utad

Inteligência Artificial

Clustering – Parte I

Paulo Moura Oliveira
Departamento de Engenharias
Gabinete F2.15, ECT-1
UTAD
email: oliveira@utad.pt

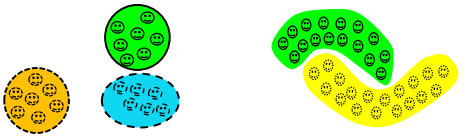
Introdução

Em que consiste o *clustering*?

Clustering, consiste na organização (ou classificação) de um conjunto de dados em vários grupos a que se chamam *clusters*.

Como se faz o *clustering*?

Utiliza-se um dado **critério de similaridade** para agrupar os dados similiares no mesmo grupo (ou de **critério de dissimilaridade** para os distinguir dos outros grupos).



Introdução

Medidas de Proximidade

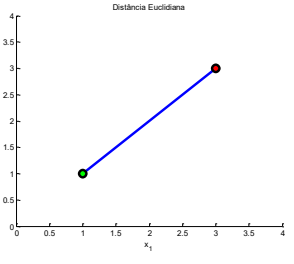
Há muitas formas de determinar a distância entre dois pontos. Das mais conhecidas temos:

Distância Euclidiana

$$dist(x_i, x_j) = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2}$$

d: dimensão

Exemplo:
d=2
x_1 =(1,1) ; x_2 =(3,3)
dist= 2.83



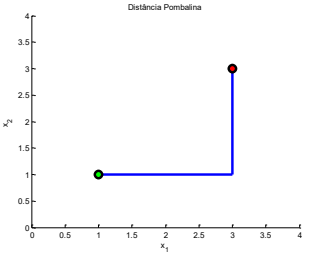
Introdução

Medidas de Proximidade

Distância Pombalina (conhecida como Manhattan ou *CityBlock*)

$$dist(x_i, x_j) = \sqrt{\sum_{k=1}^{d} |x_{ik} - x_{jk}|}$$

Exemplo:
x_1 =(1,1) ; x_2 =(3,3)
d= 4



Introdução

Medidas de Proximidade

Distância Chebychev

$$dist(x_i, x_j) = max_d |x_{id} - x_{jd}|$$

Exemplo:
x_1 =(1,1) ; x_2 =(3,3)
d= 2

Distância Minkowski

$$dist(x_i, x_j) = \sqrt[p]{\sum_{k=1}^{d} (x_{ik} - x_{jk})^p}$$

Exemplo:
p=2
x_1 =(1,1) ; x_2 =(3,3)
d= 2.83

p=2, Igual à Euclidiana

Exemplo:
p=5
x_1 =(1,1) ; x_2 =(3,3)
d= 2.297

Introdução

Como Avaliar os Clusters (Agrupamentos)?

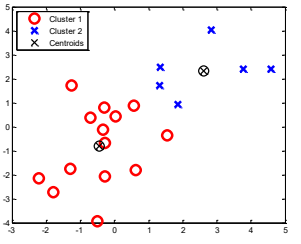
✓ **Coesão Intra-cluster**: avalia a proximidade dos seus pontos ao centróide do cluster

Uma medida muito utilizada é o Somatório do Erro Quadrático (SSE):

$$SSE = \sum_{r=1}^{d} dist^2(x_{ir} - x_{cr})$$

c: centróide do cluster

SSE_1= 47.0279
SSE_2= 14.3666
SSE= 61.3945

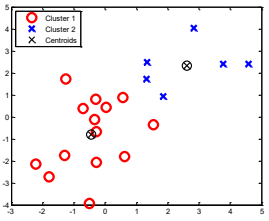


✓ **Separação Inter-cluster**: avalia a separação dos centróides dos vários clusters.

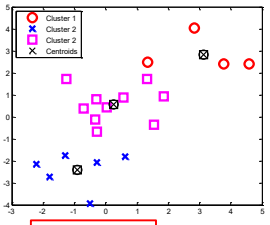
Introdução

Qual o número de Clusters?

✓ Consideremos o mesmo exemplo do diapositivo anterior utilizando o k-means. Em vez de 2 clusters vamos agora considerar 3:



SSE_1= 47.0279
SSE_2= 14.3666
SSE= 61.3945



SSE_1= 7.7042
SSE_2= 8.9528
SSE_3= 15.4975
SSE= 32.1545

Introdução

Qual o número de Clusters?

✓ Um procedimento possível é o seguinte:

1. Definir um **número fixo** de clusters
2. Executar o método de *clustering* e obter o melhor resultado para uma dada função de custo (função objetivo).
3. Voltar a 1 e **aumentar (ou diminuir) o número de clusters**

Introdução

Quais as técnicas de *Clustering*?

