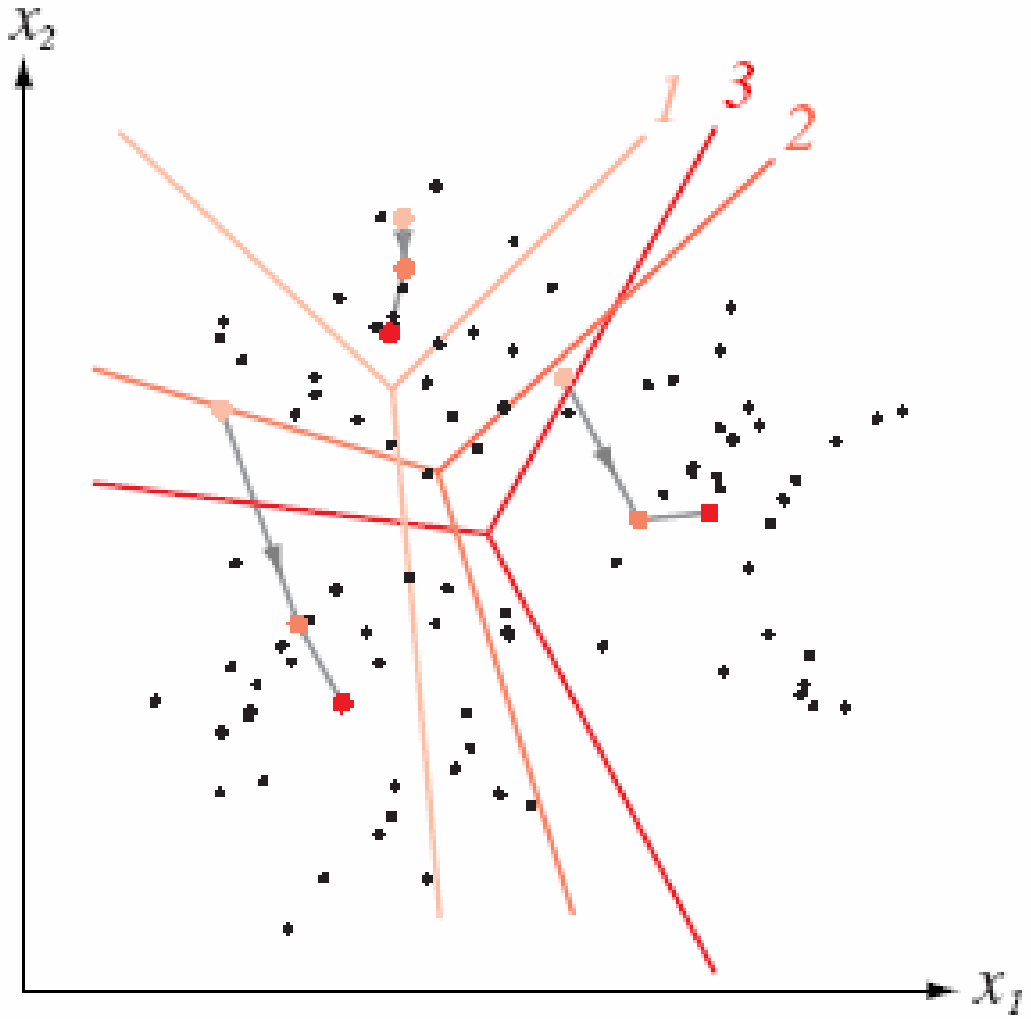
Procurar novos centróides



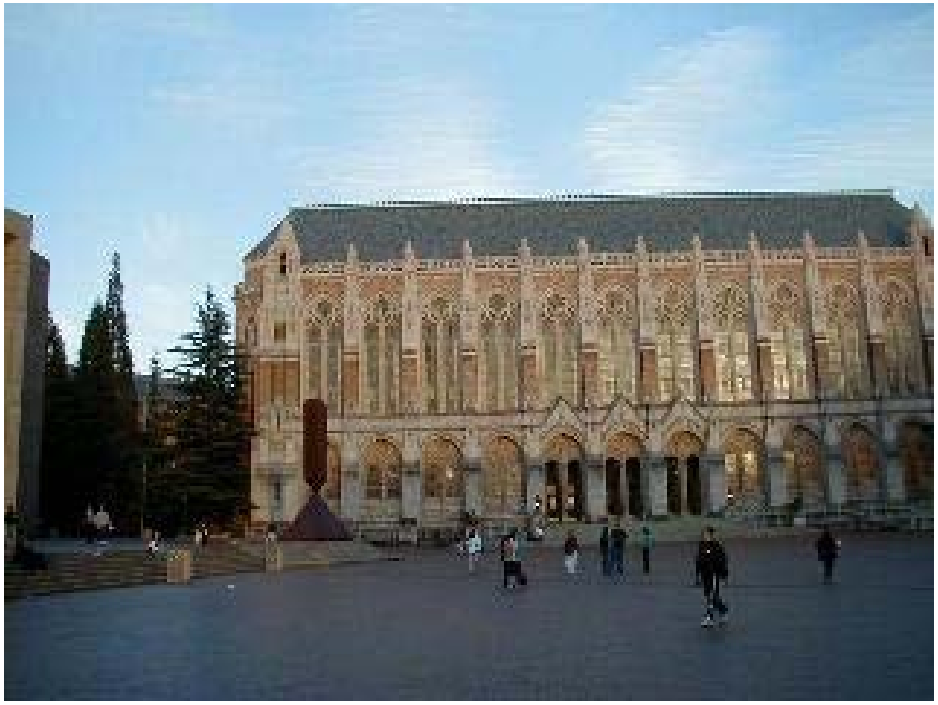
Novos clusters

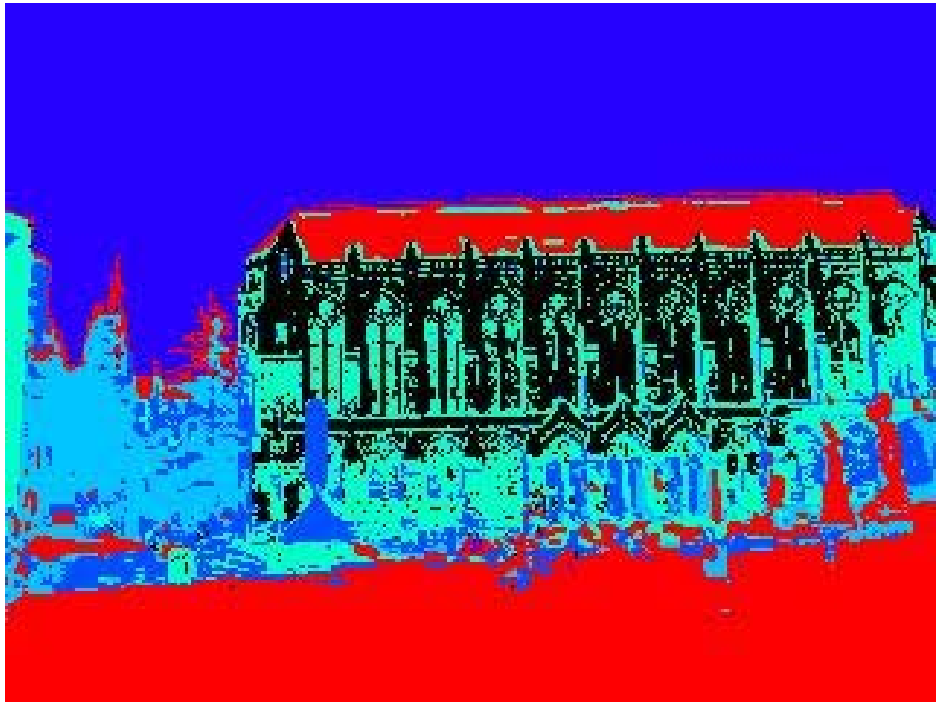


Algoritmo de k-médias (três iterações)



Algoritmo de k-médias (segmentação de imagem)





Algoritmo LBJ (Linde, Buzo e Gray)

Semelhante ao algoritmo de k-médias:

1. Selecção inicial do n.º de clusters e de cada centróide.
2. Iteração nas amostras; para cada amostra:
   - Procura-se o centróide mais próximo.
   - Atribui-se a amostra ao cluster correspondente.
3. Depois da iteração de todas as amostras, recalcula-se os centróides de acordo com a nova atribuição de pontos.
4. Volta-se ao passo 3, até um critério de convergência ser cumprido.

Algoritmo de k-médias difuso

- O algoritmo de k-médias difuso baseia-se na noção de partição difusa permitindo que um padrão pertença a várias classes.
- O objectivo do algoritmo de k-médias difuso é procurar um mínimo para a seguinte função de custo:

$$J_{fuz} = \sum_{i=1}^{c} \sum_{j=1}^{n} \left[\widehat{P}(w_i \mid x_j, \hat{\theta}) \right]^b \left\| x_j - u_i \right\|^2$$

- Para b=0, a função de custo é igual ao erro quadrático médio.
- Para b>1, permite-se que cada padrão pertença a várias classes.

Algoritmo de k-médias difuso

1. Inicialização: n, c, b, μ_1, μ_2,..,μ_c, $P(w_i \mid x_j)$
2. Normalização de $P(w_i \mid x_j)$:

$$\sum_{i=1}^{c} P(w_i \mid x_j) = 1$$

3. Calcula-se u_i através de:

$$u_j = \frac{\sum_{j=1}^{n} \left[P(w_i \mid x_j) \right]^b x_j}{\sum_{j=1}^{n} \left[P(w_i \mid x_j) \right]^b}$$

