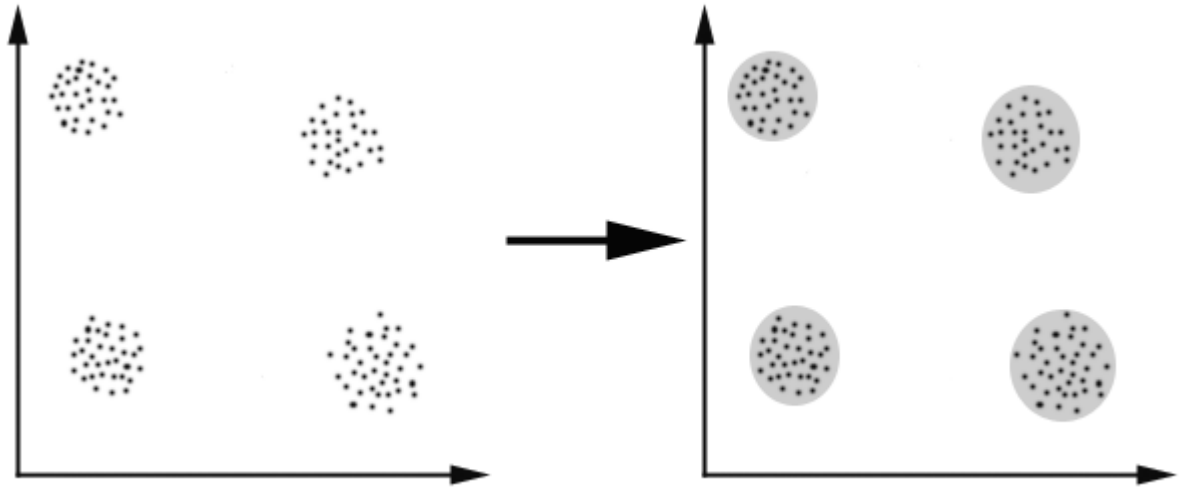
Clusterização

PROF. RICARDO SUYAMA

PROF. LUNEQUE JUNIOR

TITO CACO CURIMBABA SPADINI

O que é Clusterização?



“Encontrar K agrupamentos (ou uma classificação que consiste de K agrupamentos) de maneira que os objetos em cada agrupamento (cluster) sejam similares, e objetos de clusters distintos sejam dissimilares”

Como definer “similaridade”?

Depende da aplicação – por exemplo, no caso de classificação de textos, apresentamos uma forma de quantificar a similaridade entre os objetos com a medida do cosseno.

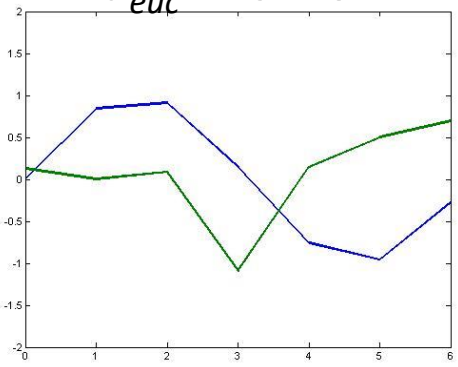
A definição da métrica de similaridade, muitas vezes, é mais importante do que o próprio algoritmo que realizará a clusterização.

Distância Euclideana

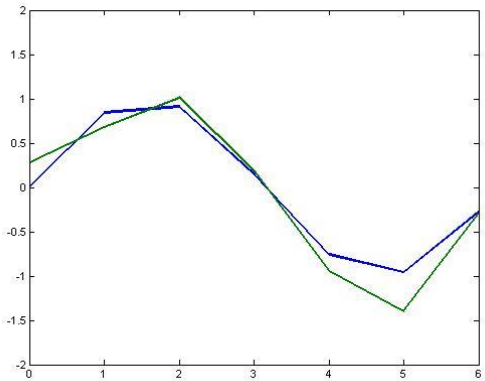
Reflete a nossa noção usual de distância entre dois pontos, e é definida matematicamente por

$$d_{euclid}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$

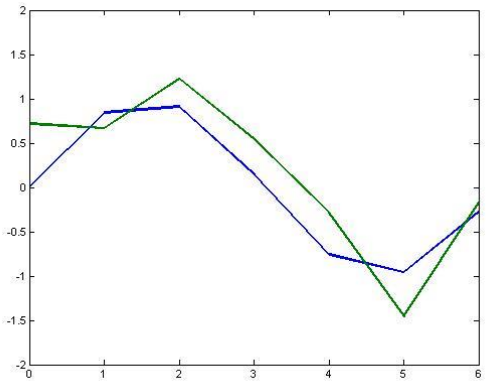
d_{euc}=2.6115



d_{euc}=0.5846



d_{euc}=1.1345



Correlação

Se o interesse não for tanto nos valores das amplitudes das components do vetor, pode-se utilizar o conceito de correlação

A Correlação linear de Pearson é definida como

$$\rho(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

É invariante a escalas e deslocamentos verticais (soma de constants)

Possui valores entre -1 e 1

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i}^{n} y_i$$

Objetivo Global

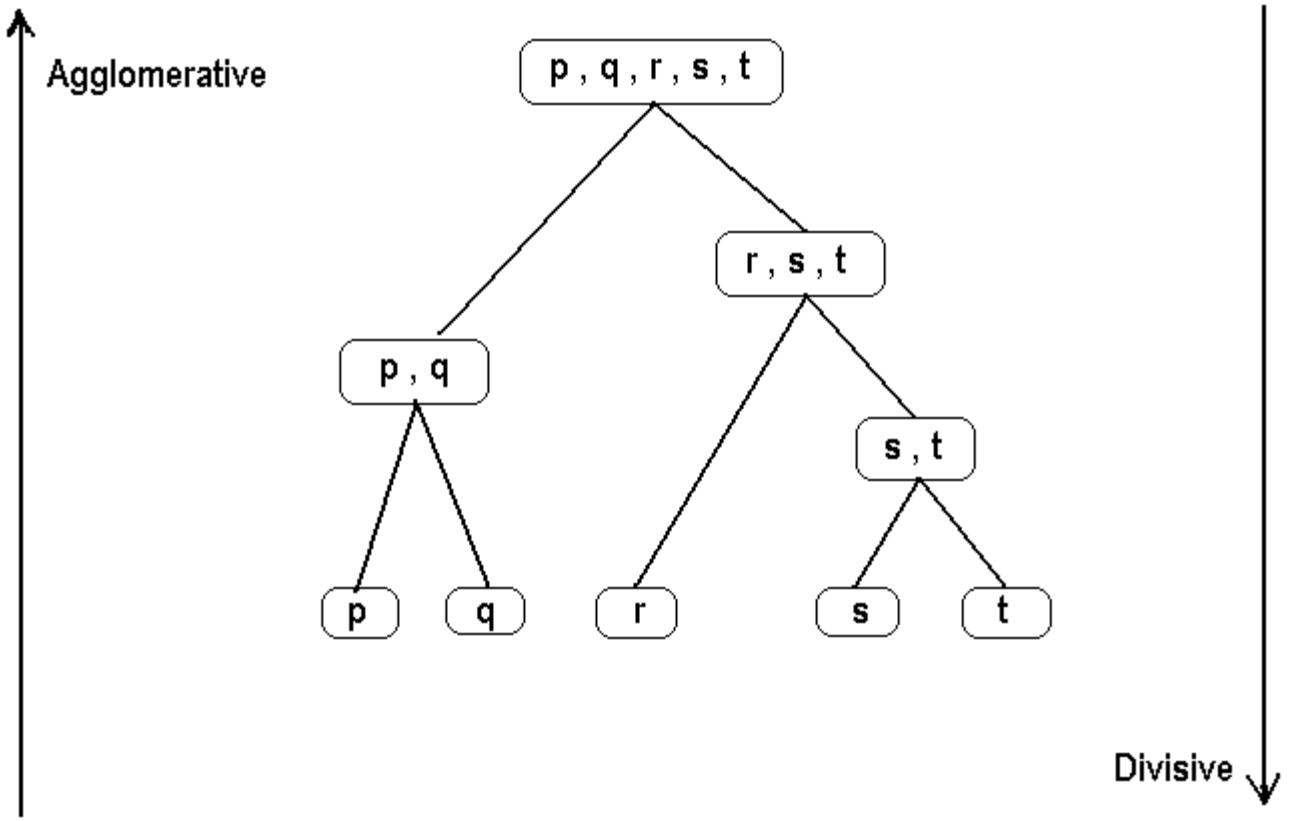
Determinar o agrupamento intrínseco dos dado.

Um bom método de clusterização deve produzir agrupamentos com

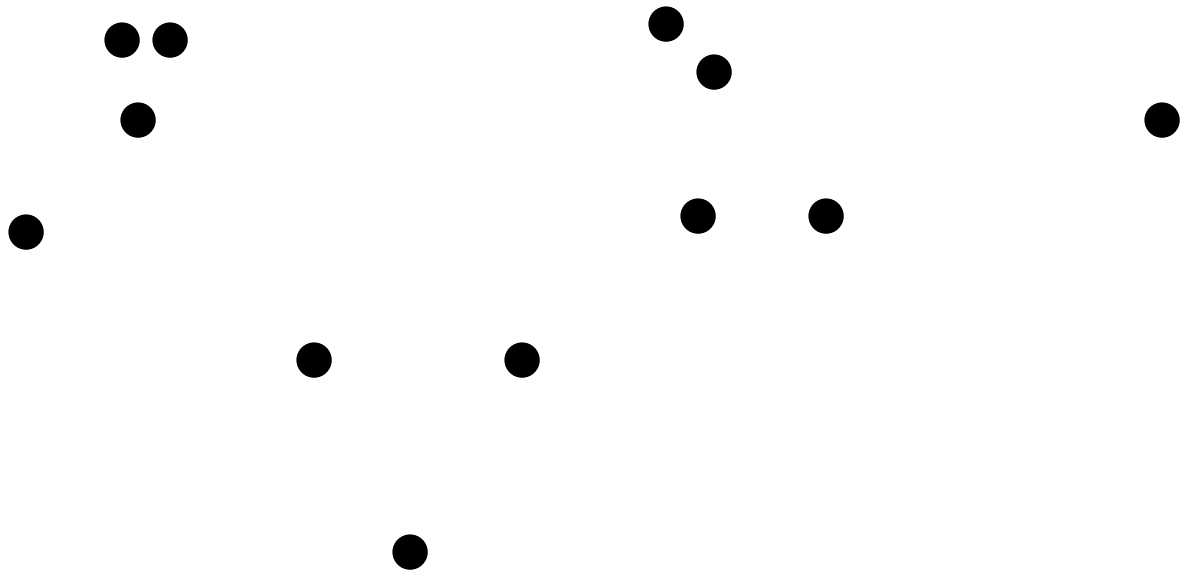
- alta similaridade intra-classe
- baixa similaridade entre classes

A qualidade dos resultados da clusterização depende tanto da medida de similaridade quanto do algoritmo.

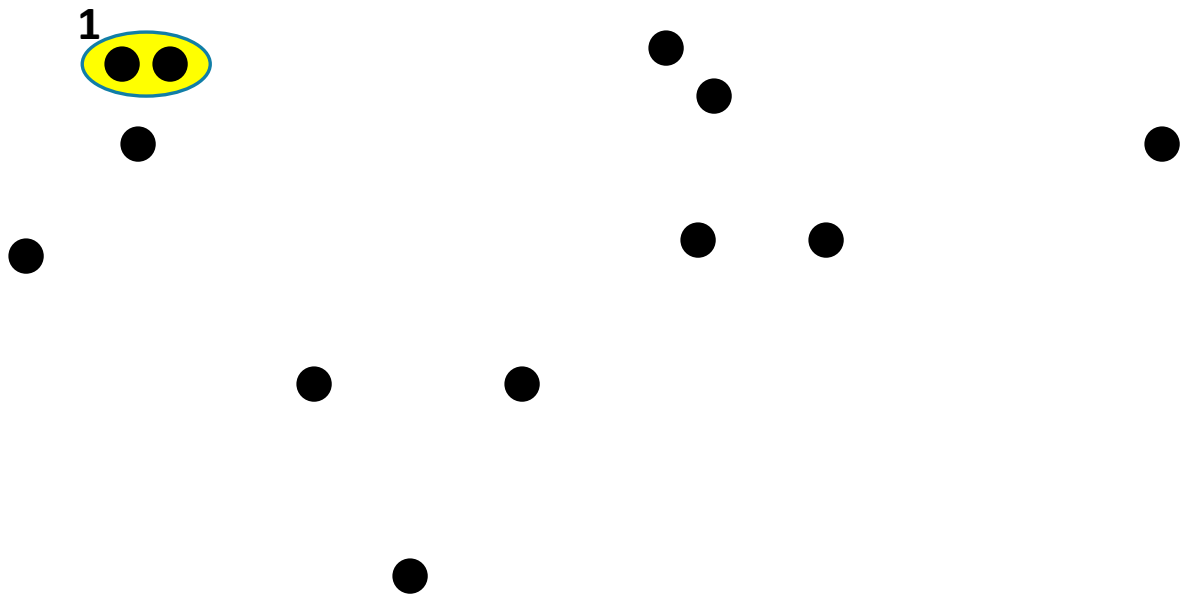
Clusterização Hierárquica



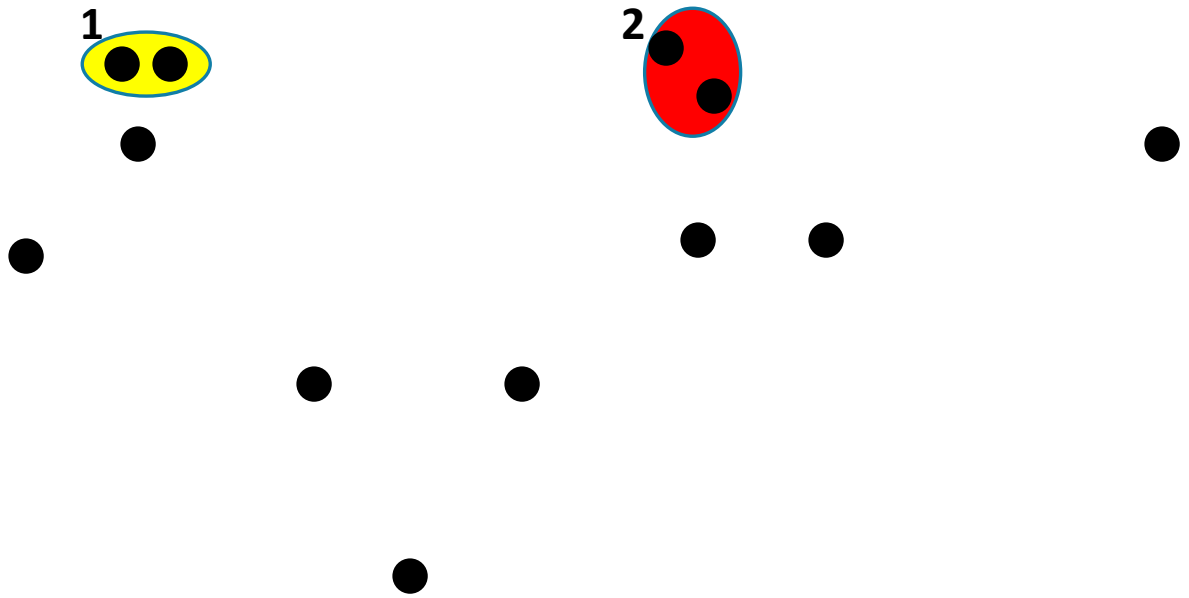
Clusterização Hierárquica – Aglomerativa (Bottom Up)



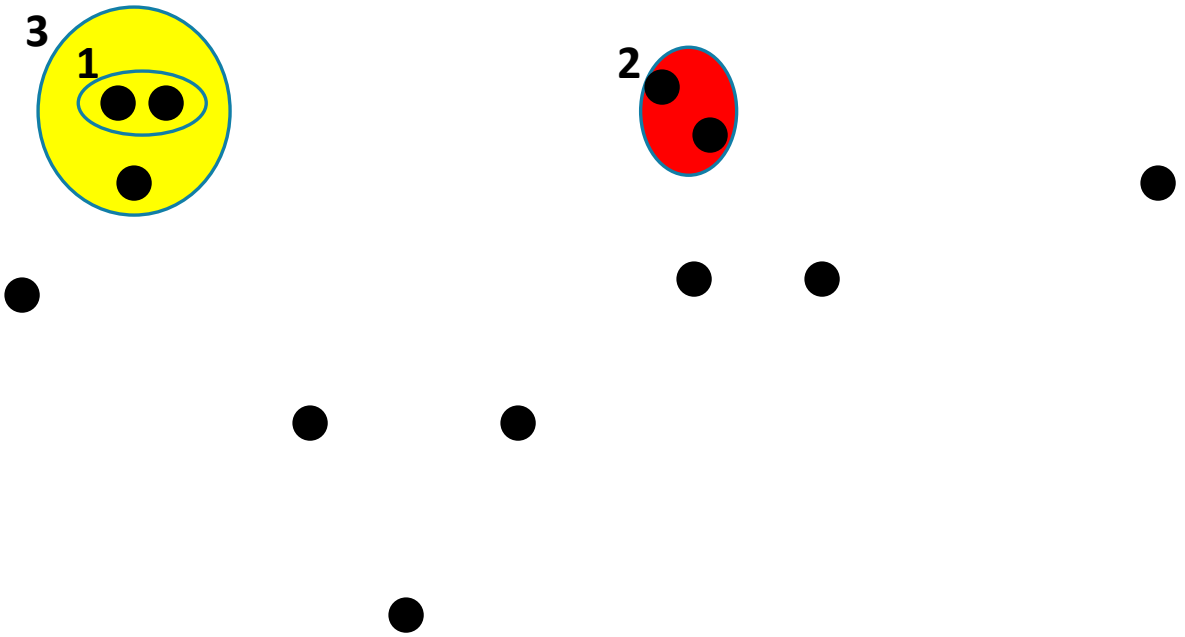
Clusterização Hierárquica – Aglomerativa (Bottom Up)



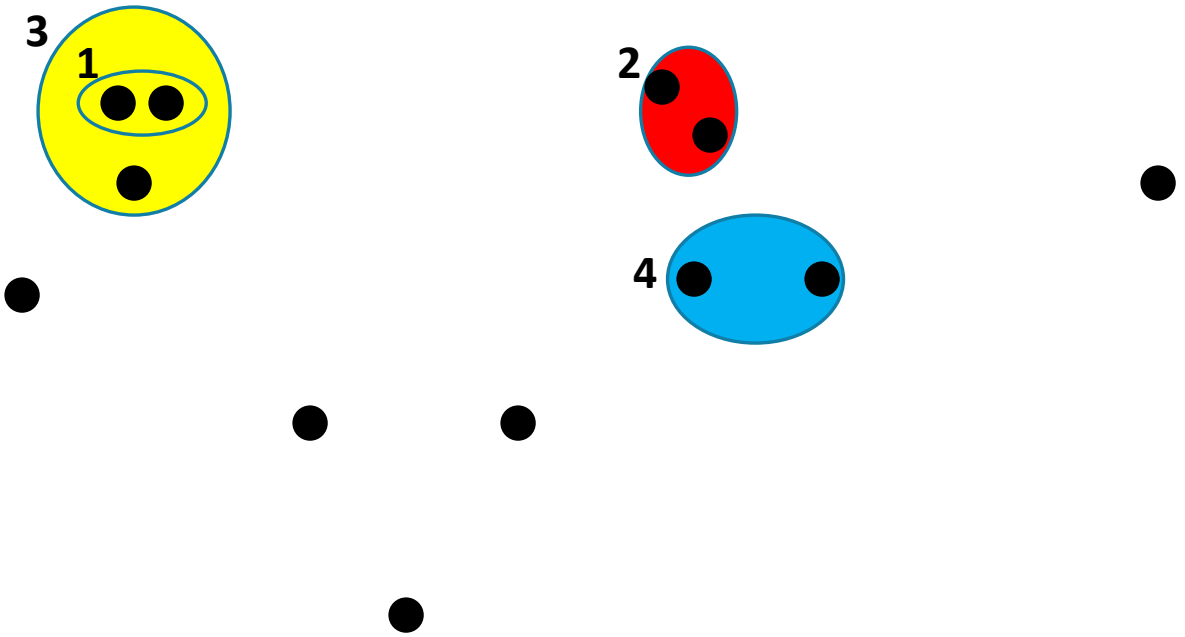
Clusterização Hierárquica – Aglomerativa (Bottom Up)