Algoritmo de k-médias difuso

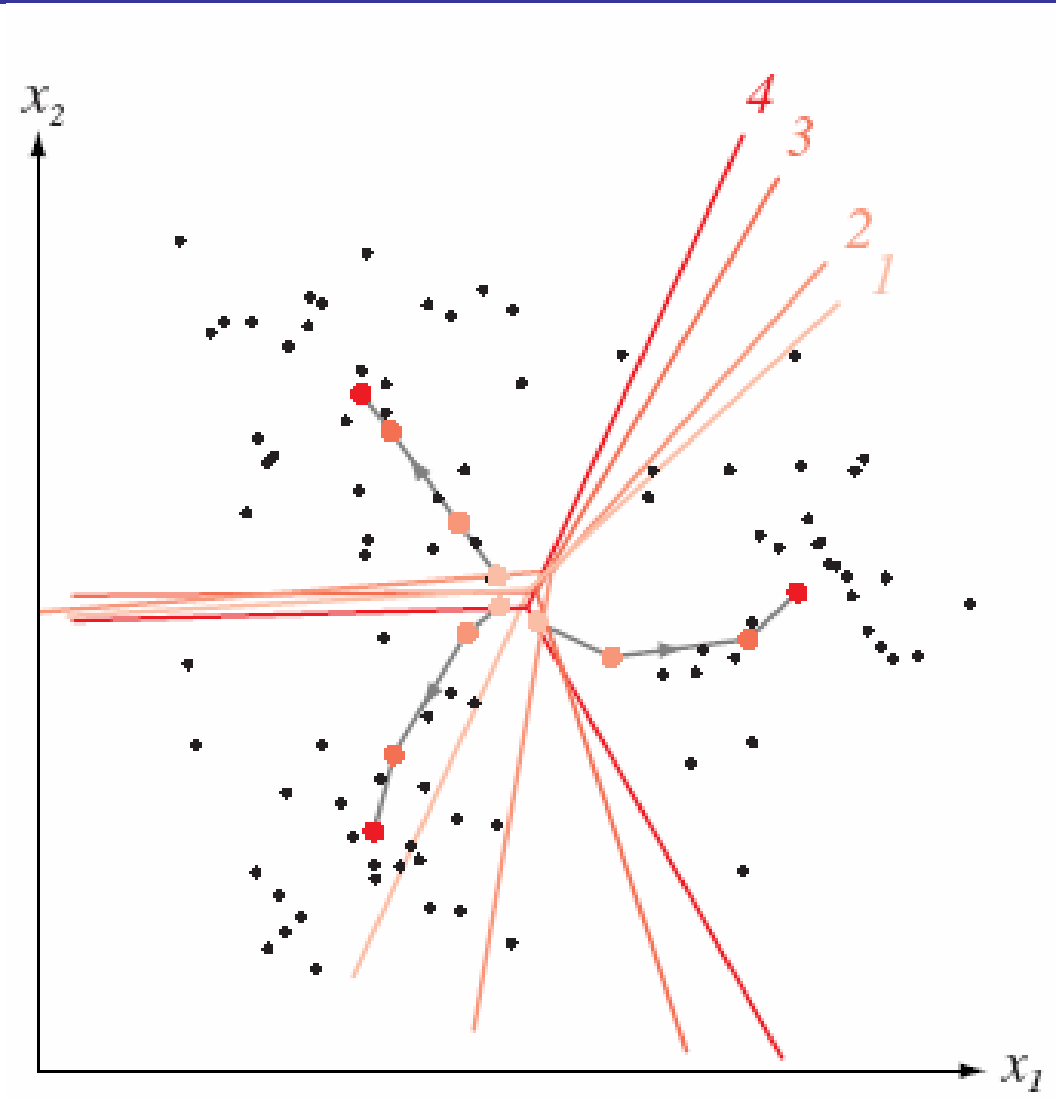
4. Recalcula-se $P(w_i \mid x_j)$:

$$P(w_i \mid x_j) = \frac{(1/d_{ij})^{1/(b-1)}}{\sum_{r=1}^{c} (1/d_{ij})^{1/(b-1)}} \quad \text{com} \quad d_{ij} = \left\| x_j - u_i \right\|^2$$

5. Volta-se ao passo 3, até um critério de convergência ser cumprido, p.e. pequenas alterações em u_i e $P(w_i \mid x_j)$.

- **Interpretação:** A função de custo é minimizada quando os centróides u_j estão perto dos pontos com probabilidade estimada de pertencer ao cluster j.

Algoritmo de k-médias difuso



Discussão: k-médias

- Aplicável a grandes conjuntos de dados
- Sensível às condições de inicialização:
  - Pode-se utilizar outras heurísticas para procurar os centróides iniciais de uma forma adequada.
- Converge para um mínimo local.
- A especificação do número de centróides é muito subjectiva.

Algoritmo ISODATA

- ISODATA, representa *Iterative Self-Organizing Data Analysis Technique Algorithm.*
  - Extende o algoritmo de k-médias para seleccionar automaticamente o número de clusters.
  - Permite eliminar clusters com poucas amostras.
  - Permite dividir clusters com amostras pouco semelhantes entre si.
  - Permite juntar clusters muito perto uns dos outros.
- Parâmetros típicos:
  - Nmin_ex : número mínimo de exemplos por cluster
  - Nd : número de clusters desejados (valor aproximado)
  - σ_s^2 : valor máximo de variância para a divisão de um cluster em 2.