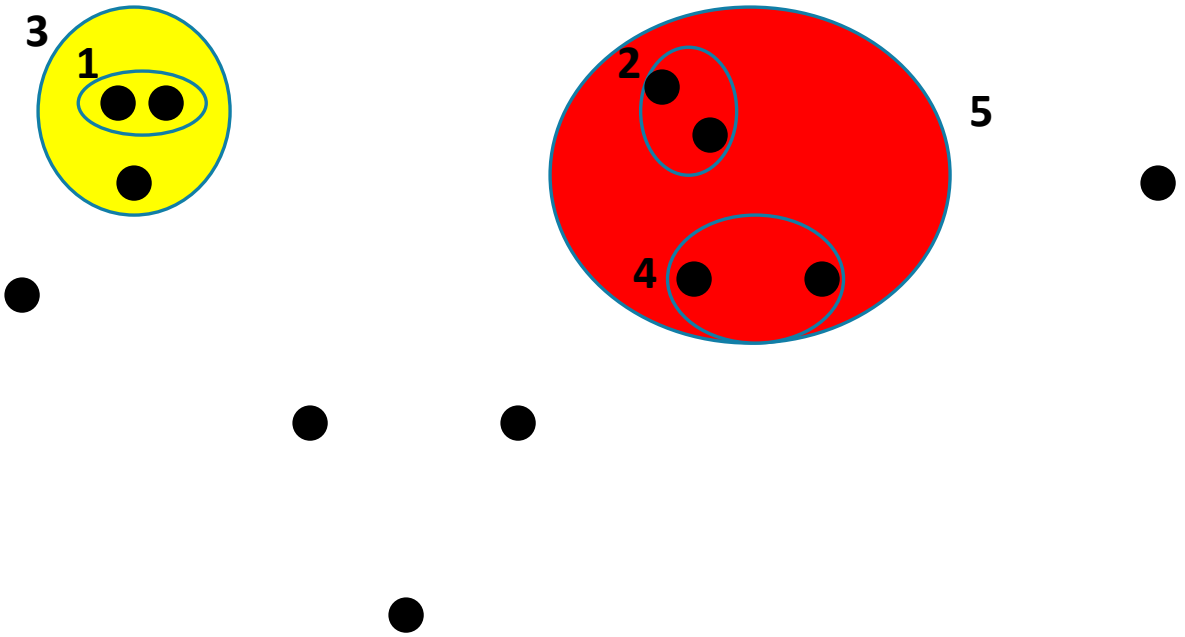
Clusterização Hierárquica – Aglomerativa (Bottom Up)



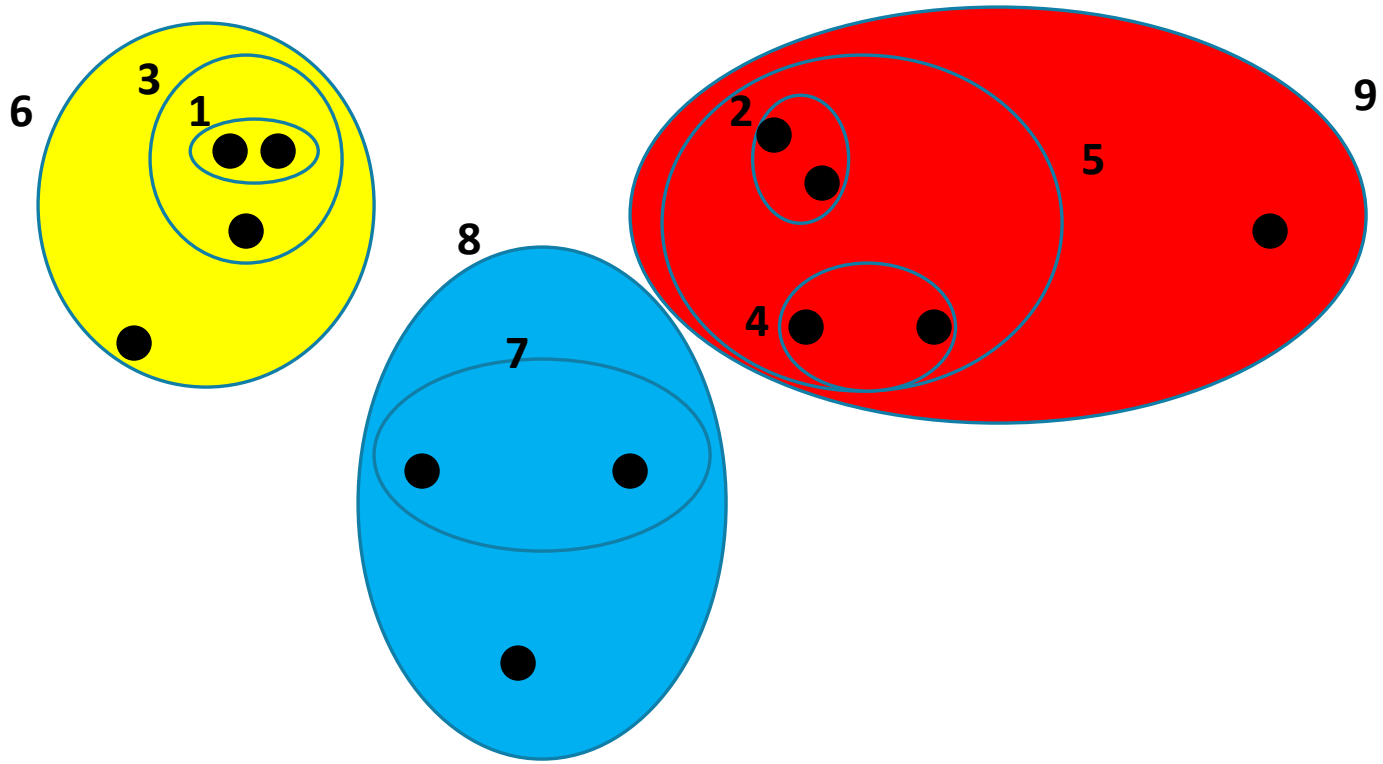
Clusterização Hierárquica – Aglomerativa (Bottom Up)



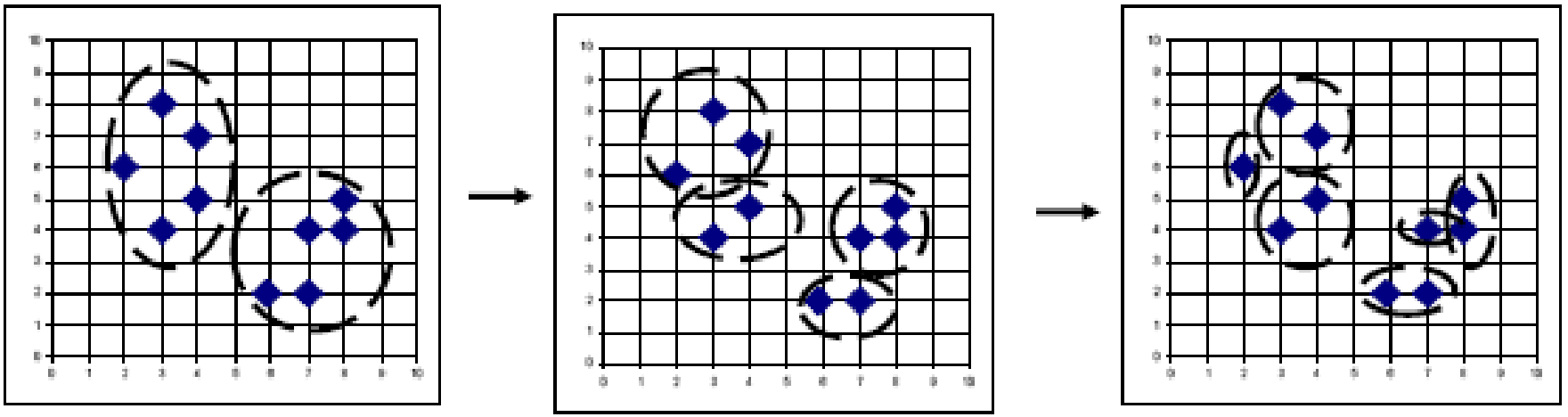
Clusterização Hierárquica – Aglomerativa (Bottom Up)



Clusterização Hierárquica – Aglomerativa (Bottom Up)



Clusterização Hierárquica – Divisiva (Top Down)



K-means

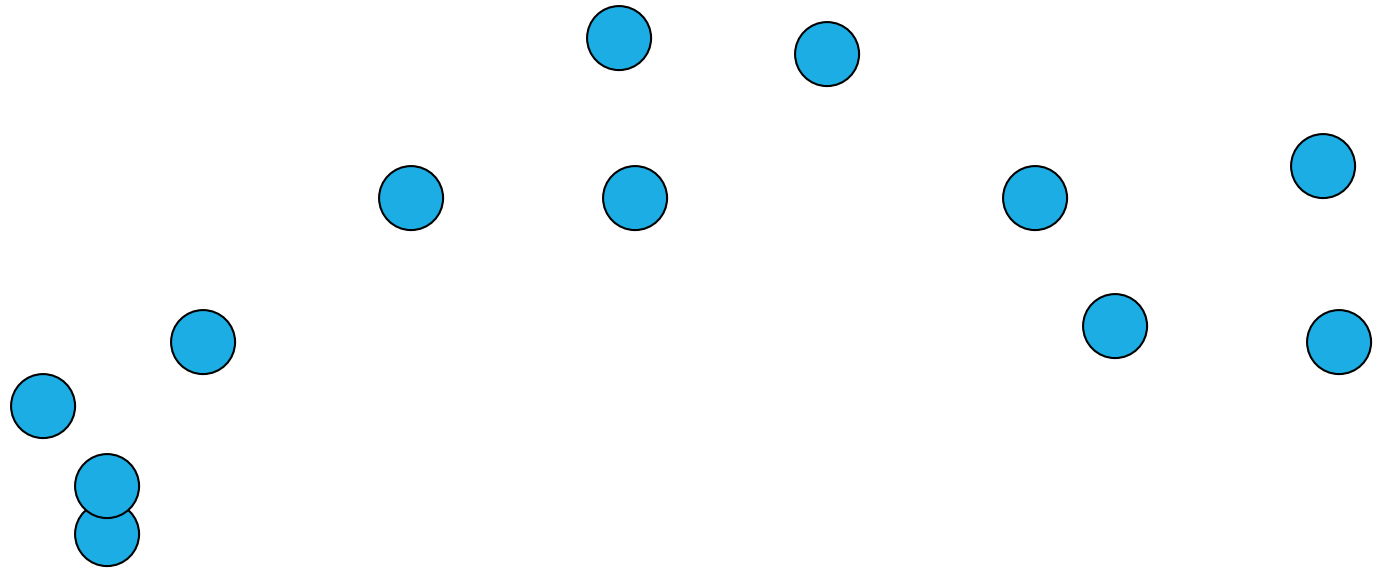
Um dos métodos mais populares para clusterização de dados

Ideia consiste em agrupar os dados de acordo com a sua proximidade a vetores denominados centros.

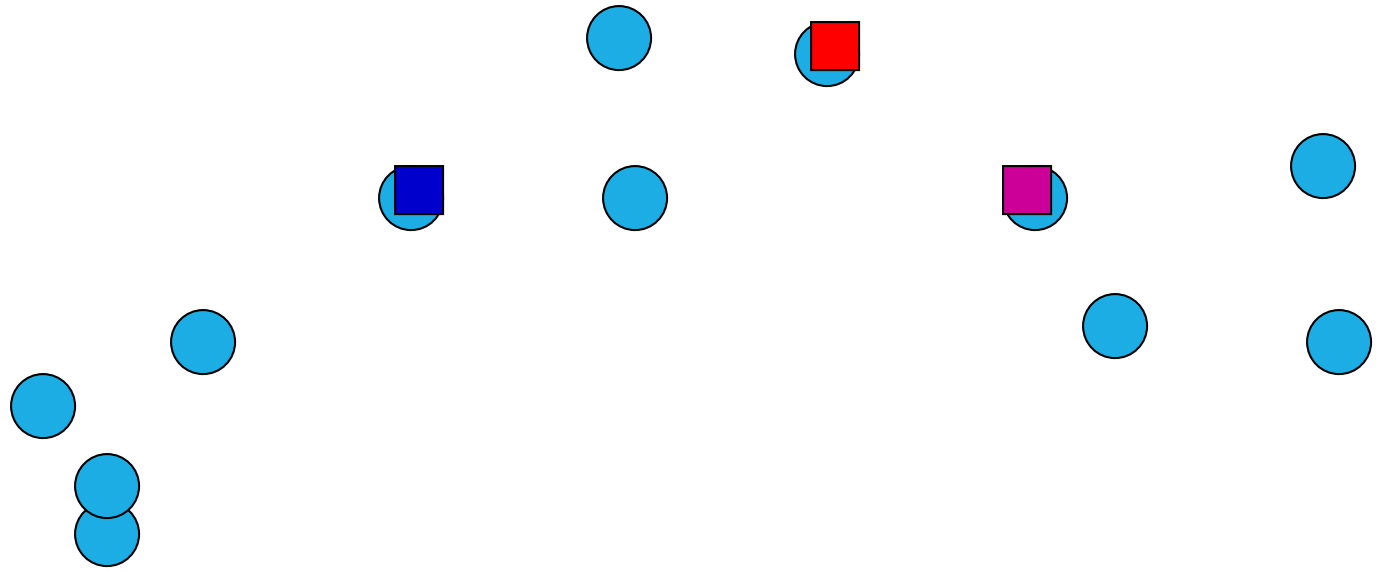
O procedimento é simples:

- Inicie com um número K de centros, inicializados aleatoriamente
- A cada iteração do algoritmo:
  - Atribua cada um dos dados a um agrupamento, que é definido por um dos centros (pela sua proximidade a um dos centros)
  - Recalcule a posição dos centros como a média dos pontos do agrupamento

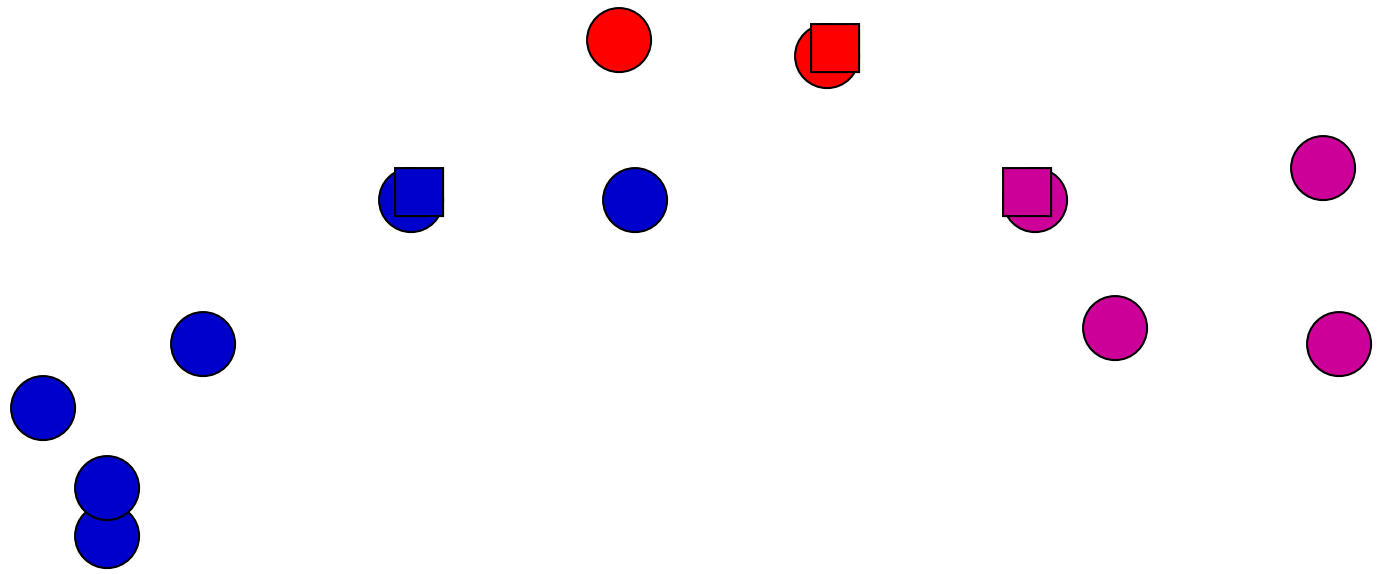
Examplo do k-means



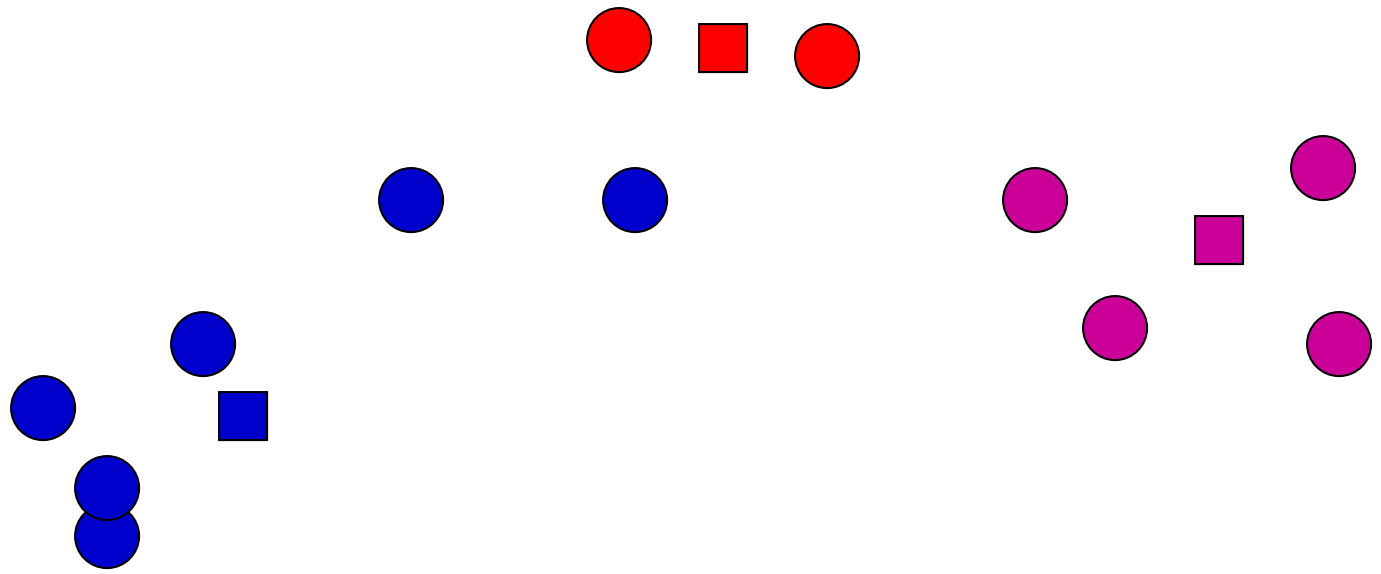
Examplo do k-means – Incializando os centros



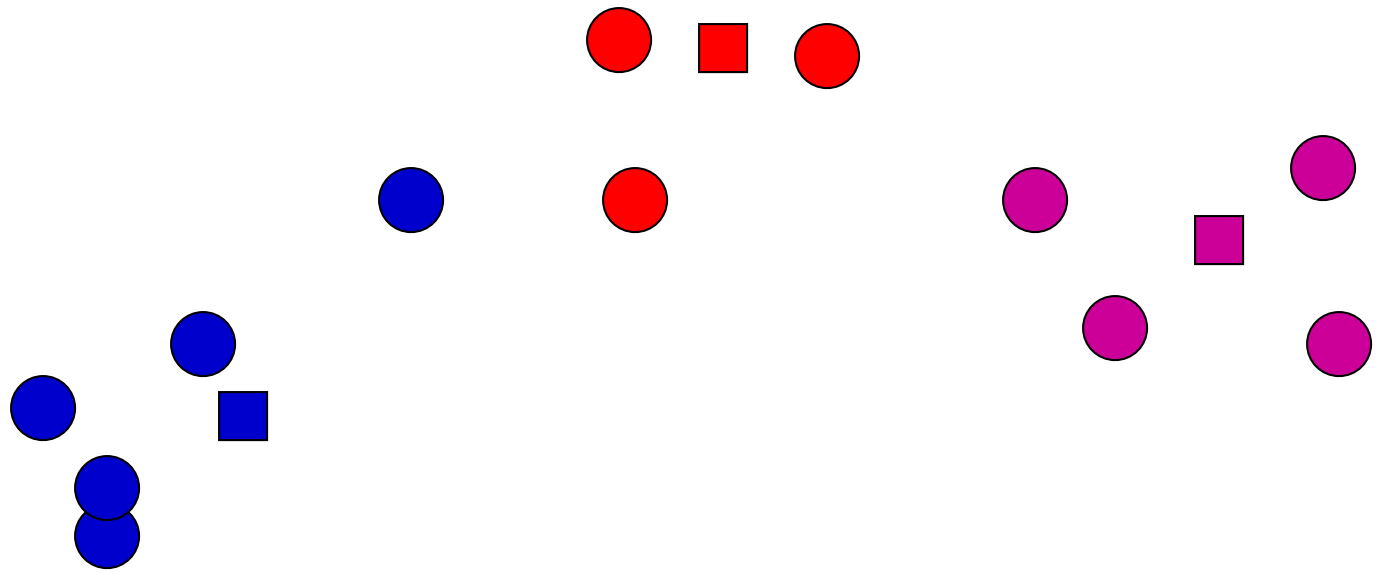
Examplo do k-means – Atribuindo os clusters