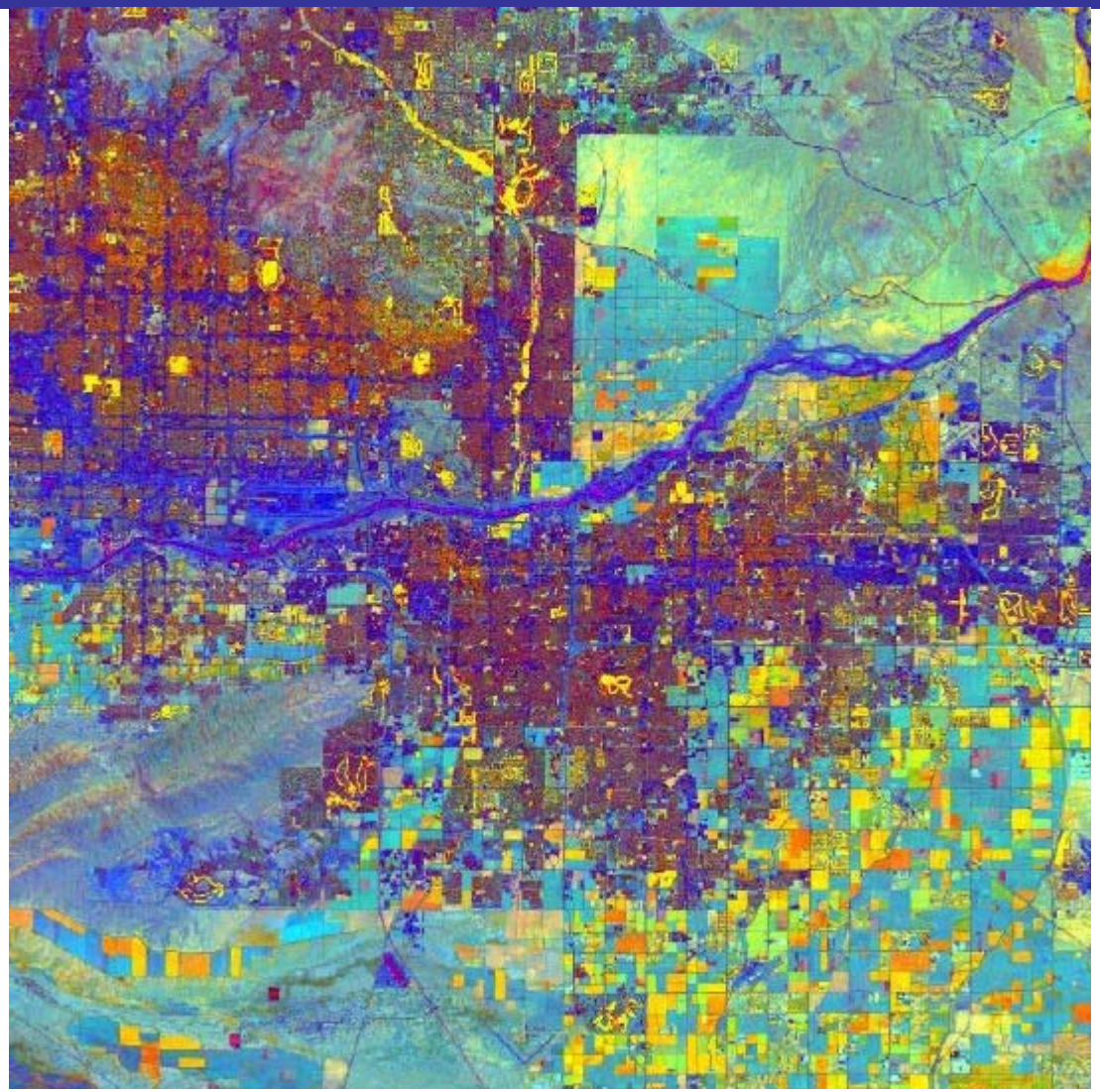
Algoritmo ISODATA

- Mais parâmetros:
  - Dmerge: distância máxima da separação de clusters para agrupar dois clusters.
  - Nmerge: número máximo de clusters que podem ser agrupados.
- O algoritmo funciona de uma forma iterativa:
  1. Efectuar o clustering com o algortimo k-means
  2. Dividir quaisquer clusters cujas amostras são suficientemente diferentes
  3. Agrupar dois clusters suficientemente semelhantes (perto)
  4. Voltar a (1)