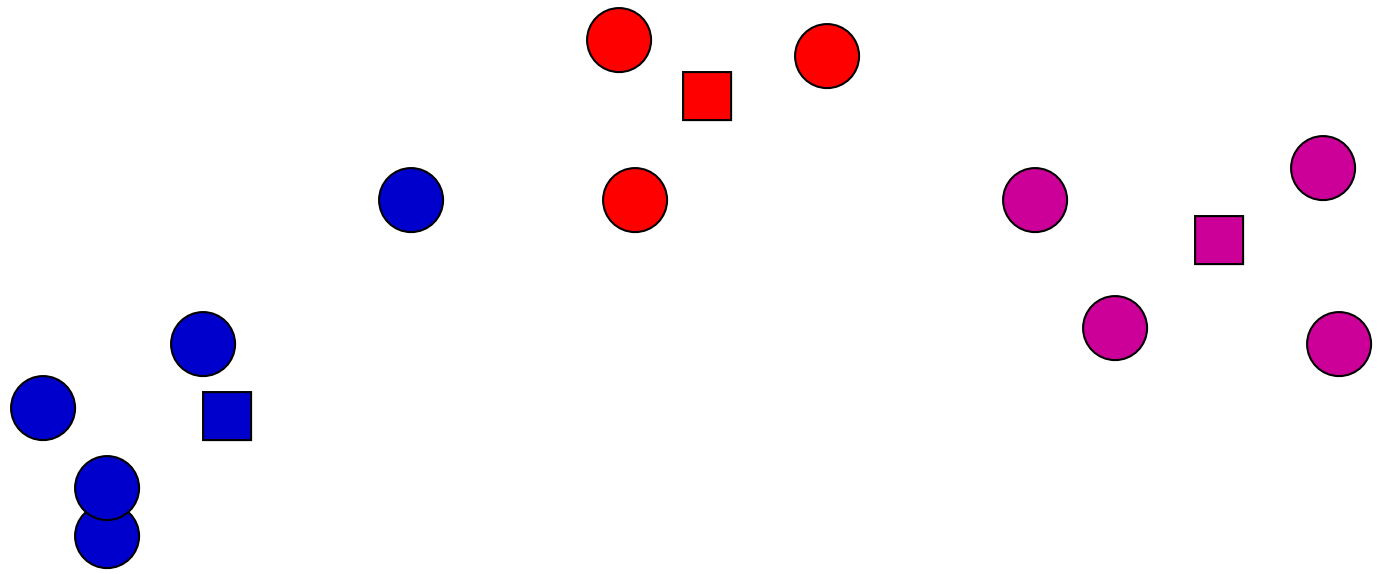
Examplo do k-means – Reajustando os centros



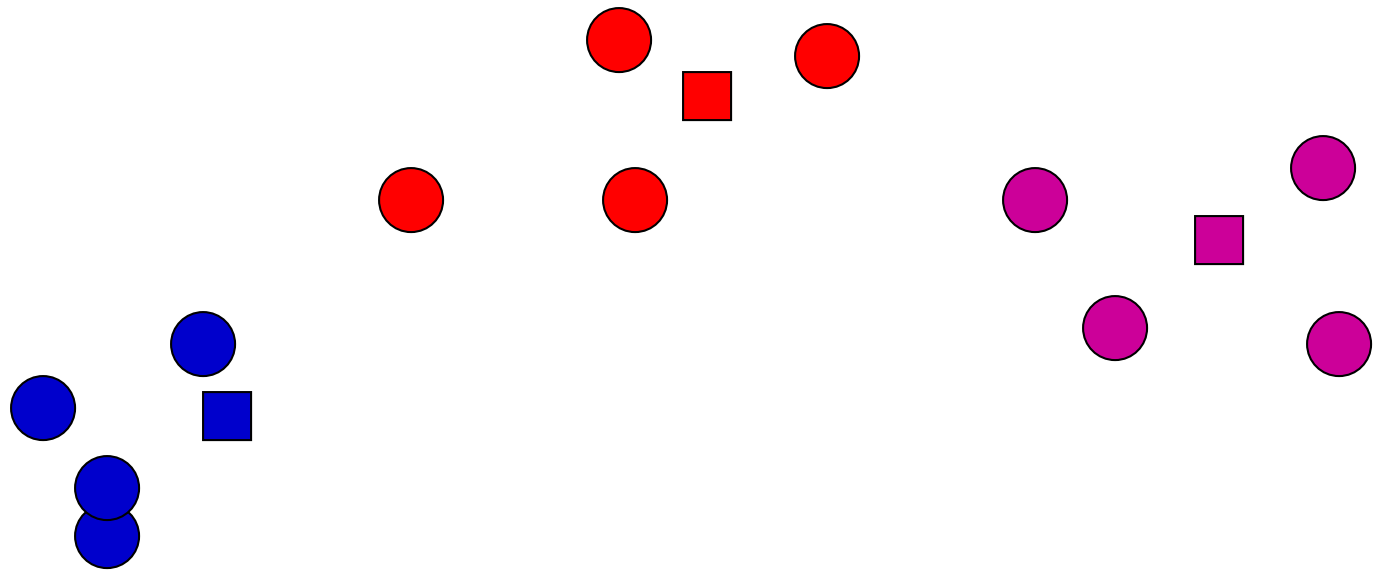
Examplo do k-means – Novos agrupamentos



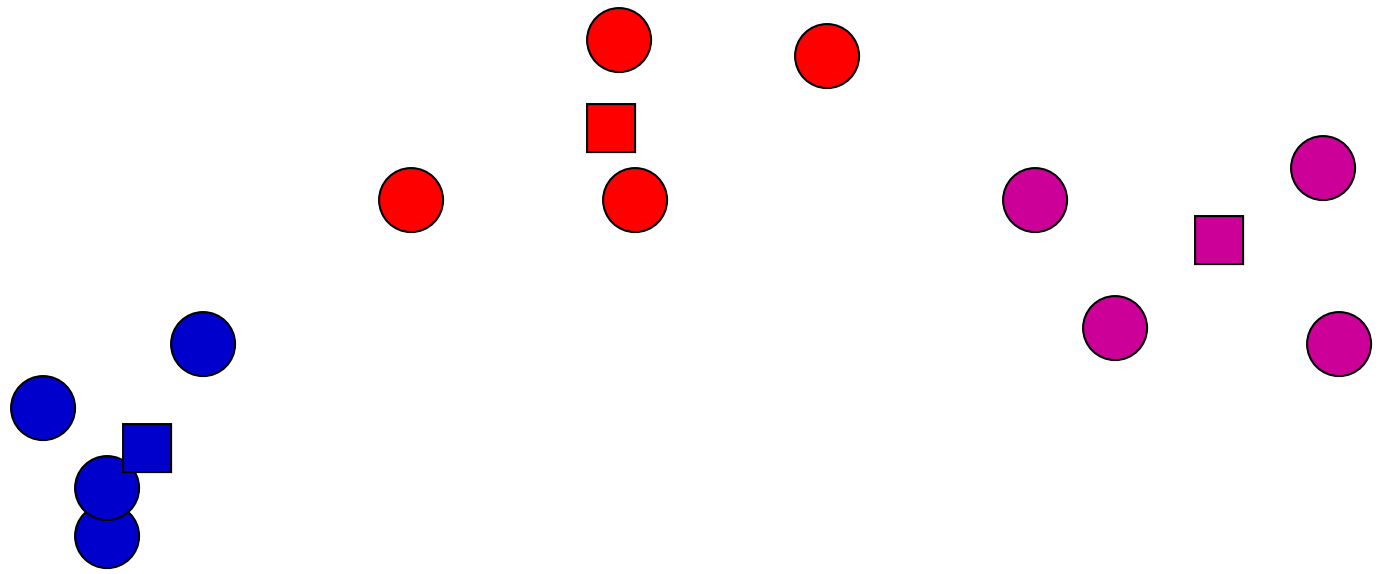
Examplo do k-means – Reajustando os centros



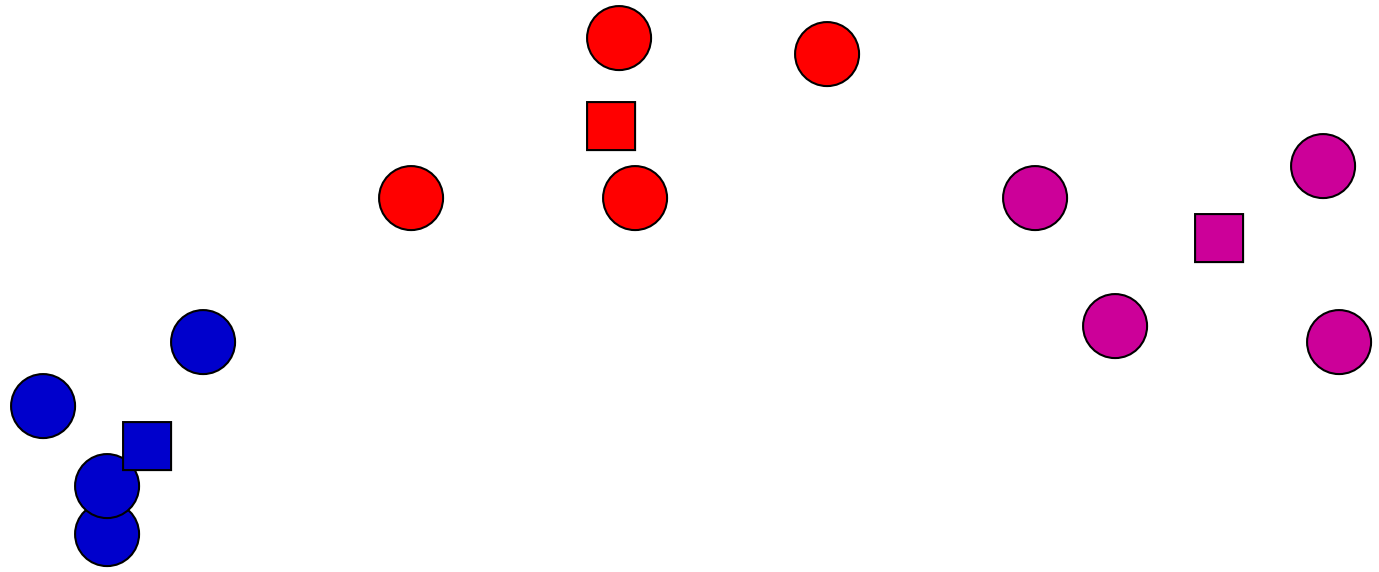
Examplo do k-means – Novos agrupamentos



Examplo do k-means – Reajustando os centros

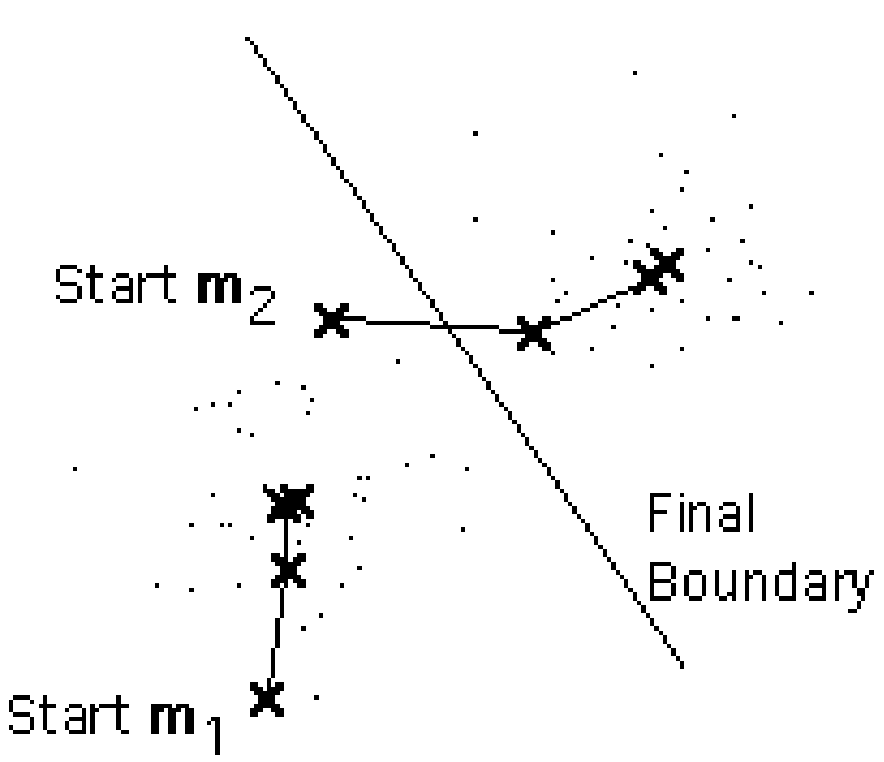


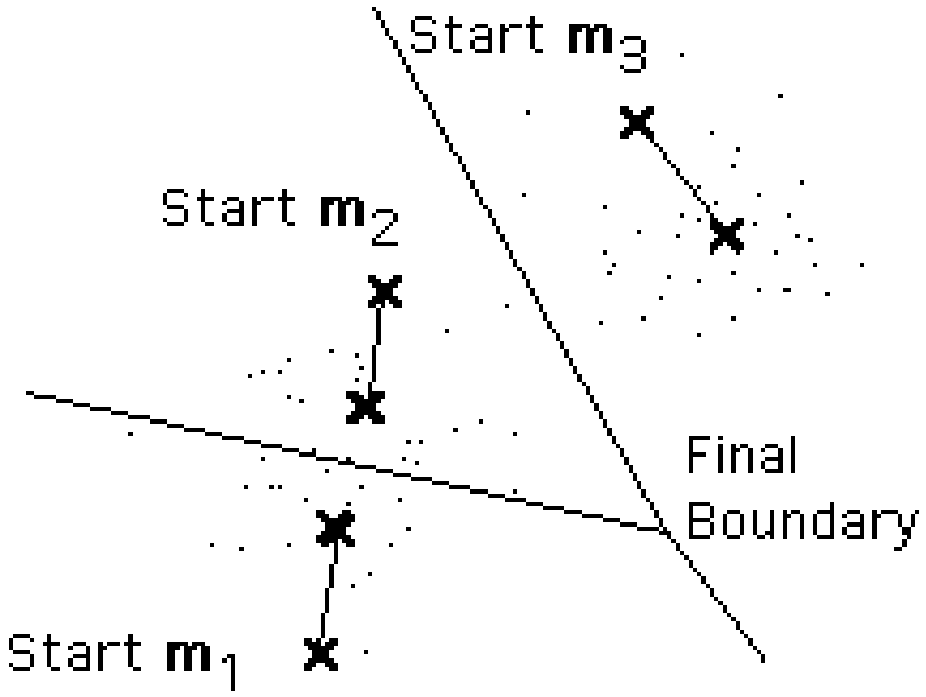
Examplo do k-means – Novos Agrupamentos



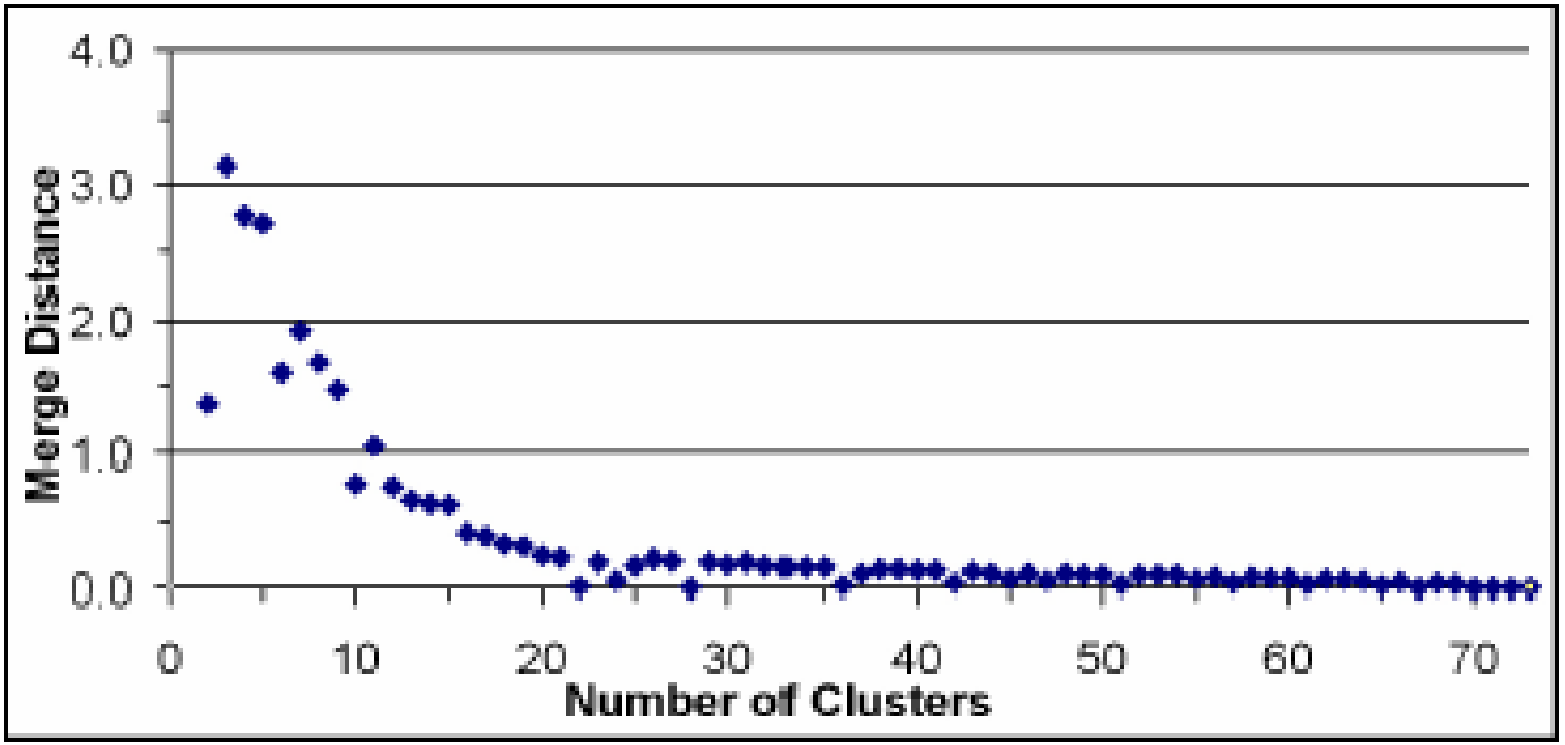
Sem mudanças

Quantos agrupamentos devo considerar?





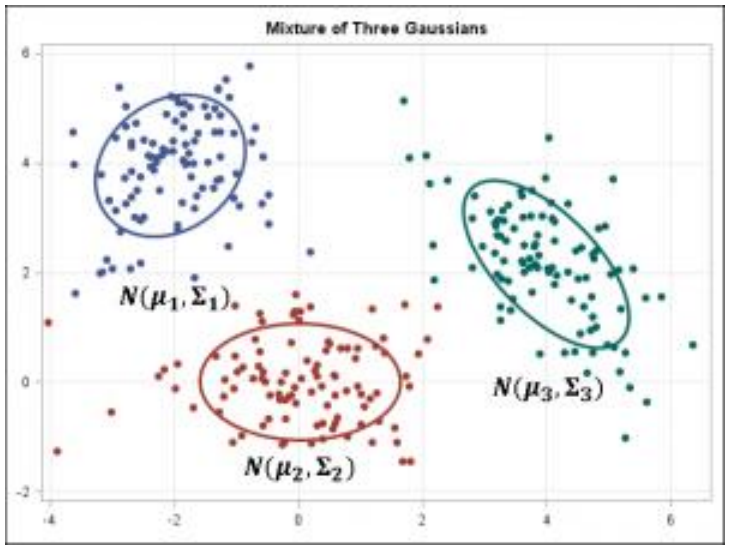
Número de Clusters

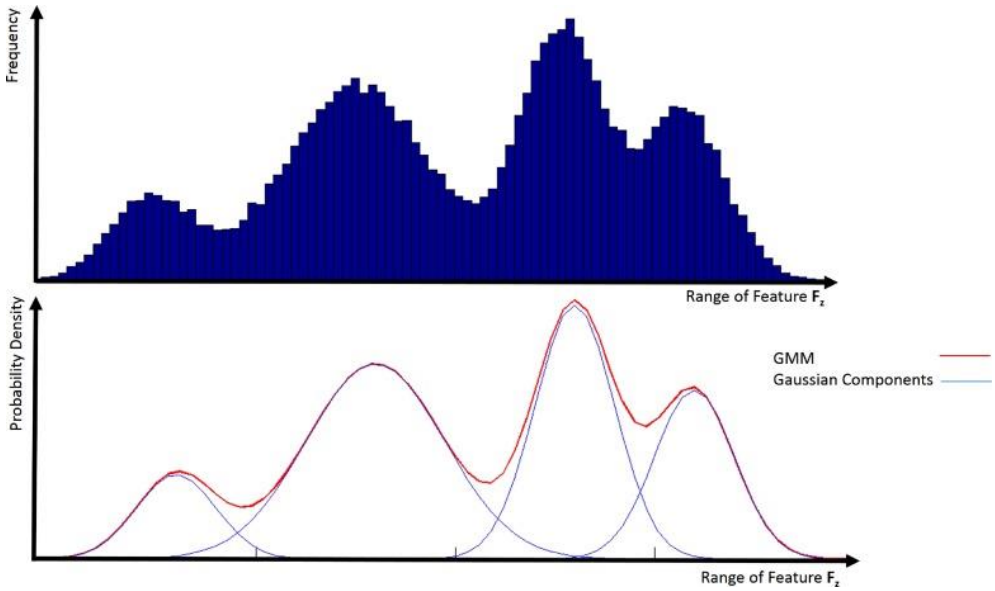


Clusterização baseada em modelo probabilístico – Mistura de Gaussianas

Nessa abordagem, assume-se que os dados observados foram gerados por um modelo probabibilístico, que pode ser construído como uma mistura de distribuições gaussianas

O desafio, nesse caso, consiste em obter os parâmetros que melhor representam os dados.





Como obter os parâmetros ótimos?

No modelo de mistura de gaussianas, é necessário definir

- O Número de componentes
- A média de cada função normal
- A variância (ou matriz de covariância) de cada componente

Um dos métodos mais utilizados para isso é o algoritmo Expectation Maximization (EM)

Affinity Propagation

No algoritmo de propagação de afinidade, os dados são considerados nós em uma rede, e os agrupamentos são formados por meio da troca de mensagens entre os nós.

Inicialmente, todos os nós são considerados como “centroides” (exemplares dos agrupamentos)

Entradas:

- Conjunto de similaridades, $\{s(i,k)\}$, onde $s(i,k)$ é um número real que indica quão bem um dado k representa o dado i, e.g.

$$s(i,k) = -\|\mathbf{x}_i - \mathbf{x}_k\|^2, i \neq k$$

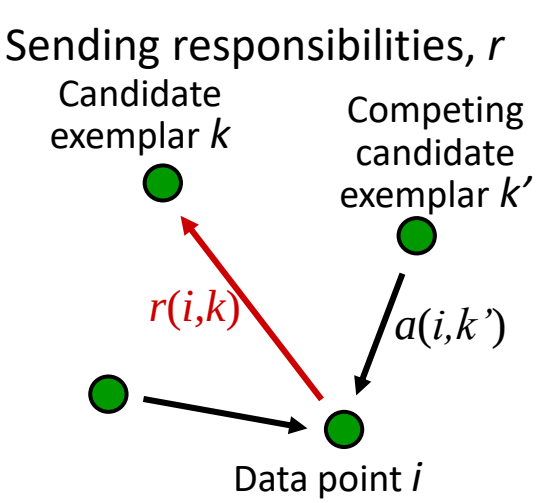
- Para cada dado k, um número real $s(k,k)$ deve indicar a preferência (a priori) de que ele seja selecionado como um centroide, e.g.

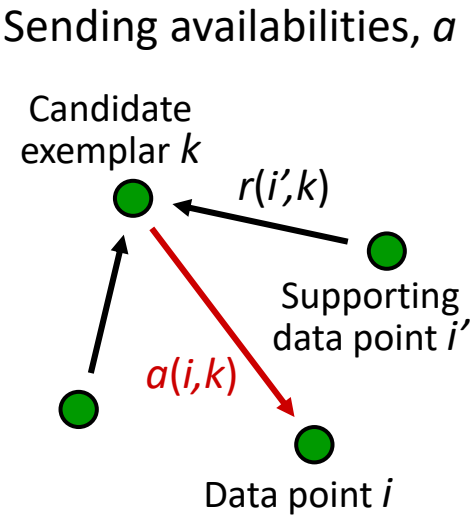
$$s(k,k) = p \; \forall k$$

Propagação das mensagens

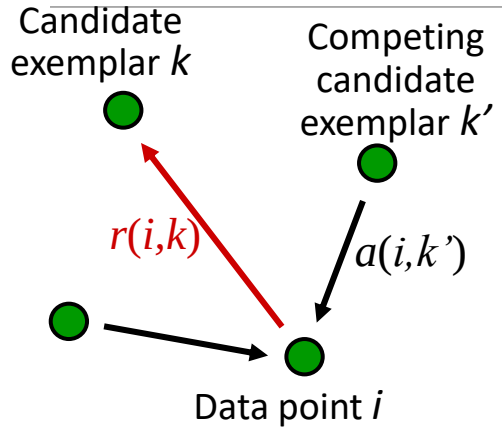
As mensagens que são trocadas entre os nós da rede são

- Responsabilidade $r(i,k)$, enviada do nó i para o nó k, que representa a evidência acumulada de que k é um exemplar de i (i.e., de que i pertence ao cluster cujo exemplar é k)
- Disponibilidade $a(i,k)$, enviada do nó k para i, que representa a evidência acumulada de que k é o exemplar do cluster ao qual pretence i

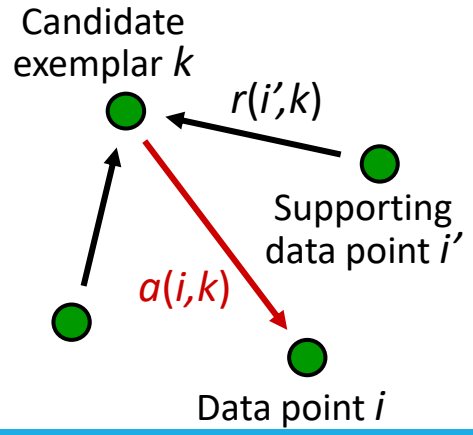




Atualização das Mensagens



Sending availabilities



Inicialmente todas as responsabilidades de disponibilidades são definidas como sendo nulas

Depois, as mensagens são atualizadas de acordo com

$$r(i,k) \leftarrow s(i,k) - \max_{k' \text{ s.t. } k' \neq k} \{a(i,k') + s(i,k')\}$$

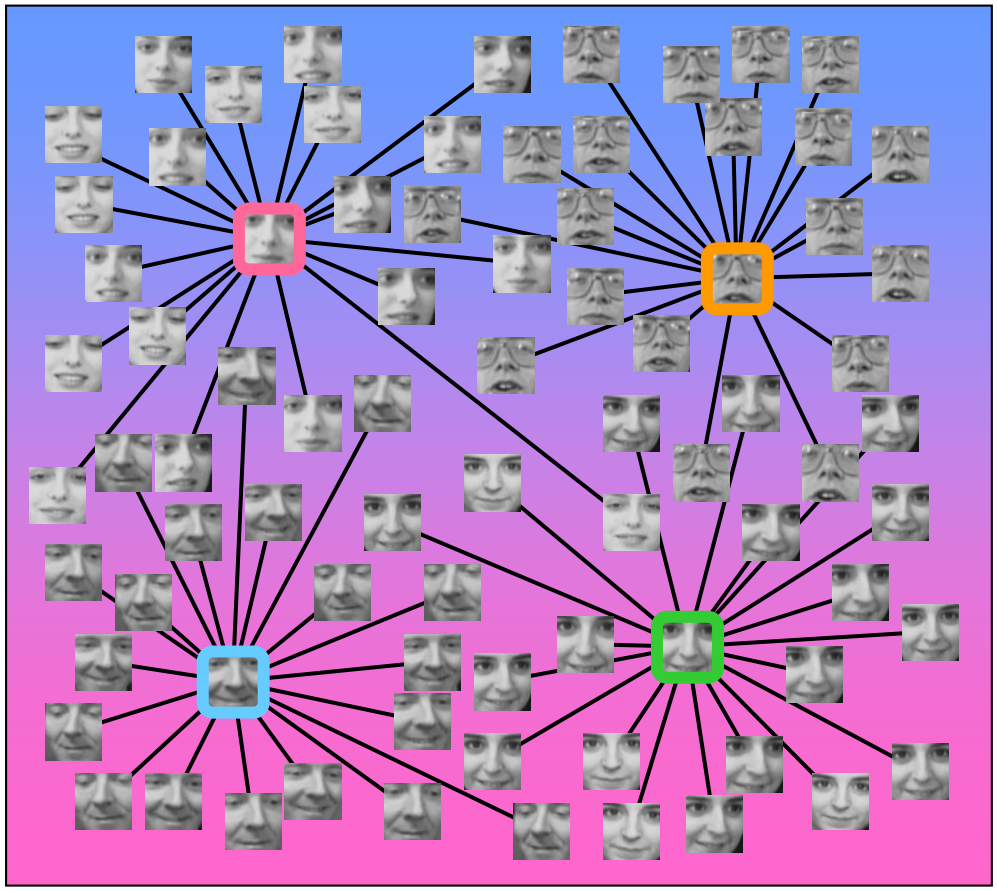
$$a(i,k) \leftarrow \min\left\{0, r(k,k) + \sum_{i' \text{ s.t. } i' \notin \{i,k\}} \max\{0, r(i',k)\}\right\}$$

$$a(k,k) \leftarrow \sum_{i' \text{ s.t. } i' \neq k} \max\{0, r(i',k)\}$$

Making decisions:

$$\text{argmax}_k \{a(i,k) + r(i,k)\}$$

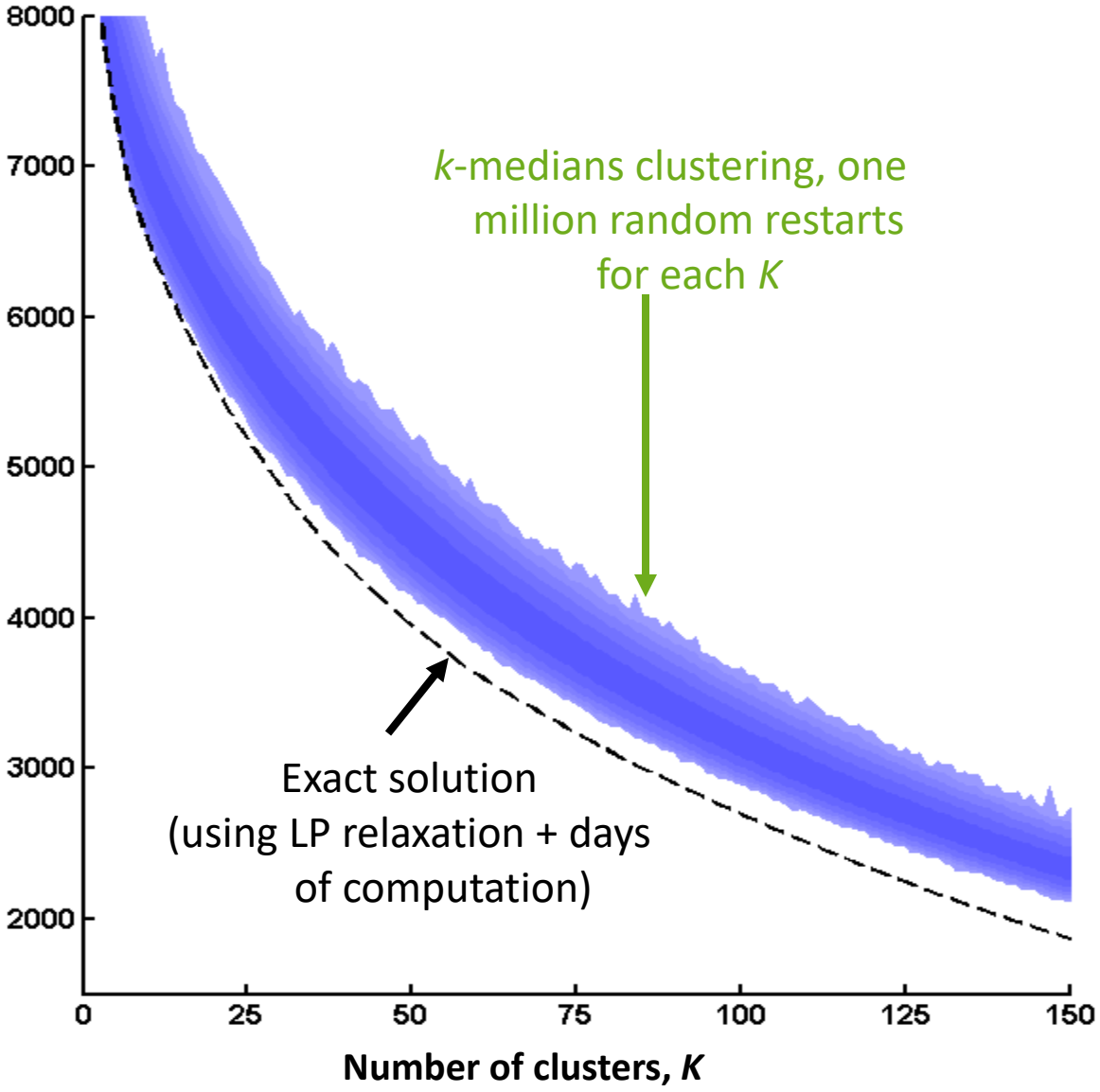
Examplo: Olivetti face images



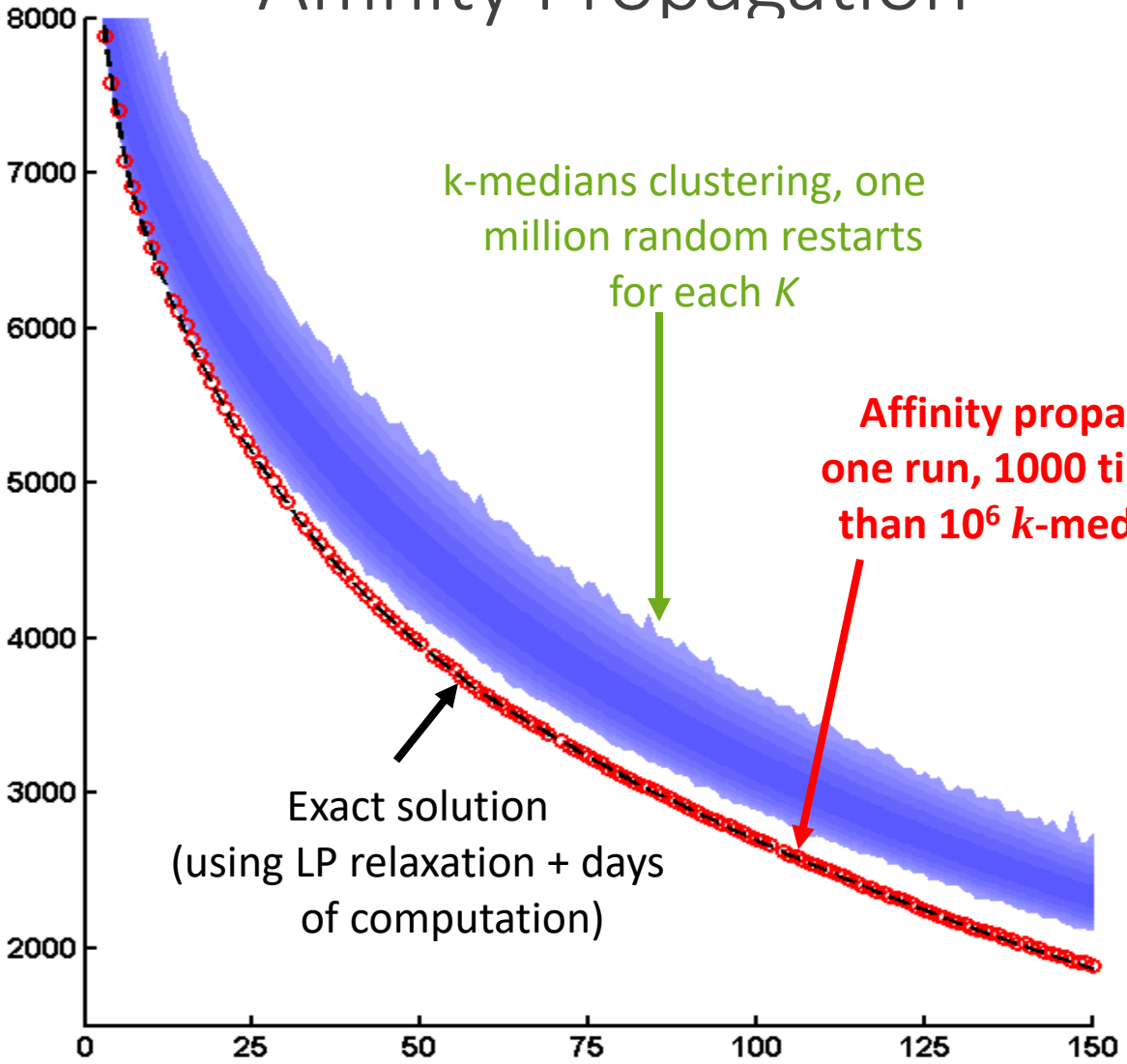
400 Imagens em tons de cinza, 64×64 pixels, de 40 people

- Similaridade based na soma dos distâncias quadráticas, usando uma janela central de 50×50 pixels
- Problema pode de ser resolvido por força bruta (pequena dimensão)

Recall Olivetti faces: squared error achieved by 1 million runs of k-medians clustering



Olivetti faces: squared error achieved by Affinity Propagation



Mapas Auto-Organizáveis

O Self-Organizing Map (SOM), ou Mapas Auto-Organizáveis foram desenvolvidos por Kohonen a partir de 1982

Aprendizado não-supervisionado, diferente de todas as redes neurais artificiais desenvolvidas até então

Possui forte inspiração neurofisiológica

É baseado em Aprendizagem Competitiva

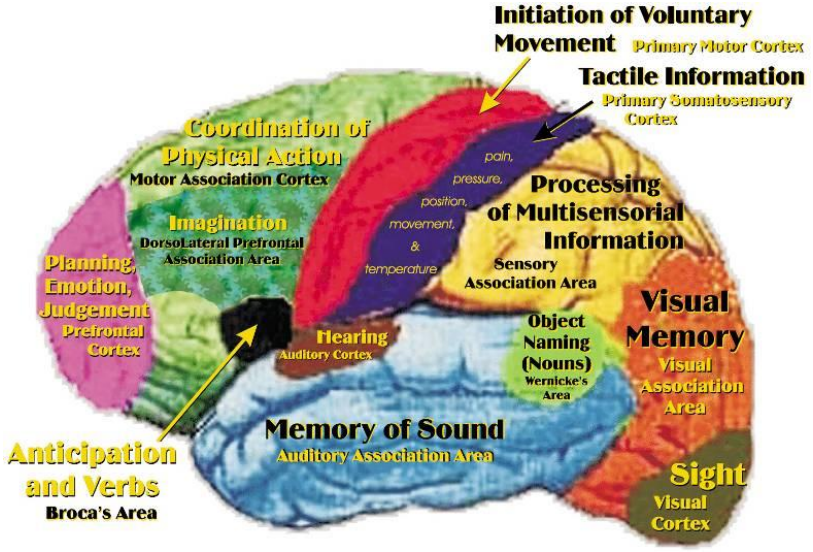
Inspiração Neurofisiológica

Observação de imagens

- Ressonância magnética (MRI)
- Tomografia Computadorizada (CT)

Diferentes estímulos geram

- Regiões de excitação
- Organização topográfica

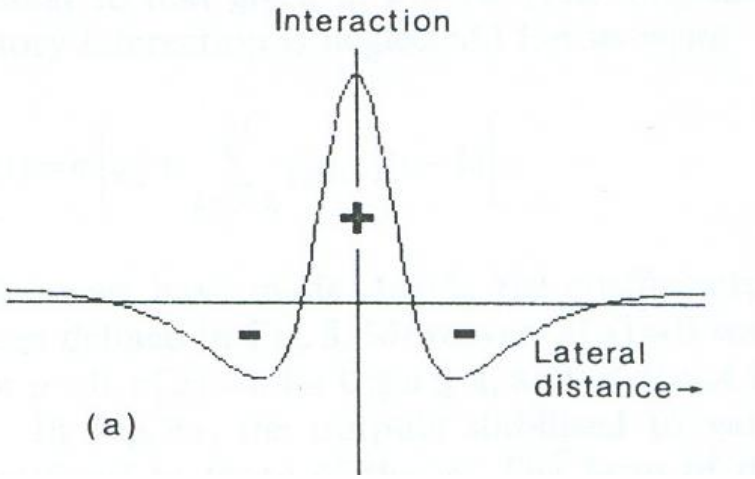


Inspiração Neurofisiológica

Quando um neurônio é excitado, ao redor uma área entre 50 e 100 µm também sofre excitação

Ao redor, uma área sofre inibição para impedir a propagação do sinal a áreas não relacionadas

- **A figura ilustra a iteração lateral entre os neurônios**



Aprendizagem Competitiva

Neurônios de saída da RNA competem entre si para se tornar ativos

Apenas um neurônio de saída está ativo em um determinado instante

Três elementos básicos:

- Neurônios com mesma estrutura, diferente pelos pesos, de forma que tenham respostas diferentes a uma entrada
- Um limite imposto sobre a força de cada neurônio
- Mecanismo de competição entre neurônios, de forma que um neurônio é vencedor em um dado instante.