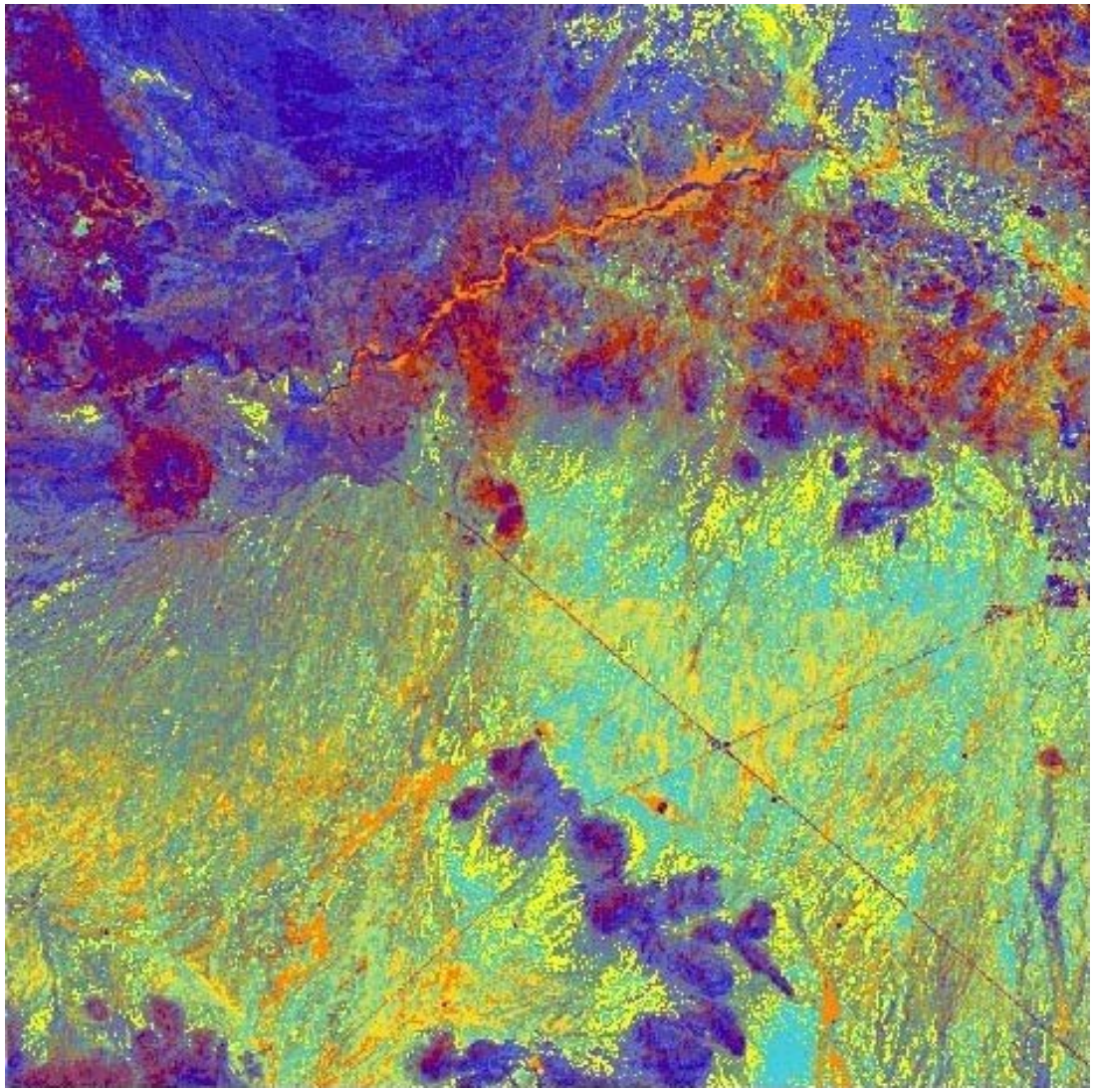
Algoritmo ISODATA

1. Select an initial number of clusters N_C and use the first N_C examples as cluster centers μ_k, k=1..N_C
2. Assign each example to a cluster according to each one's minimum distance to a cluster center
   a. Exit the algorithm if the classification of examples has not changed
3. Eliminate clusters that contain less than N_{MIN_EX} examples and
   a. Assign those examples to the other clusters using minimum distance to cluster centers
   b. Decrease N_C accordingly
4. For each cluster k,
   a. Compute the center μ_k as the sample mean of all the examples assigned to that cluster
   b. Compute the average distance between examples and cluster centers $d_{AVG} = \frac{1}{N}\sum_{k=1}^{N_C} N_k d_k$ and $d_k = \frac{1}{N_k}\sum_{x=\mu_k} |x - \mu_k|$
   c. Compute the variance of each axis and find the axis n* with maximum variance $\sigma_k^2(n^*)$
6. For each cluster k with $\sigma_k^2(n^*) > \sigma_S^2$, if $\{d_k > d_{AVG}$ and $N_k > 2N_{MIN\_EX}+1\}$ or $\{N_C < N_D/2\}$
   a. Split that cluster into two clusters where the two centers μ_{k1} and μ_{k2} differ only in the coordinate n*
      i. $\mu_{k1}(n^*) = \mu_k(n^*) + \sigma_k(n^*)/2$ (all other coordinates remain the same)
      ii. $\mu_{k2}(n^*) = \mu_k(n^*) - \sigma_k(n^*)/2$ (all other coordinates remain the same)
   b. Increment N_C accordingly
   c. Reassign the cluster's examples to one of the two new clusters based on minimum distance to cluster centers
7. If $N_C > 2N_D$ then
   a. Compute all distances $D_{ij} = d(\mu_i, \mu_j)$
   b. Sort D_{ij} in decreasing order
   b. For each pair of clusters sorted by D_{ij}, if (1) neither cluster has been already merged, (2) the distance D_{ij} satisfies $D_{ij} < D_{MERGE}$ and (3) not more than N_{MERGE} pairs of clusters have been merged in this loop, then
      i. Merge i^{th} and j^{th} clusters
      ii. Compute the cluster center $\boldsymbol{\mu}' = \frac{N_i\boldsymbol{\mu}_i + N_j\boldsymbol{\mu}_j}{N_i + N_j}$
      iii. Decrement N_C accordingly
8. Go to step 1

Algoritmo ISODATA (exemplo)



Algoritmo ISODATA (exemplo)



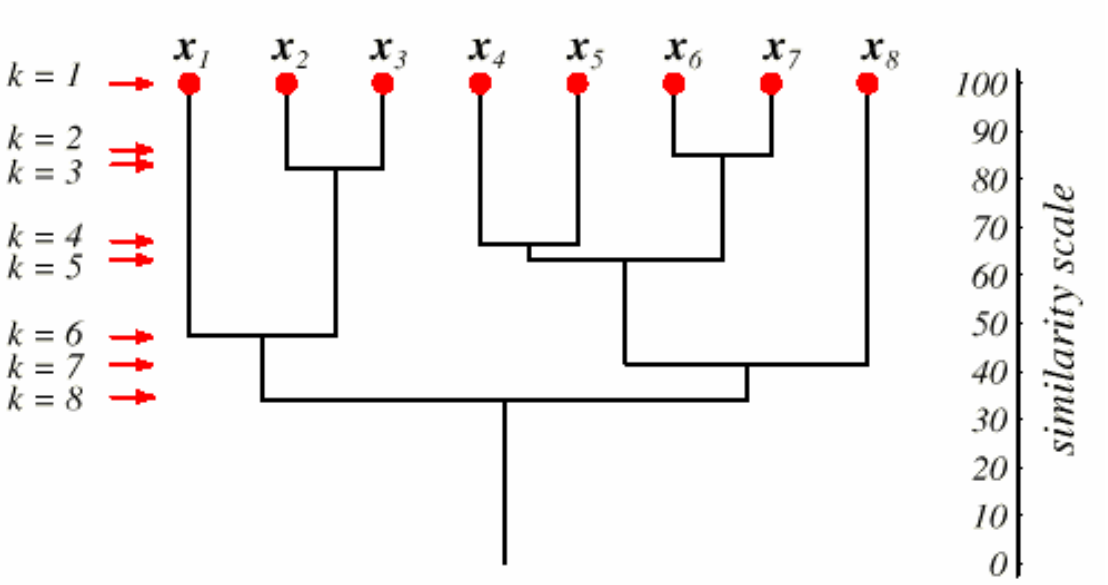
densely vegetated (hillslope or channel) 1
varnished or basaltic rock / soil 1
iron -rich bedrock or soil 1
iron-rich bedrock or soil 2
iron-rich bedrock or soil 3
iron-rich bedrock or soil 4
granular soil +/- granitoid bedrock 1
granular soil +/- varnished bedrock 1
granular soil +/- mixed rock types +/- vegetation 1
densely vegetated (hillslope/channel/cultivated) 1
desert veg. +/- granular soil +/- mixed rock types 1
veg. + wet soil 1
desert veg. +/- granular soil +/- mixed rock types 2
desert veg +/- granular soil +/- mixed rock types 3
desert veg. +/- granular soil +/- mixed rock types 4
desert veg. +/- granular soil +/- mixed rock types 5
granular soil +/- mixed rock types +/- vegetation 2
desert veg. +/- granular soil +/- mixed rock types 6
granular soil +/- mixed rock types +/- vegetation 3
desert veg. +/- granular soil +/- mixed rock types 7