Em cada momento o neurônio vencedor:

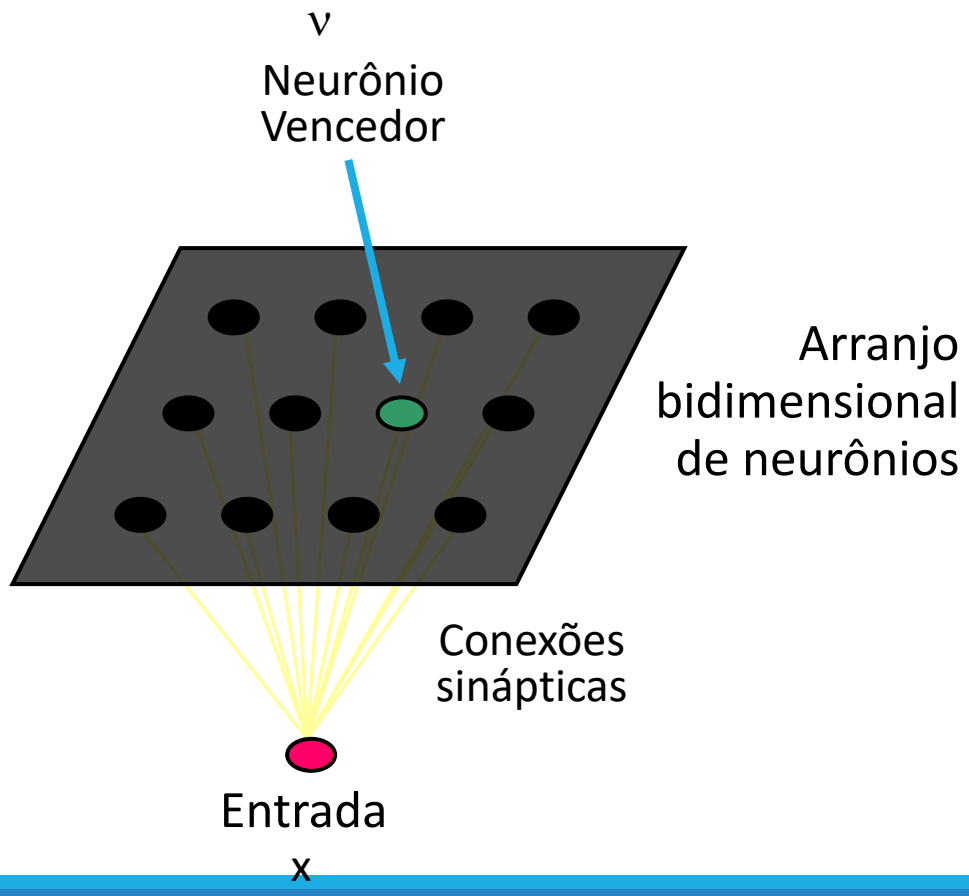
- aprende a se especializar em agrupamentos de padrões similares
- tornam-se detectores de características para classes diferentes de padrões de entrada

O número de unidades de entrada define a dimensionalidade dos dados

Modelo de Kohonen

O modelo de Kohonen

- Produz um mapeamento topológico
- Transforma um padrão de dimensão arbitrária em um mapa discreto uni- ou bidimensional
- Preserva a relação de vizinhança entre os neurônios



Aprendizagem Competitiva (Exemplo / Interpretação)

2 entradas (espaço 2D)

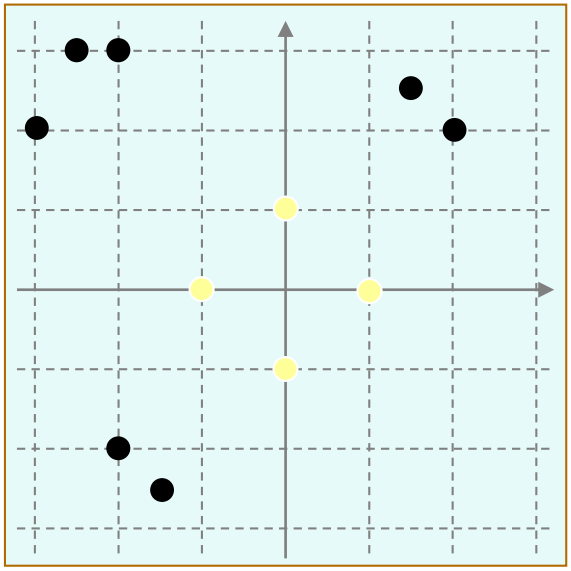
- 7 padrões de entrada

4 neurônios de saída

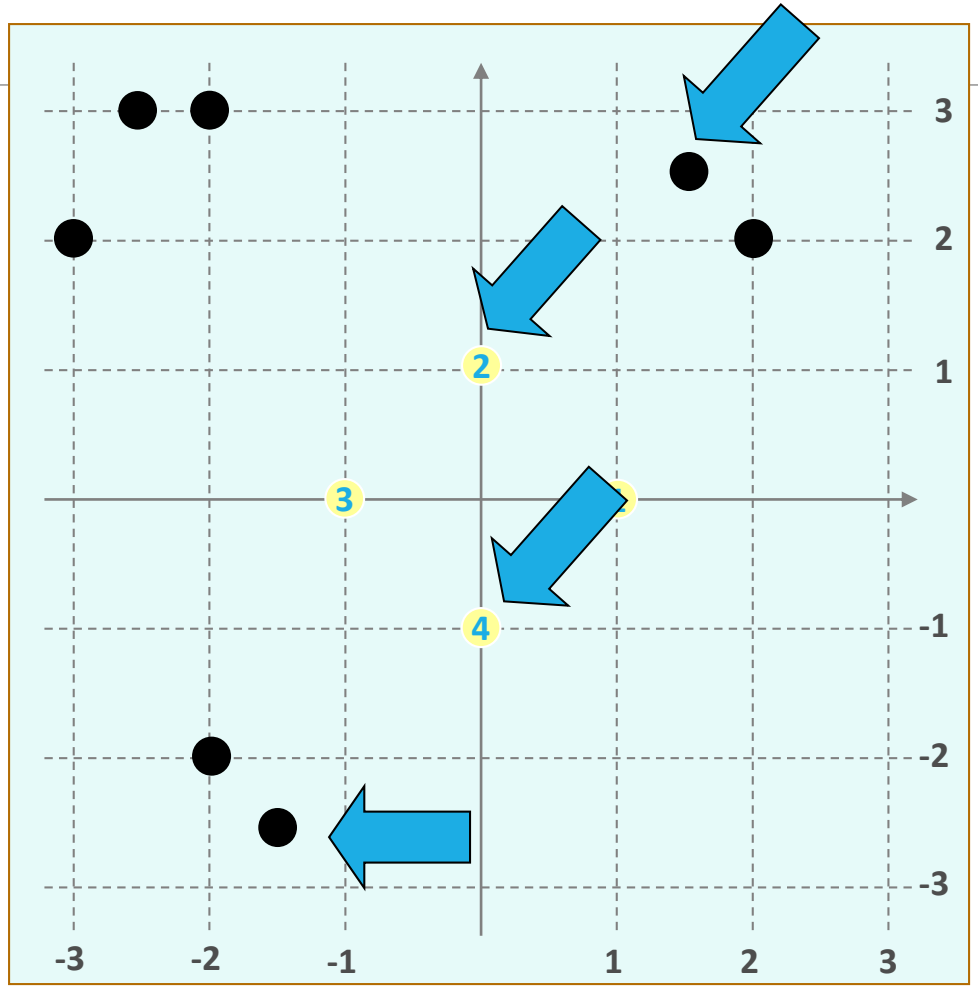
$\alpha = 0.5$

7 iterações

Na Figura, os pontos pretos representam as entradas e os amarelos os vetores dos pesos sinápticos dos 4 neurônios de saída



Estado Inicial da Rede



Entrada aleatória

Neurônio vencedor

Aprendizado

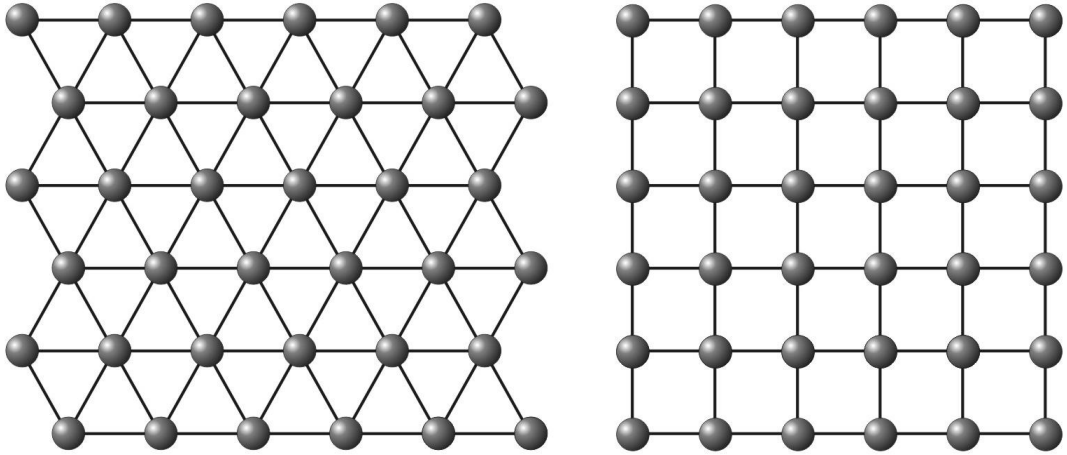
como α = 0.5, o deslocamento é referente à metade da distância

Simulação das iterações da aprendizagem competitiva

Mapa Topológico

No caso bidimensional, dois tipos de grade são possíveis: hexagonal ou retangular.

- Na hexagonal, cada neurônio possui 6 vizinhos diretos
- Na retangular, cada neurônio possui 4 vizinhos diretos



Validação dos Agrupamentos

No caso de classificação da padrões, é possível definir facilmente métricas que quantifiquem a performance do Sistema, uma vez que sabemos qual o resultado esperado (possuímos os rótulos dos padrões)

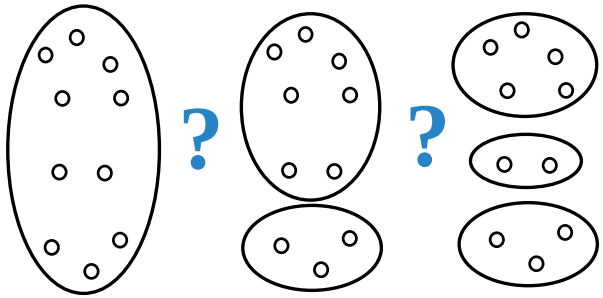
- Acurácia
- Precisão
- Recall

No caso de clusterização, é necessário definir outras métricas que não se baseiem nos rótulos

Métricas de validação dos clusters

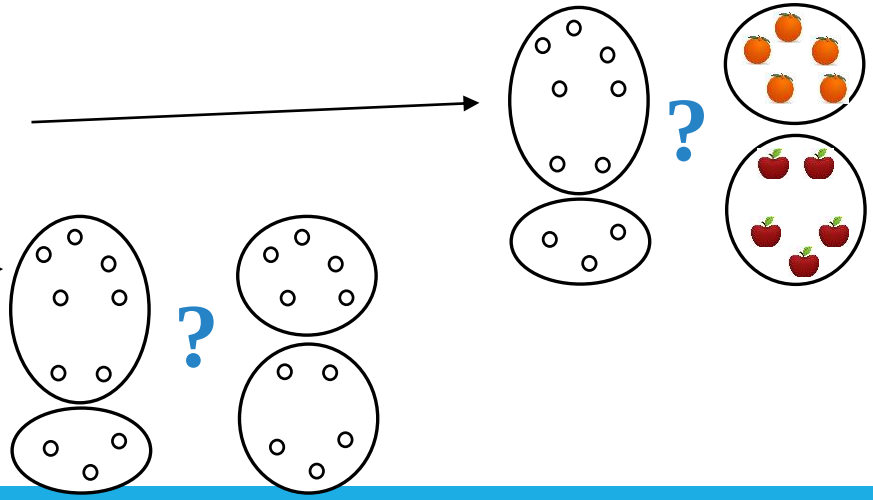
Índices internos

- Validação sem informação externa
- Com diferentes números de clusters
- Solução para o número de clusters



Índices externos

- Validação com rótulos conhecidos
- Comparação entre dois clusters



Índices Internos

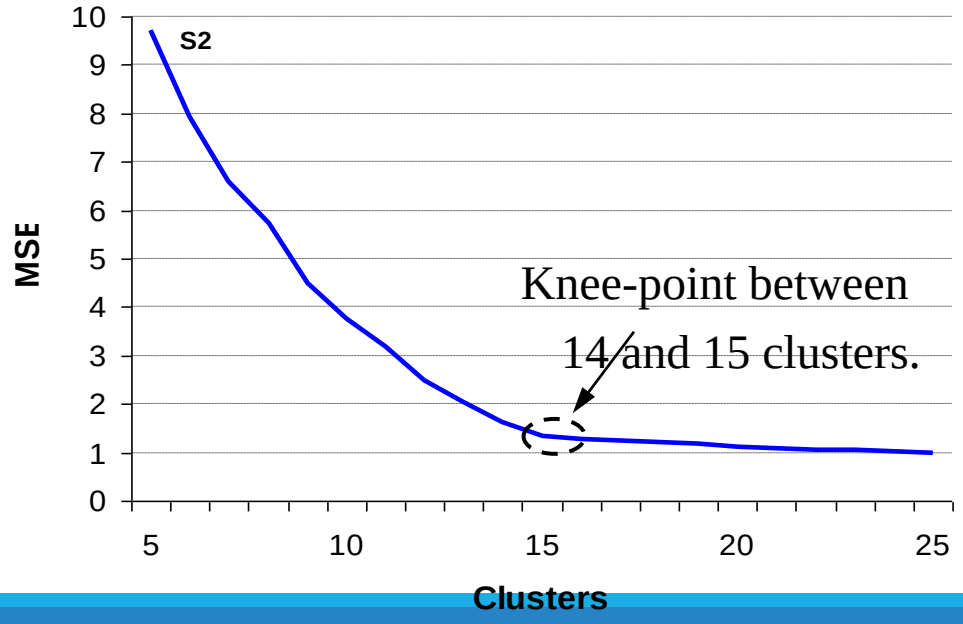
Dificilmente é possível contar com dados já classificados para avaliar os métodos de clusterização.

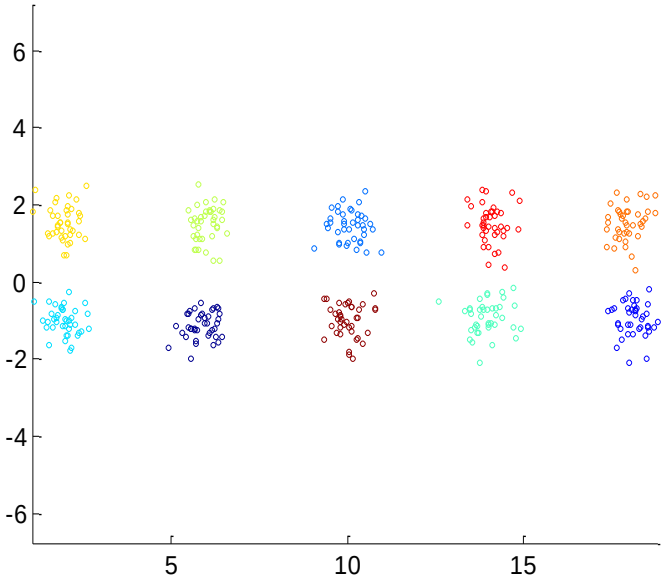
Em geral, a qualidade do agrupamento é avaliada por alguma métrica (índice) que reflete a noção de um "agrupamento coeso" ou "bem definido"

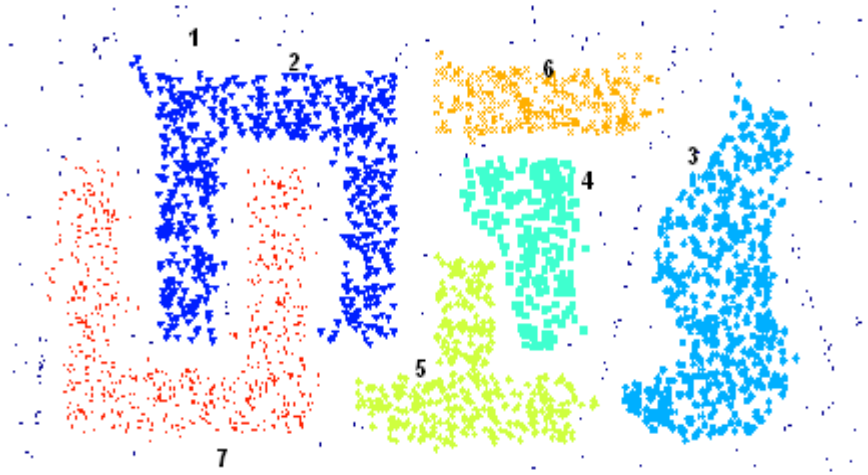
- Variância dos dados dentro de um mesmo cluster e entre clusters distintos
- Método de distorção de taxa
- Razão F (F-ratio)
- Índice *Davies-Bouldin* (DBI)
- *Bayesian Information Criterion* (BIC)
- Coeficiente de Silhueta (*Silhouette Coefficient*)

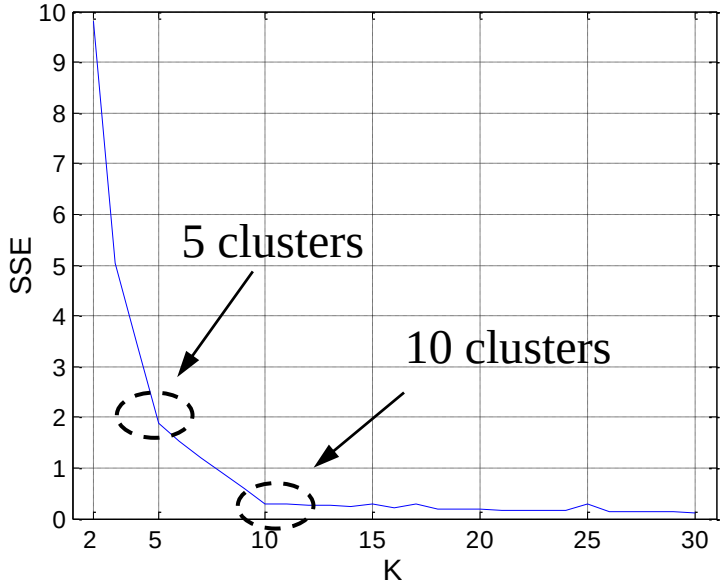
Soma dos erros quadráticos

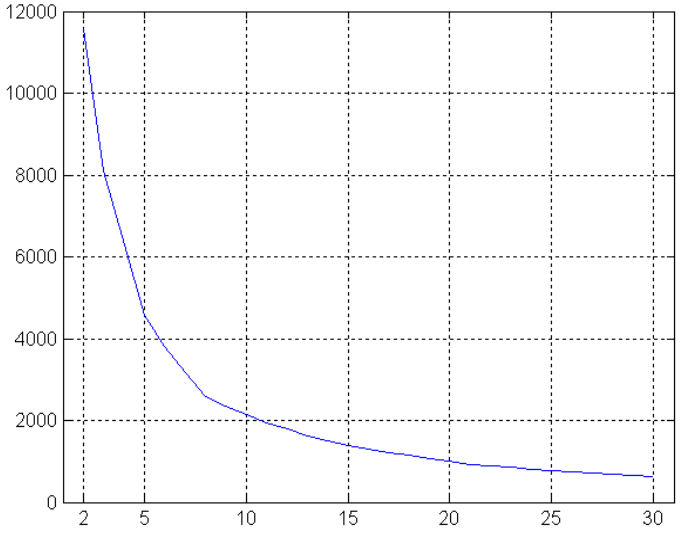
- Quanto maior o número de agrupamentos menor será o erro
- Como detectar o “joelho” para determiner o número de clusters ótimo?