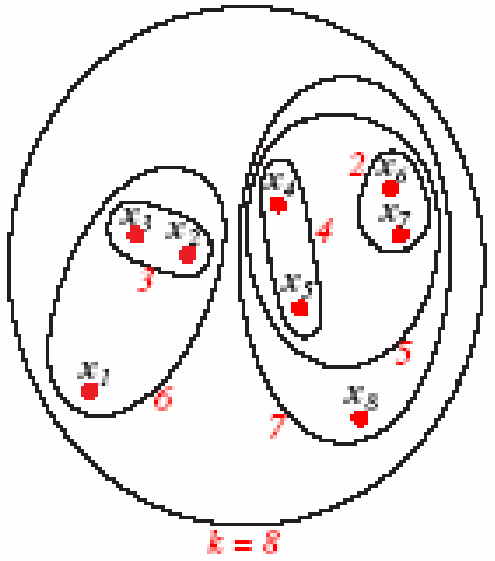
Clustering Hierárquico

- Os algoritmos k-médias, k-médias difuso, LBG e ISODATA criam clusteres disjuntos, resultando numa representação "plana" dos dados.
- Por vezes, é necessário obter uma representação hierárquica dos dados, com clusters e sub-clusters arranjados de uma forma estruturada em arvore.
- Os métodos de clustering hierárquicos podem ser agrupados em duas classes:
  - Aglomerativos (bottom-up, merging): Começa-se com N clusters com uma única amostra e agrupam-se sucessivamente até obter um único cluster.
  - Divisivos (bottom-down, splitting): Começa-se com um cluster (com todos os dados) e divide-se sucessivamente até obter N clusters.

Clustering Hierárquico

- A representação preferida para os clusters hierárquicos é o dendograma
  - Árvore binária que mostra a estrutura dos clusters, também permite medir a semelhança entre clusters (eixo vertical).
  - Alternativa: conjuntos de amostras (diagrama de Venn)





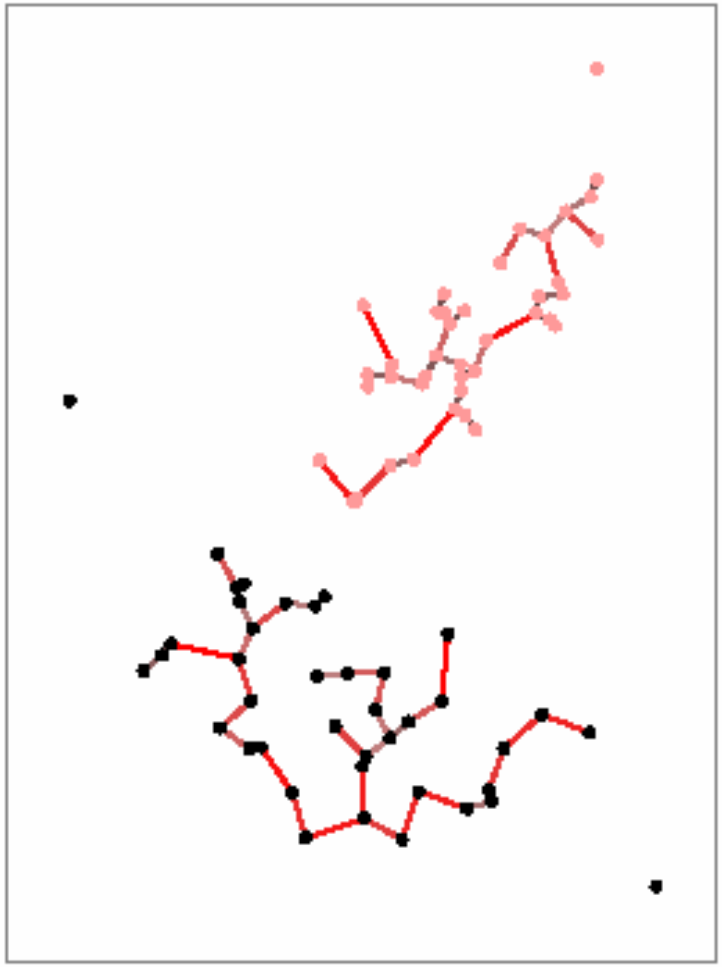
Clustering Aglomerativo

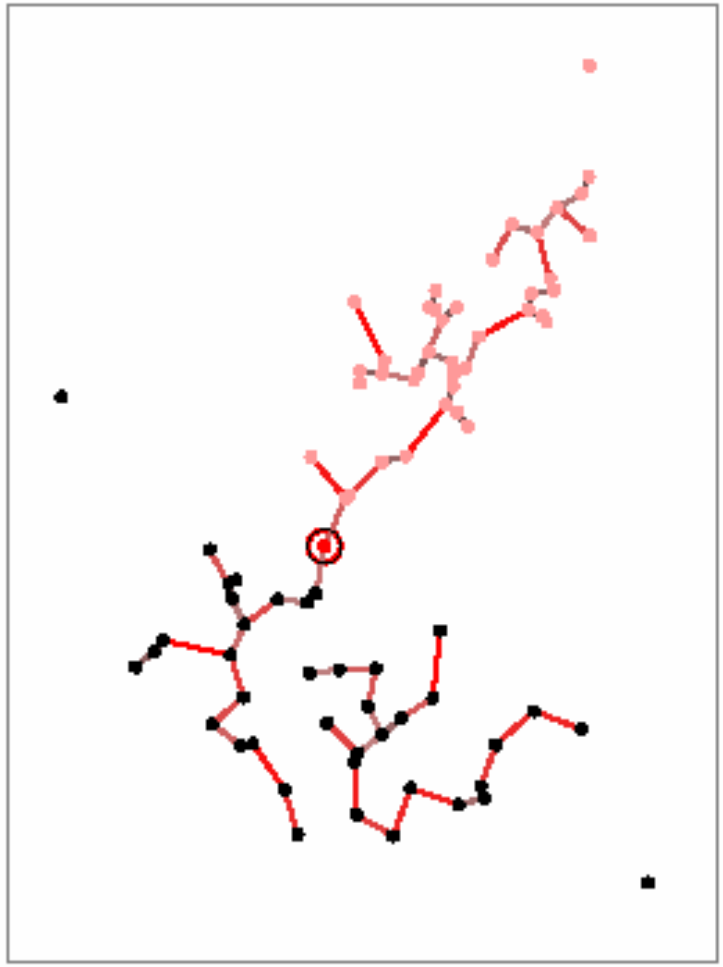
- Passo 1: Atribuir um padrão por cluster (N clusters)
- Passo 2: Encontrar o par de clusters mais semelhantes.
- Passo 3: Juntar os dois padrões num único cluster
- Passo 4: Se o número de clusters (Nc) > 1 voltar a 2.
- Como encontrar os padrões mais semelhantes ?
  - Distância mínima $$d_{\min}(w_i, w_j) = \min_{x \in w_i, y \in w_j} \|x - y\|$$
  - Distância máxima $$d_{\max}(w_i, w_j) = \max_{x \in w_i, y \in w_j} \|x - y\|$$
  - Distância média $$d_{media}(w_i, w_j) = \frac{1}{N_i N_j} \sum_{x \in w_i} \sum_{y \in w_j} \|x - y\|$$
  - Distância entre médias $$d_{entre\_medias}(w_i, w_j) = \|u_i - u_j\|$$

Clustering Aglomerativo

- Distância mínima:
  - Quando utilizada, o algoritmo é referido como o vizinho mais próximo ou single-link.
  - O resultado do algoritmo é uma árvore binária MST (minimum spanning tree)
  - Favorece classes enlongadas.
- Distância máxima:
  - Quando utilizada, o algoritmo é referido como o vizinho mais longe ou complete-link.
  - Cada cluster constitui um sub-grafo completo.
  - Favorece classes compactas.

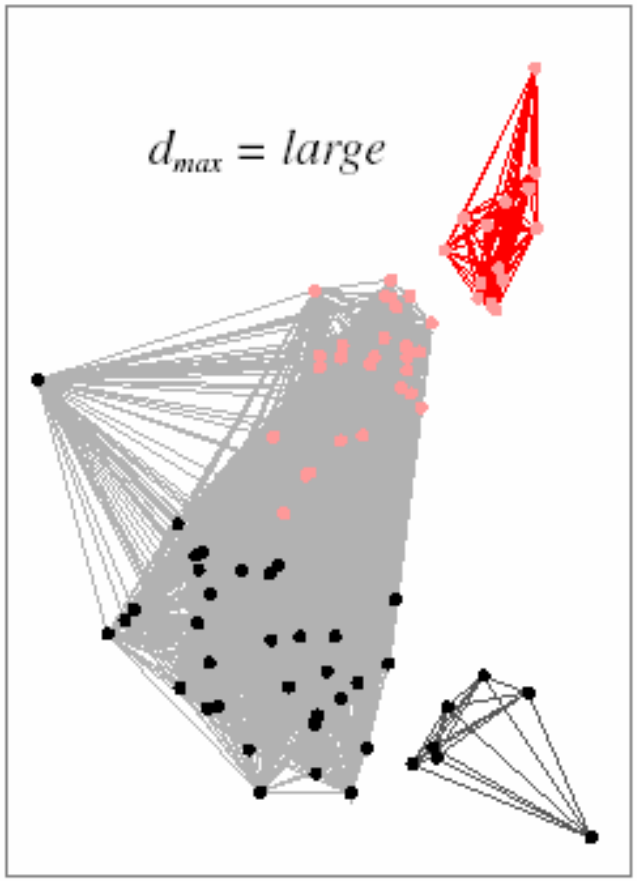
Clustering Aglomerativo (single link)