6
4
2
0
-2
-4
-6
5
10
15


1
2
6
3
4
5
7

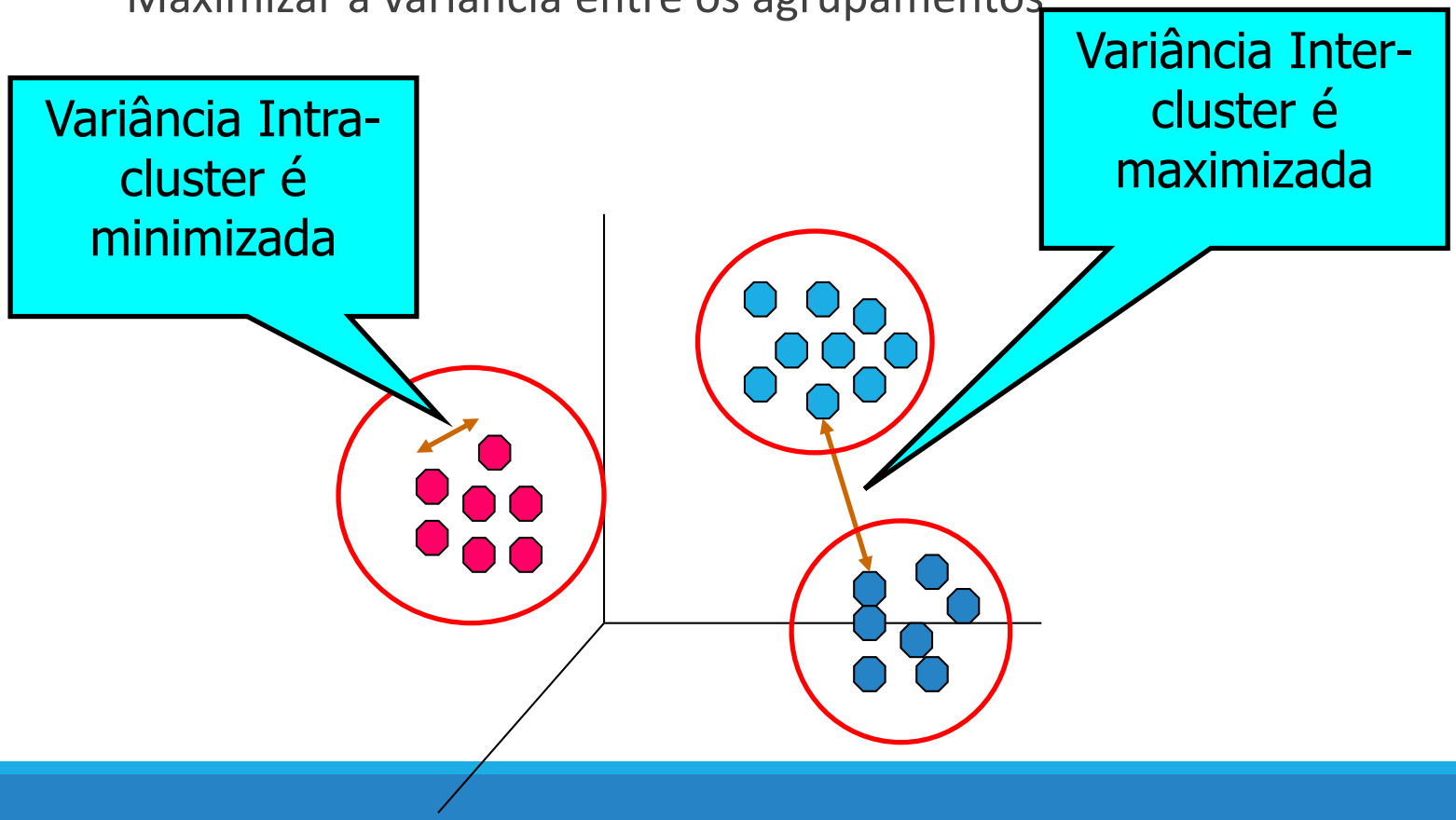

SSE
K
5 clusters
10 clusters
0
1
2
3
4
5
6
7
8
9
10
2
5
10
15
20
25
30


12000
10000
8000
6000
4000
2000
0
2
5
10
15
20
25
30

Agrupamentos “ideais”

Minimizar a variância de um agrupamento (*Total Squared Error* - TSE)

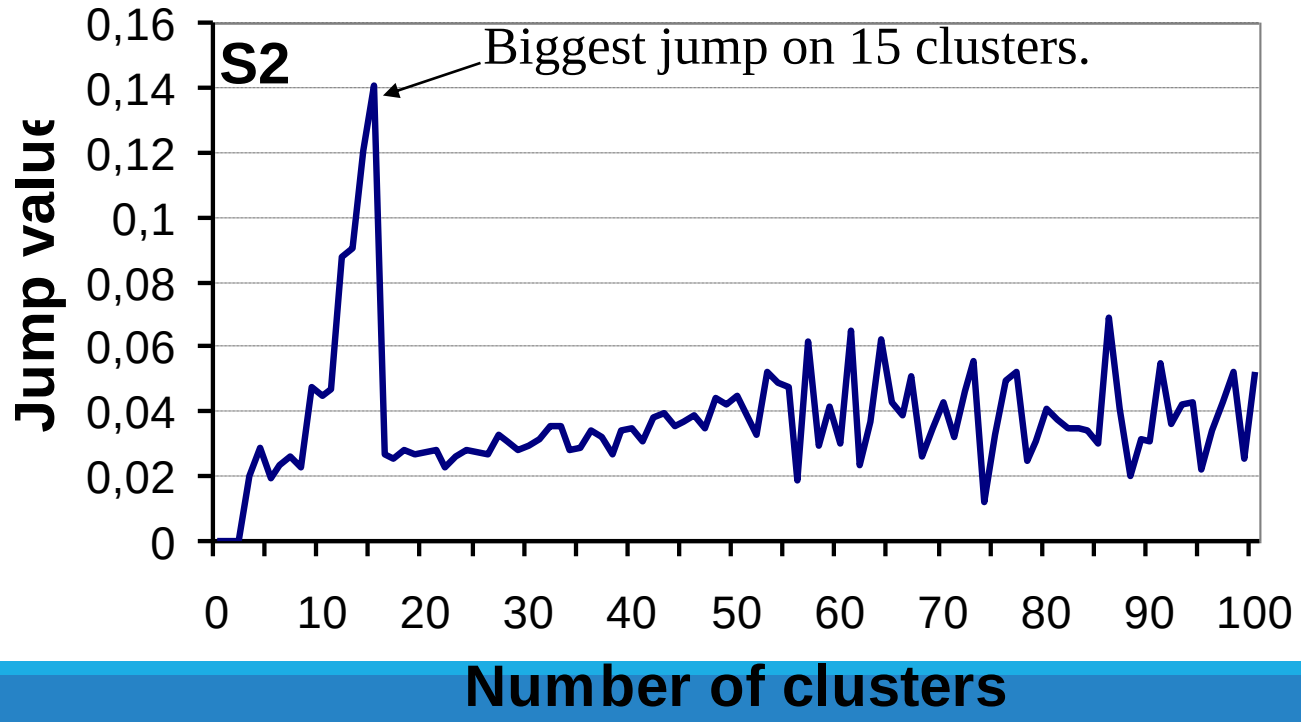
Maximizar a variância entre os agrupamentos



Abordagem “rate-distortion”

Nessa abordagem, avalia-se a variação da TSE em função do número de agrupamentos considerados. A função custo, em geral, é definida em termos de potências do valor de TSE, de acordo com

$$J(k) = TSE(k)^{-d/2} - TSE(k-1)^{-d/2}$$



Índice WB (WB-index/F-ratio)

A variância dentro de um mesmo cluster (within) é dada por

$$SSW(C,k) = \sum_{i=1}^{N} \left\| \mathbf{x}_i - \mathbf{c}_{p(i)} \right\|^2$$

onde N denota o número total de elementos. A variância entre clusters (between) é definida como

$$SSB(C,k) = \sum_{j=1}^{K} n_j \left\| \mathbf{c}_j - \bar{\mathbf{x}} \right\|^2$$

onde $\mathbf{c}_j$ e n_j denotam, respectivament, o centro e o número de elementos do cluster j, $\bar{\mathbf{x}}$ é o vetor médio dos dados. Assim, o índice WB é definido como a razão entre o SSE e SSB do resultado da clusterização

$$WB(C,k) = \frac{K \cdot SSW(C,k)}{SSB(C,k)}$$

Índices baseados no TSE

$\frac{SSW}{K}$ ---- Ball and Hall (1965)

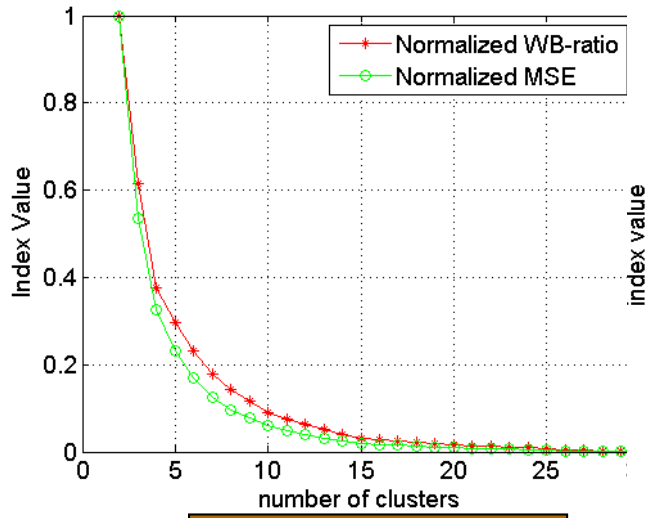
$K^2|W|$ ---- Marriot (1971)

$\frac{\frac{SSB}{K-1}}{\frac{SSE}{N-K}}$ ---- Calinski & Harabasz (1974)

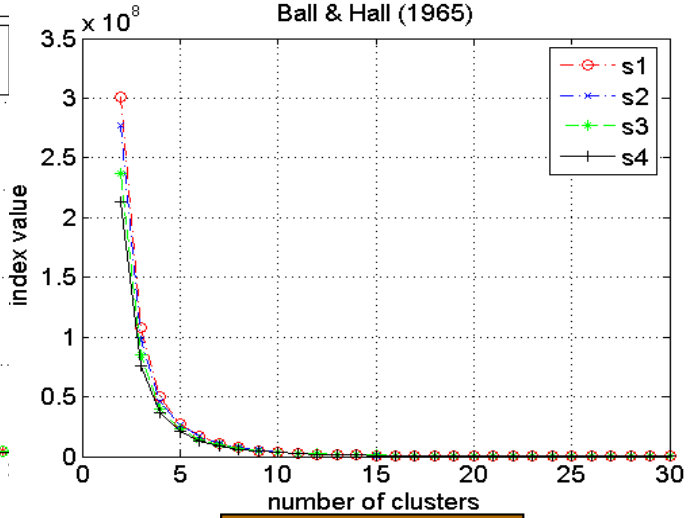
$\log(SSB/SSW)$ ---- Hartigan (1975)

$d \log\left(\sqrt{\frac{SSW}{dN^2}}\right) + \log K$ ---- Xu (1997)

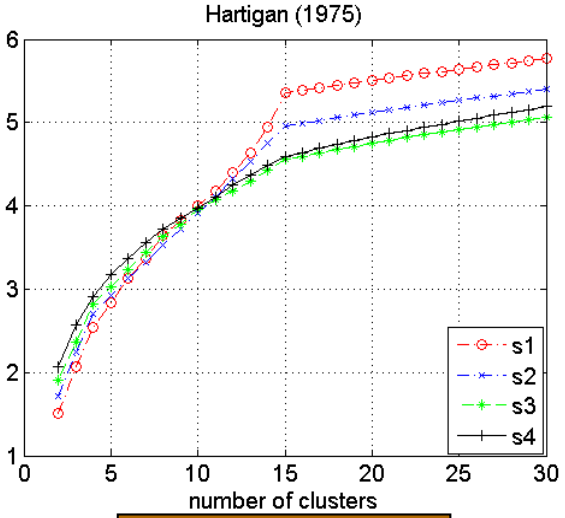
Para *d* = dimensões; *N* = número de dados; *K* = número de clusters



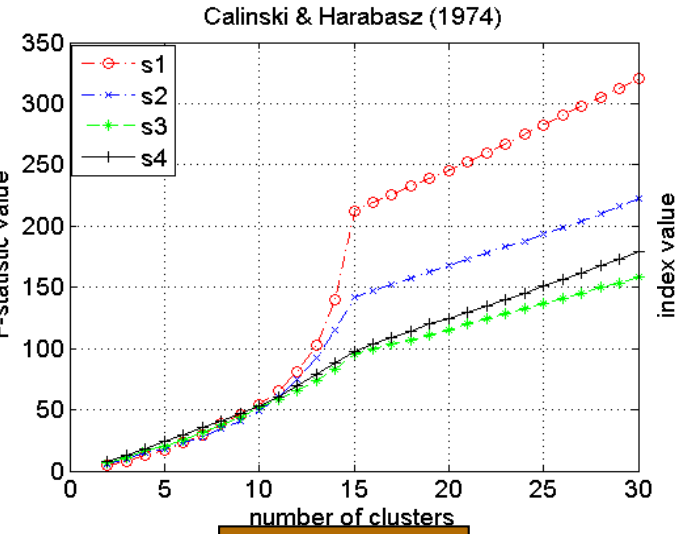
SSW / SSB & MSE



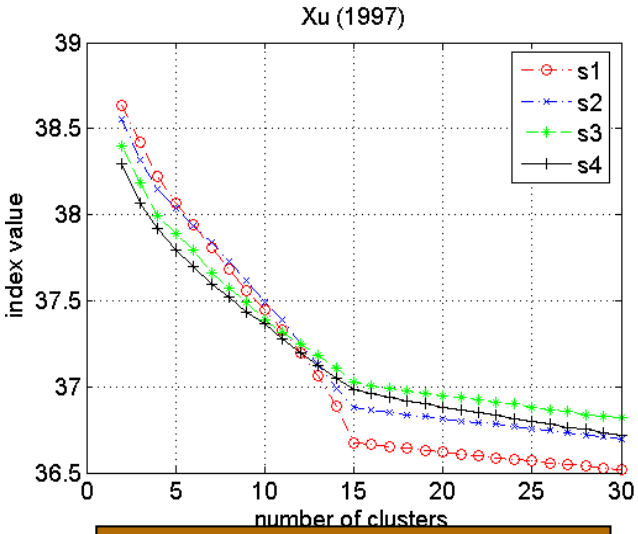
SSW / m



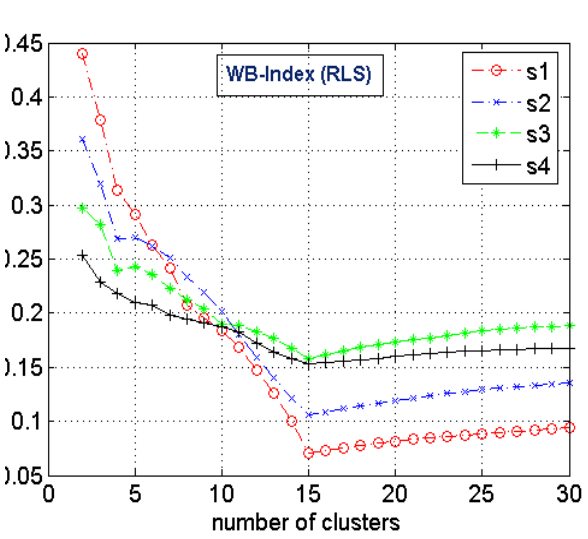
log(SSB/SSW)



$$\frac{SSB/m-1}{SSW/n-m}$$



$$d\log(\sqrt{SSW/(dn^2)})+\log(m)$$



m* SSW/SSB

Davies-Bouldin index (DBI)

O índice também avalia a dispersão dos dados de um cluster a distância entre os clusters formados. Matematicamente, seja a medida de dispersão de um cluster dada por

$$S_i = \left(\frac{1}{n_i} \sum_{j=1}^{n_i} \left| \mathbf{x}_j - \mathbf{a}_j \right|^p \right)^{1/p}$$

onde n_i é o número de elementos do cluster *i* e a_j o centroide do cluster *j,* para um valor arbitrário de *p*; seja M_{ij} uma medida de separação entre os clusters i e j, dada por

$$M_{ij} = \left\| \mathbf{a}_i - \mathbf{a}_j \right\|_p$$

Onde $\|\cdot\|_p$ denota a norma *L-p* de vetores.

Davies-Bouldin index (DBI)

Define-se

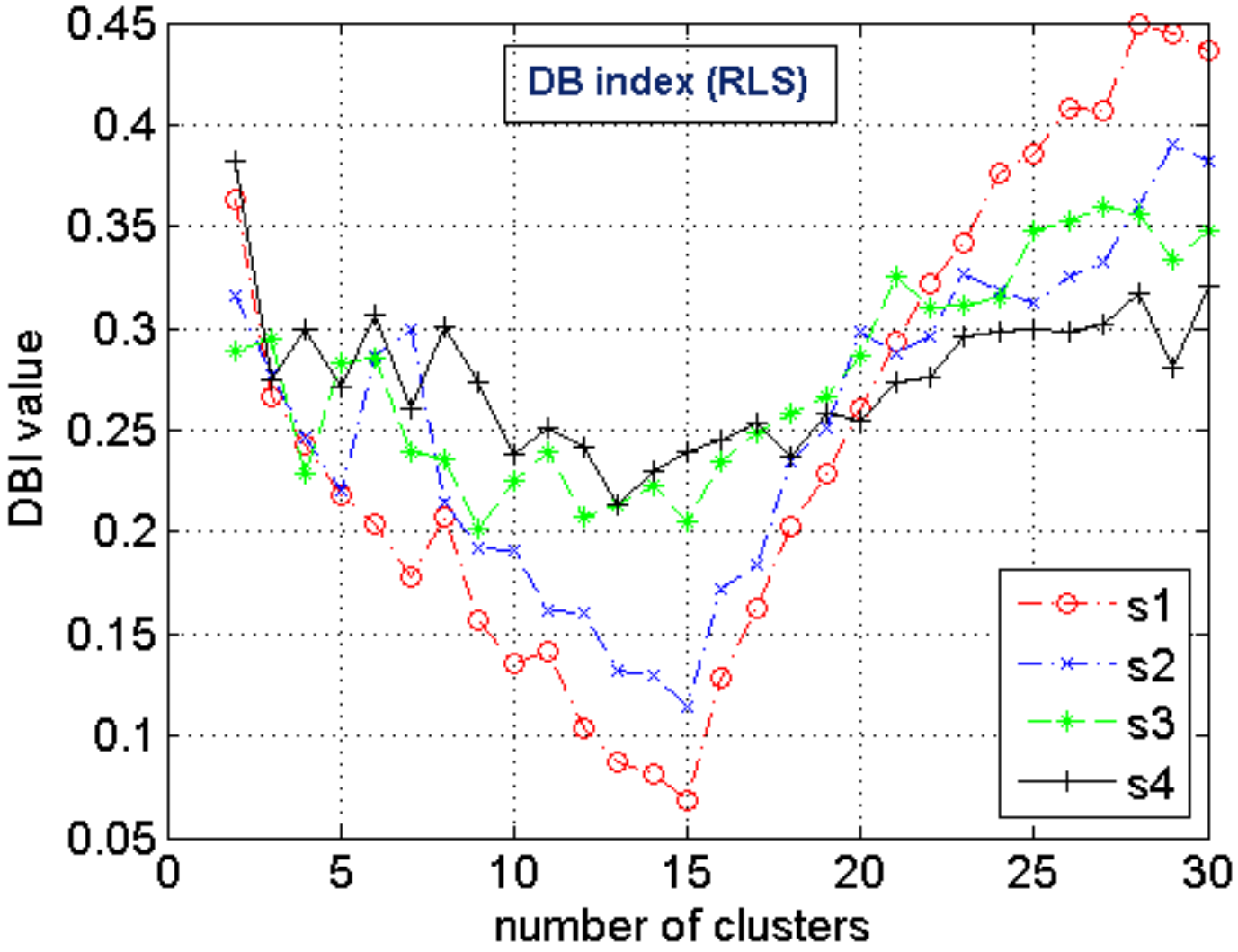
$$R_{ij} = \frac{S_i + S_j}{M_{ij}}$$

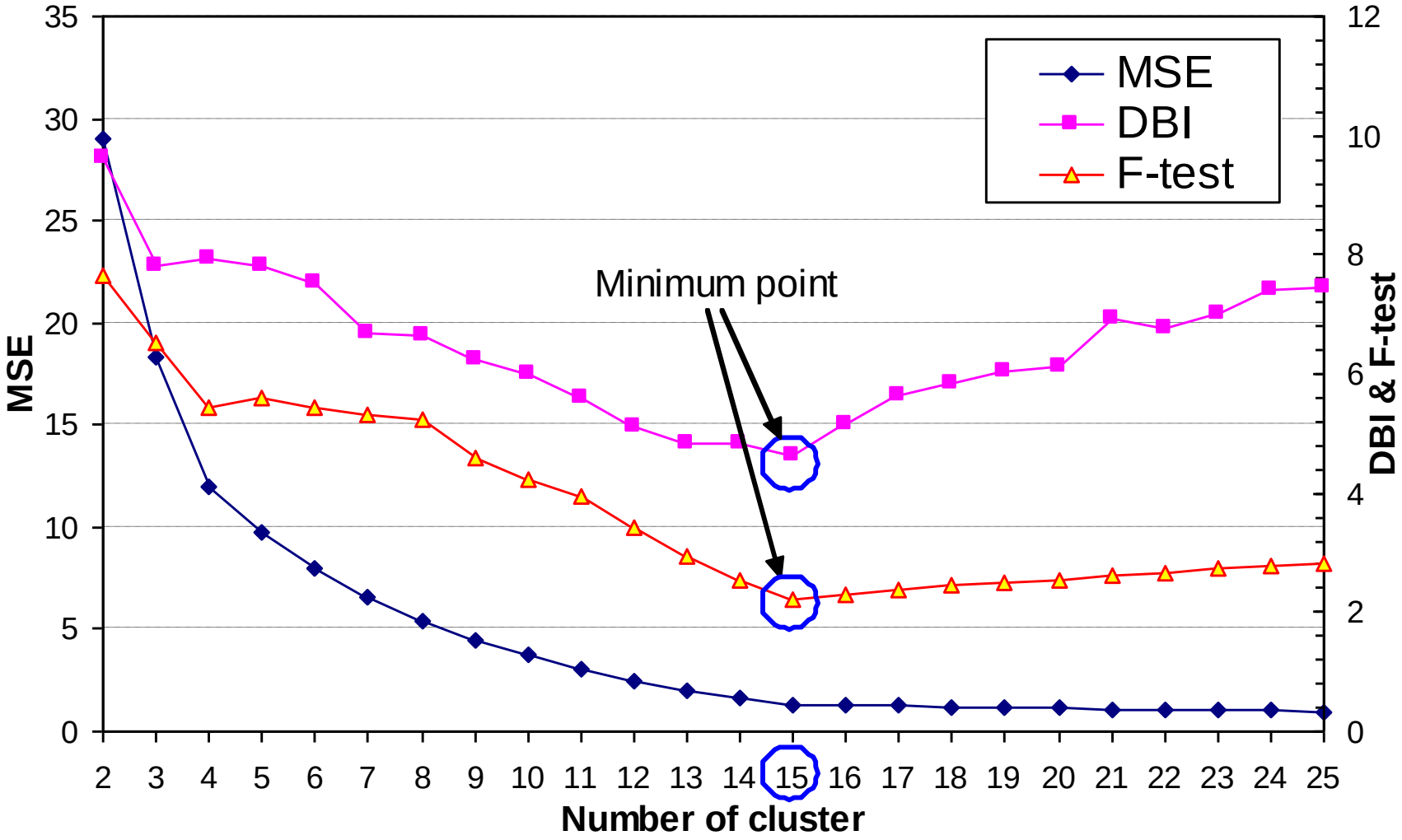
como medida da qualidade da clusterização entre os agrupamentos *i* e *j,* e

$$D_i = \max_{i \neq j} R_{ij}$$

Assim, o DBI é calculado como

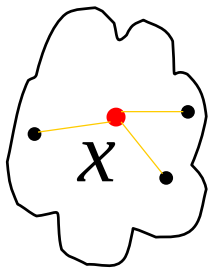
$$DBI = \frac{1}{K} \sum_{i=1}^{K} D_i$$


DB index (RLS)
DBI value
number of clusters
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
5
10
15
20
25
30
s1
s2
s3
s4


MSE
DBI
F-test
Minimum point
MSE
DBI & F-test
Number of cluster
0
5
10
15
20
25
30
35
0
2
4
6
8
10
12
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

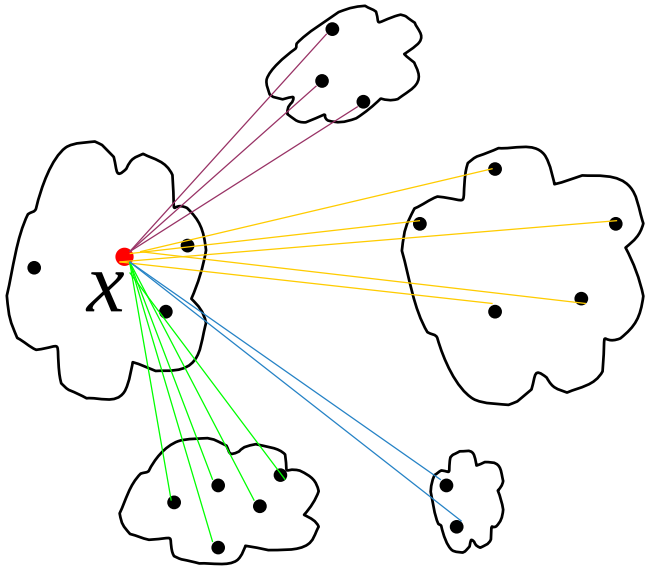
Coeficiente de Silhueta

Seguindo a mesma ideia explorada nas medidas anteriores, o coeficiente de silhueta busca medir a coesão de objetos em um mesmo cluster (proximidade dos elementos do cluster) e a separação dos clusters.