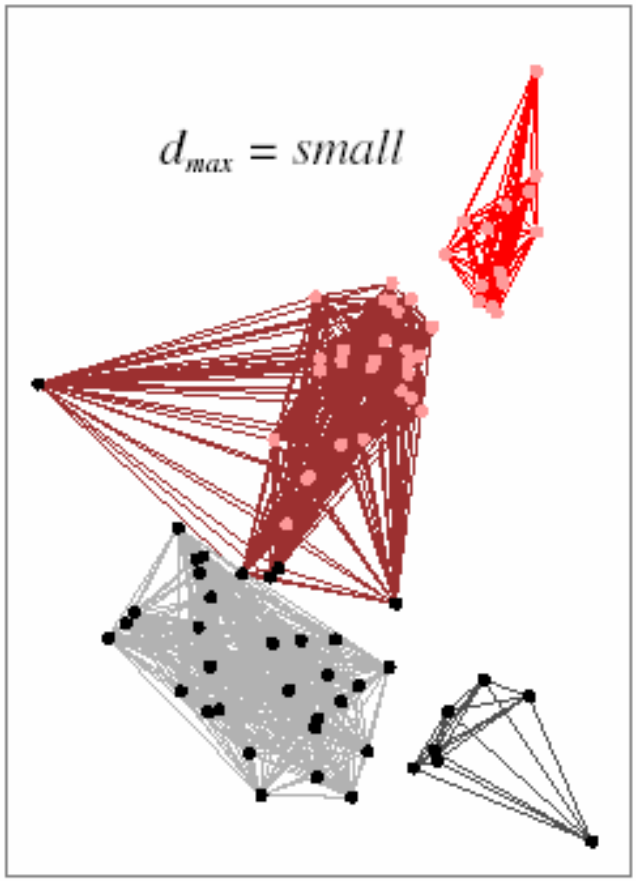




Clustering Aglomerativo (complete link)

- O algoritmo pode terminar quando a distância entre clusters é maior que um limiar (dmax).





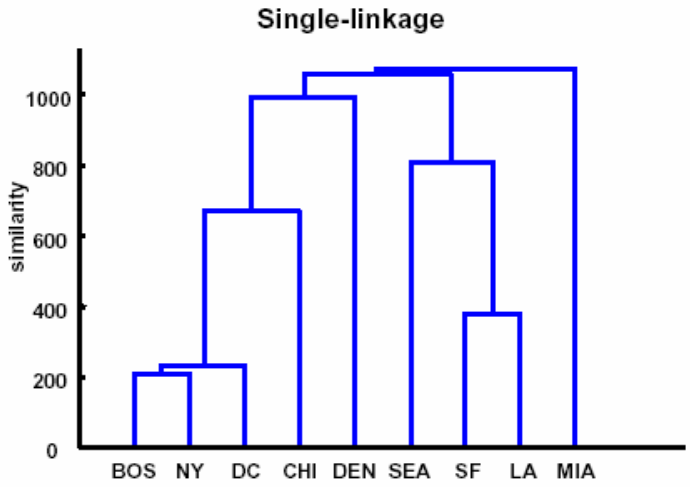
Clustering Aglomerativo

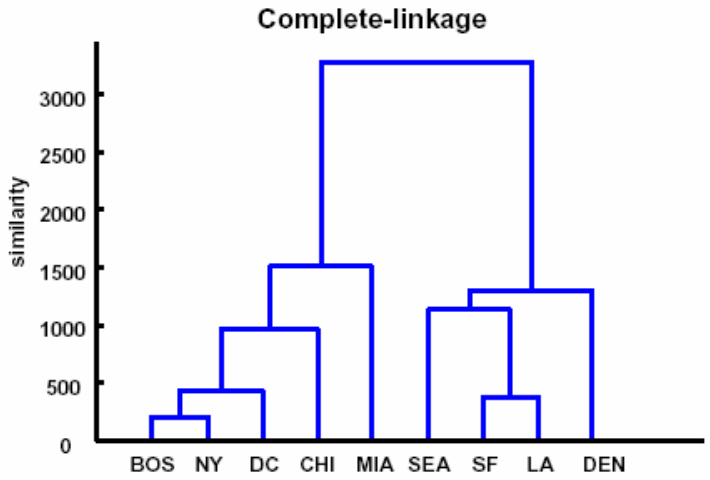
- Distância média ou distância entre médias:
  - A distância mínima ou máxima são bastante sensíveis a outliers (pontos afastados).
  - Estas distâncias permitem resolver esse problema.
  - Das duas a distância entre médias é computacionalmente mais atractiva:
  - Para a distância média é necessário calcular $N_i N_j$ distâncias para cada par de clusters.

Clustering Aglomerativo (exemplo)

- Considere o problema de agrupar as nove maior cidades dos EUA.

	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS	0	206	429	1504	963	2976	3095	2979	1949
NY	206	0	233	1308	802	2815	2934	2786	1771
DC	429	233	0	1075	671	2684	2799	2631	1616
MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
CHI	963	802	671	1329	0	2013	2142	2054	996
SEA	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	0





Clustering Divisivo

- Passo 1: Atribuir todos os padrões (N) a um cluster
- Passo 2: Encontrar o “pior” cluster.
- Passo 3: Dividi-lo em dois.
- Passo 4: Se o número de clusters (Nc) < N voltar a 2.
- Como encontrar o “pior” cluster ?
  - Maior número de amostras.
  - Maior variância
  - Maior erro quadrático médio
  - ...

Clustering Divisivo

- Como dividir um cluster ?
  - Aplicar a média ou a mediana numa das direcções do vector de características.
  - Perpendicular à direcção com maior variância.
  - ...
- Os cálculos necessários no clustering divisivo são mais intensos que o clustering hierárquico, e logo os métodos aglomerativos são mais comuns.