Coesão

$a(\mathbf{x})$



Separação

$b(\mathbf{x})$

Coeficiente de Silhueta

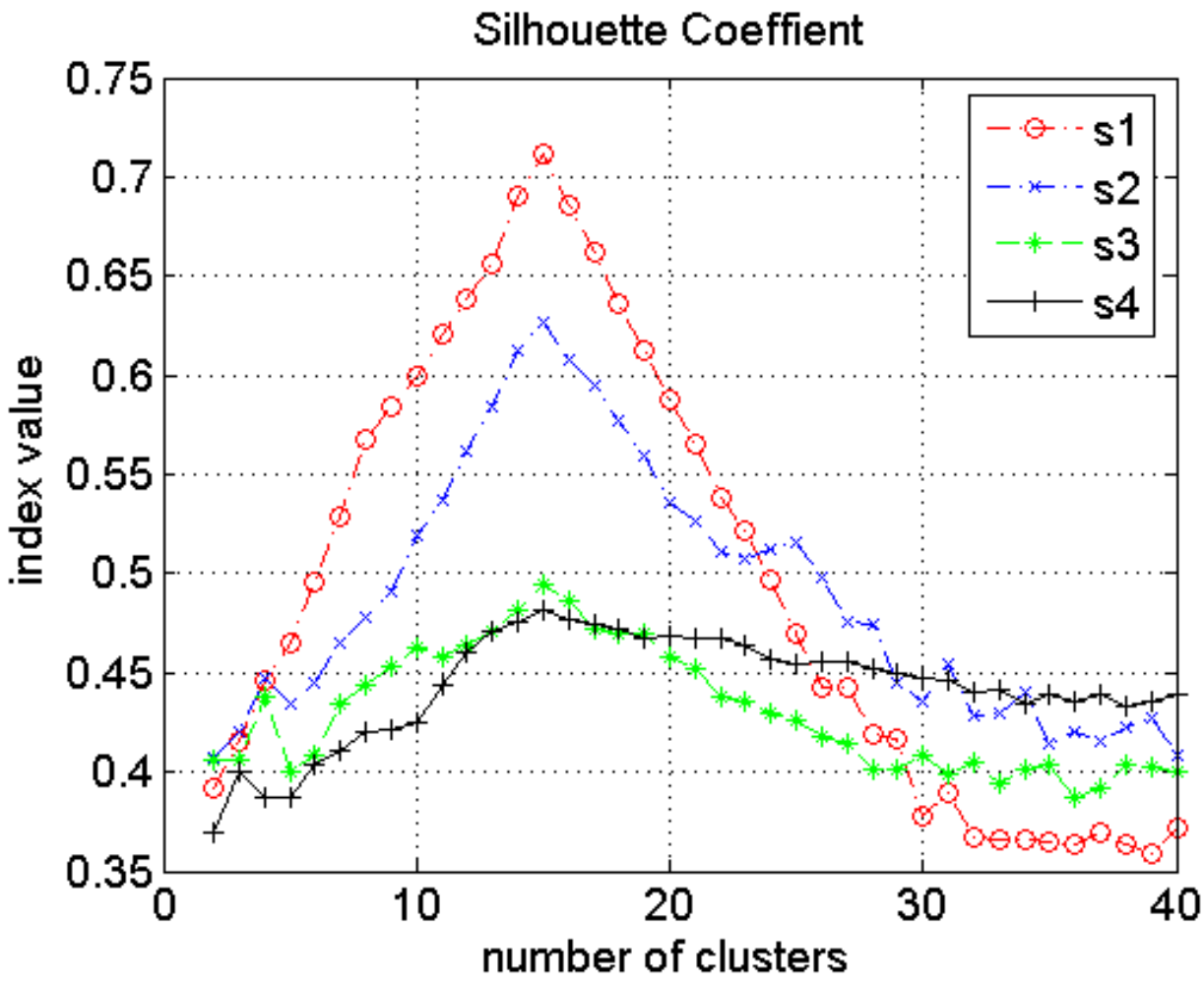
Coesão, $a(\mathbf{x}_i)$: distância média entre o elemento $\mathbf{x}_i$ e todos os demais vetores de um mesmo cluster

Separação, $b(\mathbf{x}_i)$: menor distância média entre $\mathbf{x}_i$ e vetores de outro cluster

O coeficiente de silhueta

$$s(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{\max\{a(\mathbf{x}_i), b(\mathbf{x}_i)\}}$$

possui valores entre -1 e 1, sendo -1 uma clusterização ruim, e 1 uma boa clusterização (0, indiferente)


Silhouette Coeffient
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
index value
0
10
20
30
40
number of clusters
s1
s2
s3
s4

Bayesian Information Criterion (BIC)

Suponha que os dados observados foram gerados por uma distribuição, desconhecida, $g(\mathbf{y})$. Um modelo estatístico para os dados observados é definido pela distribuição

$$p(\mathbf{y}|\boldsymbol{\theta}_k)$$

onde $\boldsymbol{\theta}_k$ denota o vetor de parâmetros do modelo M_k. Suponha que tenhamos diferentes modelos "candidatos" , $M_{k1}, M_{k2}, \ldots, M_{kL}$, para explicar os dados observados, cada um com um número arbitrário de parâmetros. O critério BIC foi proposto para indicar qual dos modelos candidatos "melhor aproxima" a distribuição verdadeira $g(\mathrm{y})$.

Sua derivação considera a abordagem Bayesiana, assumindo uma distribuição a priori para os modelos, $M_k \sim \pi(k)$. Assim, seja $g(\boldsymbol{\theta_k}|k)$ a distribuição a priori dos parâmetros dado o modelo M_k. Nesse caso, utilizando o teorema de Bayes, podemos escrever que

$$p(k, \boldsymbol{\theta}_k|\mathbf{y}) = \frac{\pi(k)g(\boldsymbol{\theta}_k|k)p(\mathbf{y}|\boldsymbol{\theta}_k)}{p(\mathbf{y})}$$

Bayesian Information Criterion (BIC)

A partir de $p(k, \boldsymbol{\theta}_k|\mathbf{y})$ pode-se obter a distribuição a posteriori do modelo M_k, i.e.

$$p(k|\mathbf{y}) = \int_{\boldsymbol{\theta}_k} \frac{\pi(k)g(\boldsymbol{\theta}_k|k)p(\mathbf{y}|\boldsymbol{\theta}_k)}{p(\mathbf{y})} d\boldsymbol{\theta}_k$$

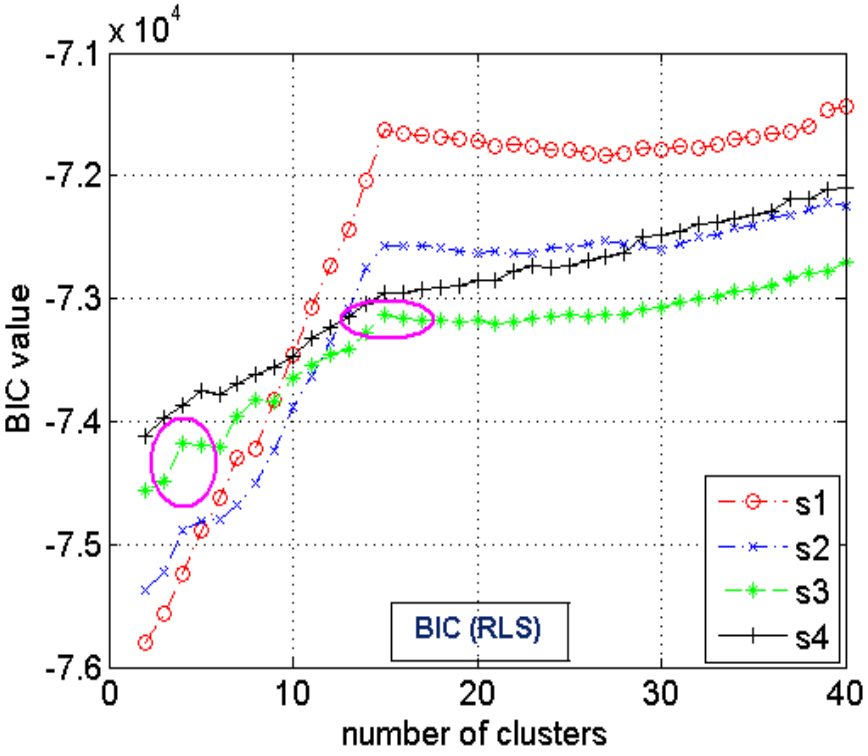
Note que o cálculo envolve a função de verossimilhança dos dados $L(\boldsymbol{\theta}_k) = p(\mathbf{y}|\boldsymbol{\theta}_k)$.

Utilizando algumas aproximações, mostra-se que $p(k|\mathbf{y})$ pode ser descrita por

$$p(\mathbf{y}|\boldsymbol{\theta}_k) = BIC = -2 \ln L(\boldsymbol{\theta}_k) + k \ln n$$

onde k representa a ordem do modelo e n o número de dados observados

Original BIC = *F*(*m*)



***SD*(*m*) = *F*(*m-1*) + *F*(*m+1*) – 2·*F*(*m*)**

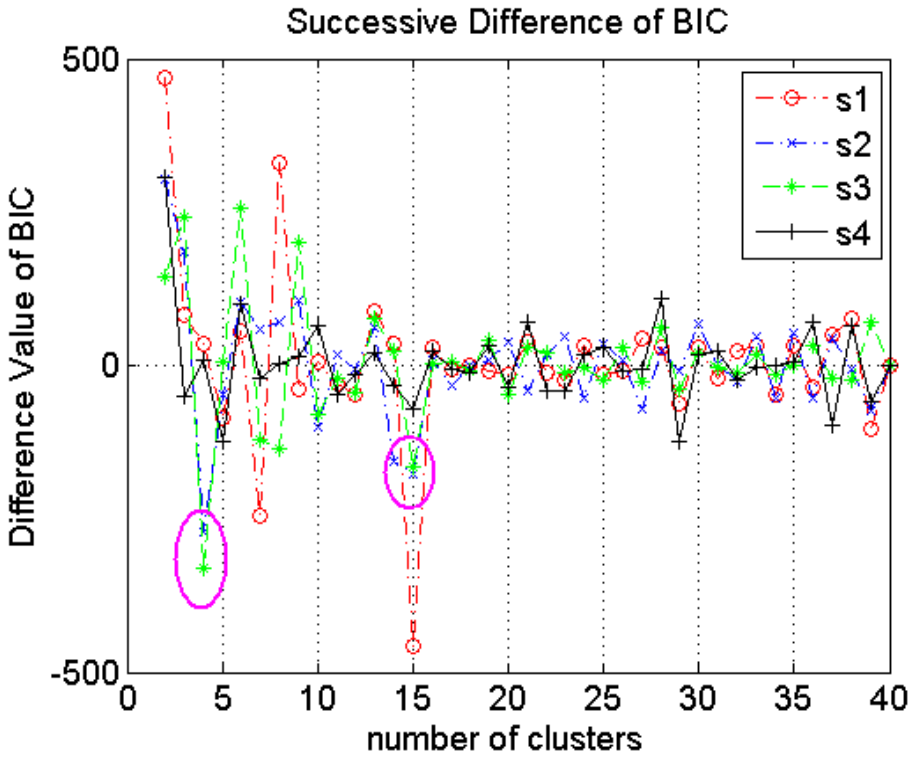
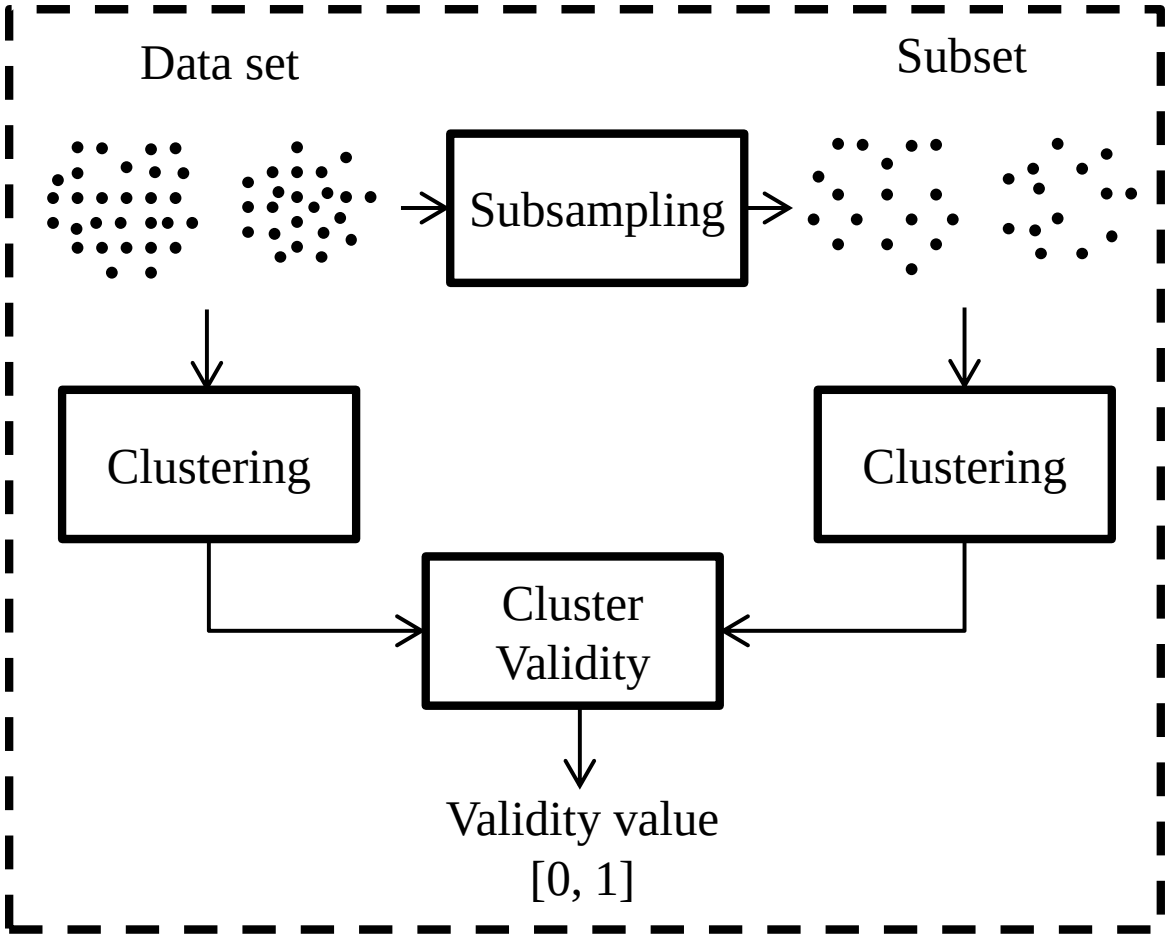


Table B.1: Formulas for internal indexes

Name	Formula
SSW	$SSW = \frac{1}{N}\sum_{i=1}^{N}\left\|x_i - C_{p_i}\right\|^2$
SSB	$SSB = \frac{2}{M(M-1)}\sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M}\left\|C_i - C_j\right\|^2$
Calinski-Harabasz index	$CH = \frac{SSB/(M-1)}{SSW(N-M)}$
Hartigan	$H_M = (\frac{SSW_M}{SSW_{M+1}} - 1)(N-M-1)$ $or: H_M = \log(SSB_M/SSW_M)$
Krzanowski-Lai index	$diff_M = (M-1)^{2/D}SSW_{M-1} - M^{2/D}SSW_M$ $KL_M = \left\|diff_M\right\| / \left\|diff_{M+1}\right\|$
Ball&Hall	$BH_M = SSW_M/M$
Xu-index	$Xu = D\log(\sqrt{SSW_M/(DN^2)}) + \log M$
Dunn's index	$Dunn = \sum_{i=1}^{M}\frac{\max(\left\|x_j - C_i\right\|^2)_{j\in C_i}}{}$
Davies&Bouldin index	$R_{ij} = \frac{S_i + S_j}{d_{ij}}, i \neq j$ $where: d_{ij} = \left\|C_i - C_j\right\|^2, S_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\left\|x_j - C_i\right\|^2$ $and, R_i = \max_{j=1,\ldots,M} R_{ij}, i = 1,\ldots,M$ $DBI = \frac{1}{M}\sum_{i=1}^{M} R_i$

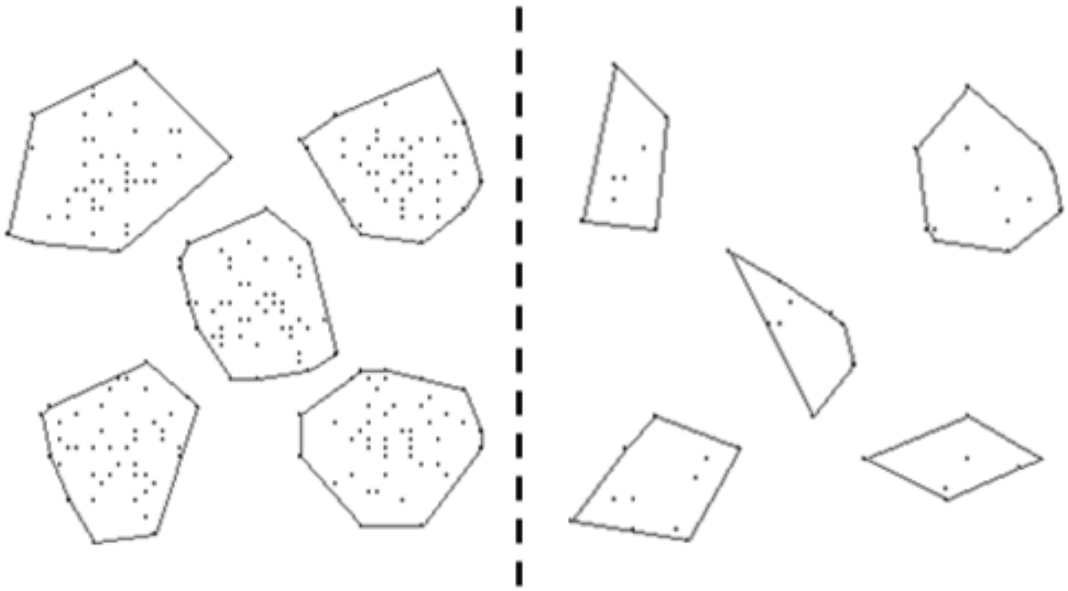
Silhouette Coefficients	$a(x_i) = \frac{1}{n_m - 1} \sum_{j=1, j \neq i}^{n_m} \left\| x_i - x_j \right\|^2_{x_i, x_j \in C_m}$ $b(x_i) = \min_t \{ \frac{1}{n_t} \sum_{j \in C_t} \left\| x_i - x_j \right\|^2 \}_{x_i \notin C_t}$ $s(x_i) = \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))}$ $SC = \frac{1}{N} \sum_{i=1}^{N} s(x_i)$ $b(x_i) = \min \{ \sum_{t \neq m} \left\| C_t - C_m \right\|^2 \}_{x_i \notin C_t} (SC'2008)$
RMSSTD	$RMSSTD = \frac{\sum_{\substack{k=1,\ldots,M \\ d=1,\ldots,D}} \sum_{i=1}^{n_{kd}} (x_i - \overline{x^d})^2}{\sum_{\substack{k=1,\ldots,M \\ d=1,\ldots,D}} (n_{kd} - 1)}$
R-square	$RS = \frac{SST - SSW}{SST} = \frac{\sum_{d \ 1,\ldots,D} \sum_{i=1}^{n_d} (x_i - \overline{x^d})^2 - \sum_{\substack{k=1,\ldots,M \\ d=1,\ldots,D}} \sum_{i=1}^{n_{kd}} (x_i - \overline{x^d})^2}{\sum_{d=1,\ldots,D} \sum_{i-1}^{n_d} (x_i - \overline{x^d})^2}$
Bayesian Information Criterion	$BIC = L * N - \frac{1}{2} M(D+1) \sum_{i=1}^{M} \log(n_i)$
Xie-Beni	$XB = \frac{\sum_{i-1}^{N} \sum_{k=1}^{M} u_{ik}^2 \left\| x_i - C_k \right\|^2}{N \min_{t \neq s} \{ \left\| C_t - C_s \right\|^2 \}}$
Partition Coefficient	$PC = \sum_{i=1}^{N} \sum_{k=1}^{M} u_{ik}^2 / N$
Partition Entropy	$PE = -(\sum_{i=1}^{N} \sum_{k=1}^{M} u_{ik} \log(u_{ik})) / N$

Validação baseada em estabilidade – Validação Cruzada



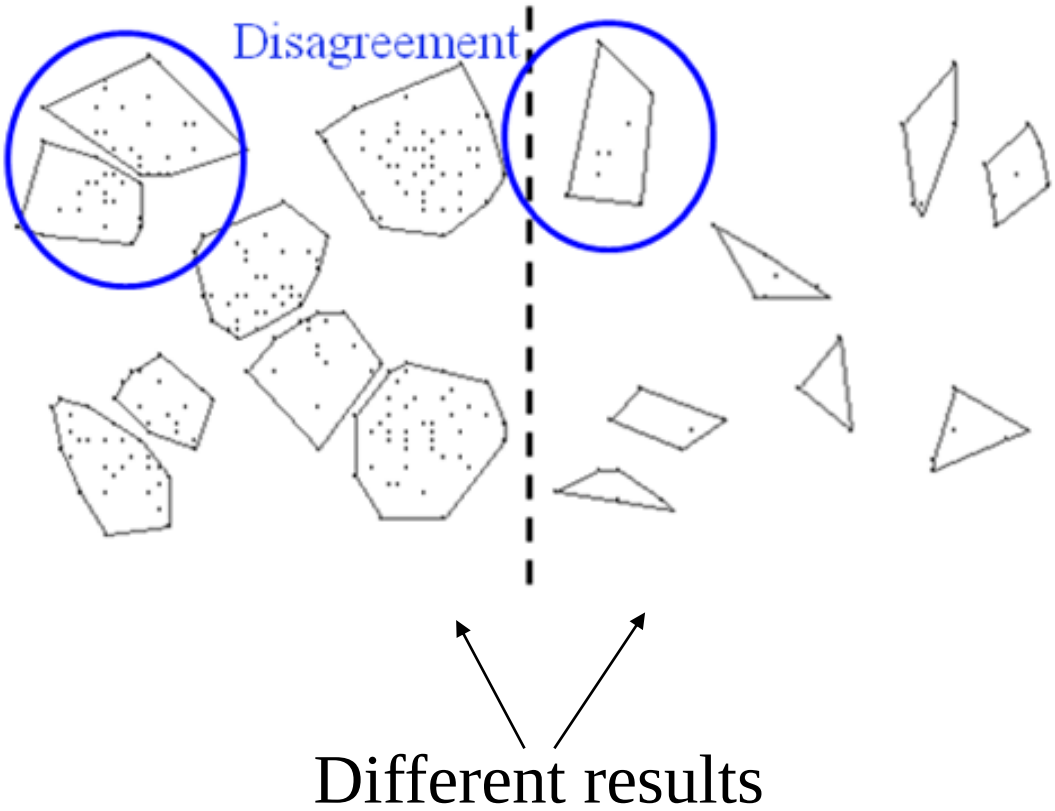
Ideia central consite em verificar se o resultado da clusterização permanece inalterado se for introduzida alguma aleatoriedade nos dados ou no algoritmo

Correct number of clusters: k=5



Same results

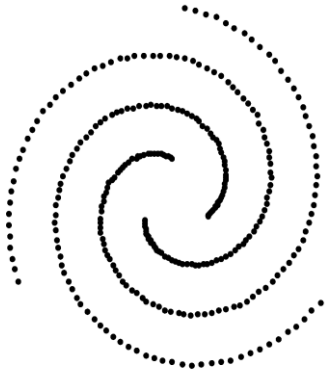
Incorrect number of clusters: k=8



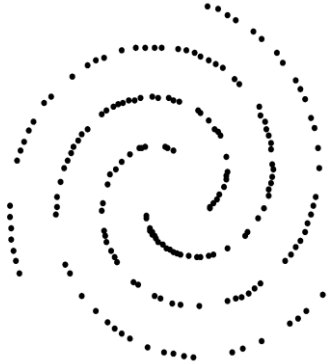
Sub-Amostragem

Uma forma de introduzir a aleatoriedade é por meio da sub-amostragem.

- Se a sub-amostragem for pequena, não há diferenças significativas nos resultados
- Se a sub-amostragem for grande, pode-se destruir completamente a estrutura dos agrupamentos
- Recomenda-se reter entre 20-40% dos dados

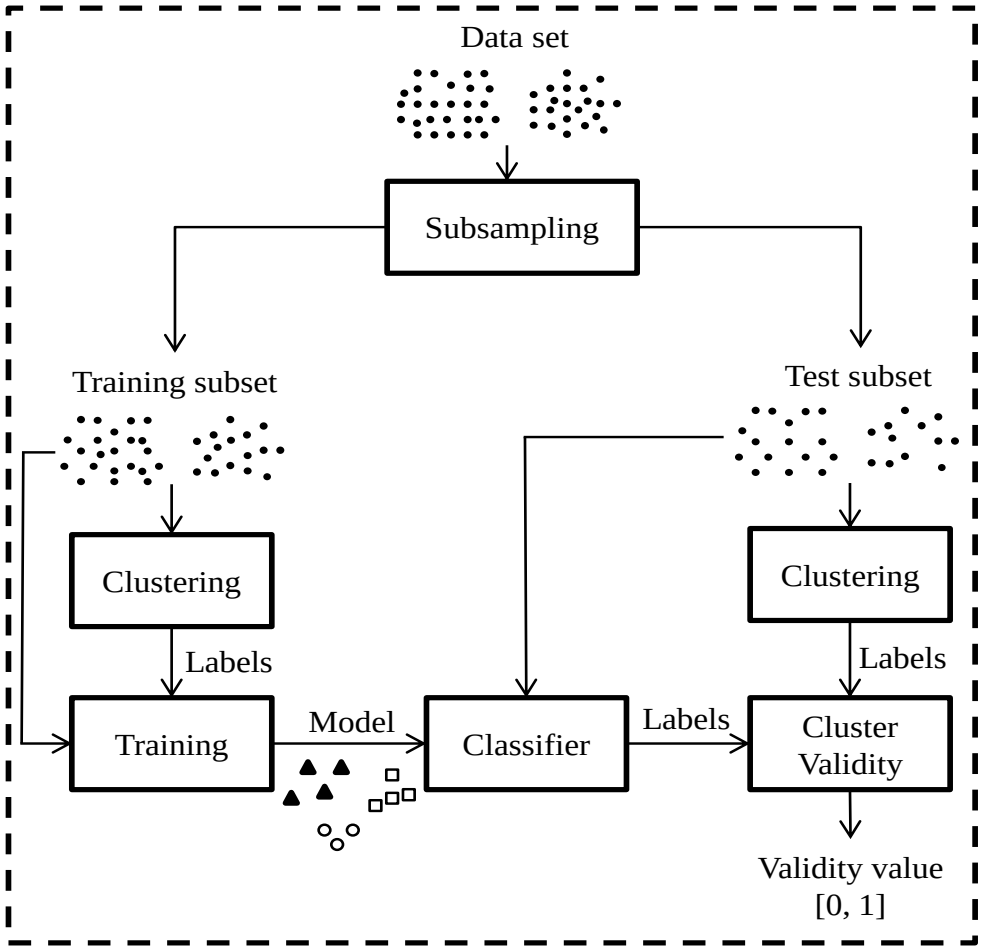


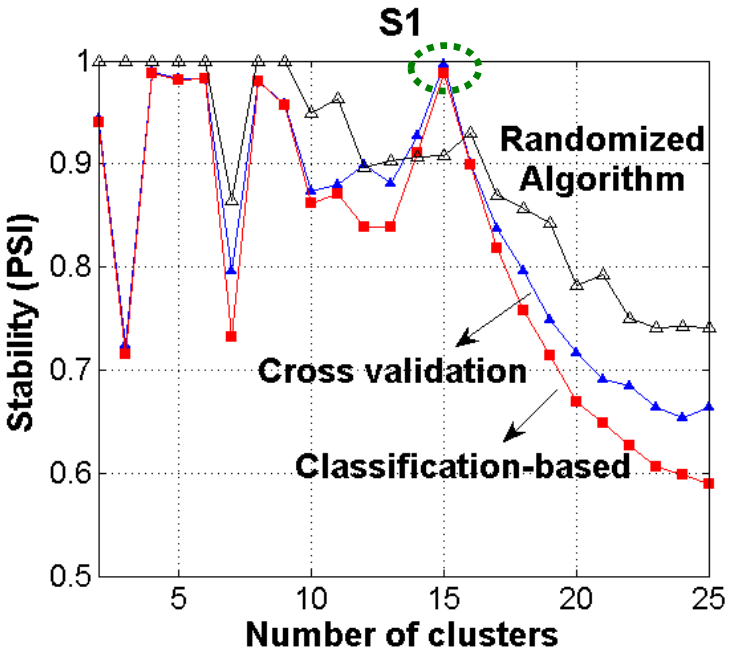
60%

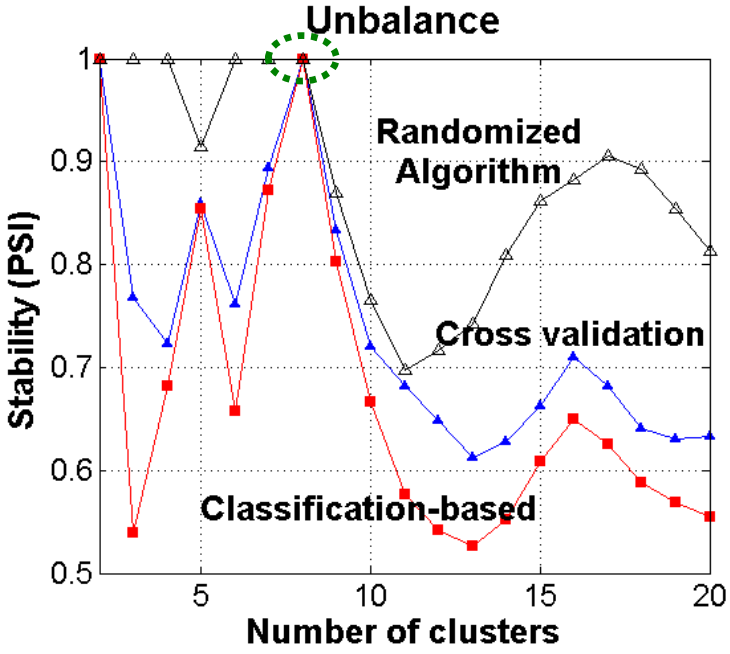


20%

Abordagem baseada em Classificação




S1
Randomized
Algorithm
Cross validation
Classification-based
Stability (PSI)
Number of clusters
1
0.9
0.8
0.7
0.6
0.5
5
10
15
20
25


Unbalance
Randomized
Algorithm
Cross validation
Classification-based
Stability (PSI)
Number of clusters
1
0.9
0.8
0.7
0.6
0.5
5
10
15
20

Índices Externos

Quando é possível contar com resultados externos ou informação sobre a classe correta para cada dado, utilizam-se algumas métricas para verificar o desempenho dos algoritmos de clusterização

- Índice de Rand e índice ajustado de Rand
- Informação Mútua
- Índices de similaridade entre clusters

Matriz de Contingência

Seja um conjunto de N elementos e dois agrupamentos (clusters) desses elementos, e.g., $X = \{X_1, X_2, \dots, X_r\}$ e $Y = \{Y_1, Y_2, \dots, Y_s\}$. A matriz de contingência mostra o número de elementos em comum entre X_i e Y_j

	Y_1	Y_2	$\cdots$	Y_s	**Soma**
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	$\boldsymbol{a_1}$
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	$\boldsymbol{a_2}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	...	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$\boldsymbol{a_r}$
	$\boldsymbol{b_1}$	$\boldsymbol{b_2}$	$\cdots$	$\boldsymbol{b_s}$	N

Medidas baseadas na contagem de pares

Número de pares de dados (n_{ij}= número de dados que pertencem ao agrupamento *i* em G e ao agrupamento j em P)

- Que pertencem à mesma classe tanto em P quanto em G

$$a = \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}(n_{ij} - 1)$$

- Que pertencem à mesma classe em P mas não em G

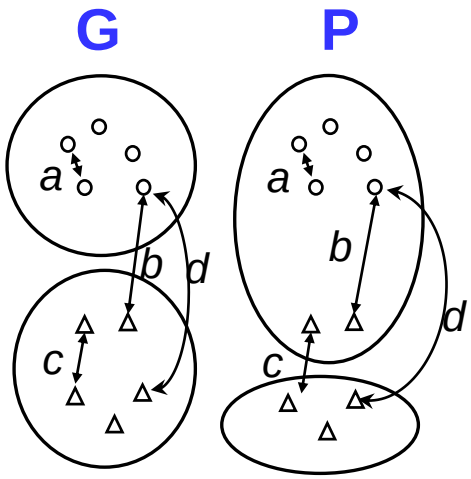
$$b = \frac{1}{2}(\sum_{j=1}^{K'} n_{jj}^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2)$$

- Que pertencem à mesma classe em G mas não em P

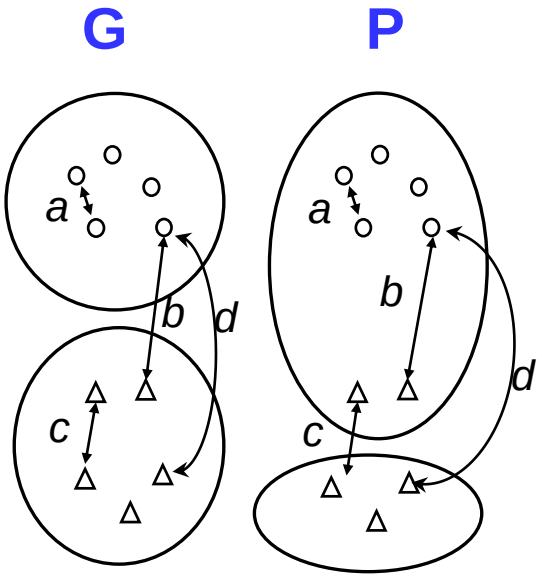
$$c = \frac{1}{2}(\sum_{i=1}^{K} n_{ii}^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2)$$

- Pertencem a classes distintas tanto em P quanto em G

$$d = \frac{1}{2}(N^2 + \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2 - (\sum_{i=1}^{K} n_{ii}^2 + \sum_{j=1}^{K'} n_{jj}^2))$$



Rand Index



$$RI(P,G) = \frac{a+d}{a+b+c+d}$$

Valores entre 0 e 1 (1 quando as partições são idênticas)

a = 20 b = 24

d = 72 c = 20

Rand index = (20+72) / (20+24+20+72) = 92/136 = **0.68**

Adjusted Rand index

O valor esperado do rand index para duas partições criadas aleatoriamente não é constante (seria interessante que para partições criadas aleatoriamente o índice apresentasse o seu menor valor)

Para corrigir esse problema foi proposto o índice ajustado de Rand, dado por

$$ARI = \frac{RI - E(RI)}{1 - E(RI)}$$

onde $E(RI)$ denota o valor esperado para o índice de Rand

Informação Mútua

Intuitivamente, pode ser entendida como a quantidade de informação que uma variável aleatória "carrega" sobre outra.

Matematicamente, é definida a partir da entropia de uma variável aleatória, $H(P) = \sum_i p(x_i) \log p(x_i)$. Intuitivamente, ela indica o grau de incerteza a respeito do valor da variável X.

A entropia condicional, $H(X|Y)$, também indica o grau de incerteza a respeito do valor da variável X, mas após termos conhecimento sobre o valor da variável Y, definida por $H(X|Y) = \sum_{i,j} p(x_i, y_i) \log \frac{p(x_i, y_i)}{p(y_i)}$

A informação mútua corresponde a

$$I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

e indica quanta informação foi obtida após a observação do valor de *Y.*

Table 1: External Cluster Validation Measures.

	Measure	Notation	Definition	Range
1	Entropy	E	$-\sum_i p_i(\sum_j \frac{p_{ij}}{p_i} \log \frac{p_{ij}}{p_i})$	$[0, \log K']$
2	Purity	P	$\sum_i p_i(\max_j \frac{p_{ij}}{p_j})$	(0,1]
3	F-measure	F	$\sum_j p_j \max_i [2\frac{p_{ij}}{p_i}\frac{p_{ij}}{p_j}/(\frac{p_{ij}}{p_i}+\frac{p_{ij}}{p_j})]$	(0,1]
4	Variation of Information	VI	$-\sum_i p_i \log p_i - \sum_j p_j \log p_j - 2\sum_i\sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j}$	$[0, 2\log\max(K, K')]$
5	Mutual Information	MI	$\sum_i\sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j}$	$(0, \log K']$
6	Rand statistic	R	$[\binom{n}{2} - \sum_i \binom{n_{i\cdot}}{2} - \sum_j \binom{n_{\cdot j}}{2} + 2\sum_{ij}\binom{n_{ij}}{2}]/\binom{n}{2}$	(0,1]
7	Jaccard coefficient	J	$\sum_{ij}\binom{n_{ij}}{2}/[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2} - \sum_{ij}\binom{n_{ij}}{2}]$	[0,1]
8	Fowlkes and Mallows index	FM	$\sum_{ij}\binom{n_{ij}}{2}/\sqrt{\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}}$	[0,1]
9	Hubert Γ statistic I	Γ	$\frac{\binom{n}{2}\sum_{ij}\binom{n_{ij}}{2} - \sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}}{\sqrt{\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}[\binom{n}{2} - \sum_i \binom{n_{i\cdot}}{2}][\binom{n}{2} - \sum_j \binom{n_{\cdot j}}{2}]}}$	(-1,1]
10	Hubert Γ statistic II	Γ'	$[\binom{n}{2} - 2\sum_i \binom{n_{i\cdot}}{2} - 2\sum_j \binom{n_{\cdot j}}{2} + 4\sum_{ij}\binom{n_{ij}}{2}]/\binom{n}{2}$	[0,1]
11	Minkowski score	MS	$\sqrt{\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2} - 2\sum_{ij}\binom{n_{ij}}{2}}/\sqrt{\sum_j \binom{n_{\cdot j}}{2}}$	$[0, +\infty)$
12	classification error	ε	$1 - \frac{1}{n}\max_\sigma \sum_j n_{\sigma(j),j}$	[0,1)
13	van Dongen criterion	VD	$(2n - \sum_i \max_j n_{ij} - \sum_j \max_i n_{ij})/2n$	[0, 1)
14	micro-average precision	MAP	$\sum_i p_i(\max_j \frac{p_{ij}}{p_i})$	(0,1]
15	Goodman-Kruskal coefficient	GK	$\sum_i p_i(1 - \max_j \frac{p_{ij}}{p_i})$	[0,1)
16	Mirkin metric	M	$\sum_i n_{i\cdot}^2 + \sum_j n_{\cdot j}^2 - 2\sum_i\sum_j n_{ij}^2$	$[0, 2\binom{n}{2})$

Note: $p_{ij} = n_{ij}/n,\ p_i = n_{i\cdot}/n,\ p_j = n_{\cdot j}/n.$