✓ Existem várias taxonomias de técnicas de *clustering* que podem ser encontradas na literatura. Uma classificação comum usa três grupos:

1. Hierárquicas (*Hierarchical*)
2. Particionais (*Partitional*) ← **k-Means**
3. *Bayseianas (Bayesian)*

✓ Como o algoritmo k-means é o mais utilizado no contexto da utilização de algoritmos evolutivos, vamos começar por esta técnica.

k-Means

O que é? Técnica de *clustering* que particiona um conjunto de dados em k clusters.

✓ Cada cluster tem um centro (centróide)
✓ O número de clusters, k, é especificado pelo utilizador.

Algoritmo k-means

Selecionar (ou Gerar) k-centros (centróides iniciais)

***while**(!(critério de paragem))*
Atribuir cada amostra de dados ao cluster cujo centróide está mais próximo.
Recalcular os centróides utilizando os clusters atuais
end while

k-Means

Critério de Paragem

✓ Alguns critérios que podem ser utilizados para parar o ciclo do k-means:

1. Um número pré-definido de iterações;
2. Variação dos centróides abaixo de um limiar mínimo;
3. Variação dos pontos nos clusters menor que um valor baixo pré-definido;
4. Soma do erro quadrático abaixo que um valor baixo pré-definido.

k- clusters ← k ; d → **Para todas as dimensões de x**

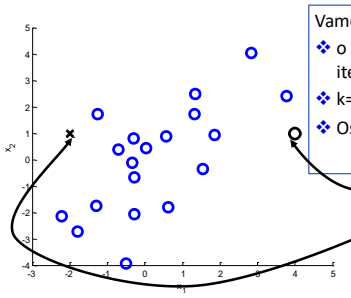
$$SSE = \sum_{j=1}^{k} \sum_{\substack{r=1 \\ x_i \in C_j}}^{d} dist^2(x_{ir} - x_{jr})$$

Elementos de cada cluster, j

k-Means

Exemplo 1

✓ Configuração inicial de um conjunto com 20 pontos.

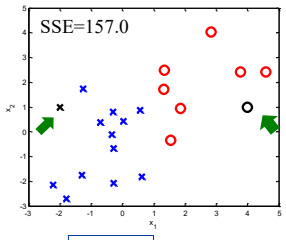


Vamos considerar
- o critério de paragem 5 iterações;
- k=2;
- Os centróides inicializados em: c_1=[-2, 1] e c_2=[4,1]

k-Means

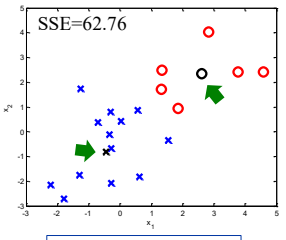
Exemplo 1

✓ Iteração 1:



c_1=[-2, 1]
c_2=[4, 1]

✓ Iteração 2:

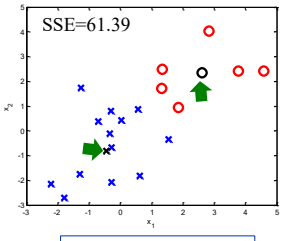


c_1=[-0.5950, -0.8475]
c_2=[2.4633, 1.9504]

k-Means

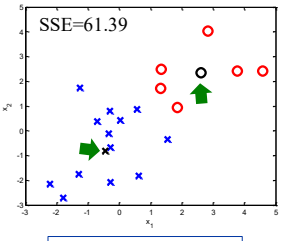
Exemplo 1

✓ Iteração 3:



c_1=[-0.4427, -0.8120]
c_2=[2.6175, 2.3338]

✓ Iteração 4:

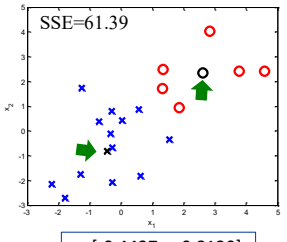


c_1=[-0.4427, -0.8120]
c_2=[2.6175, 2.3338]

k-Means

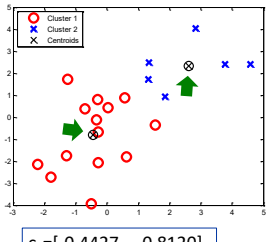
Exemplo 1

✓ Iteração 5:



c_1=[-0.4427, -0.8120]
c_2=[2.6175, 2.3338]

✓ Utilizando a função do Matlab (kmeans):



c_1=[-0.4427, -0.8120]
c_2=[2.6175, 2.3338]

✓ Neste caso deu o mesmo resultado.

Silhueta- *Silhouette*

✓ Uma forma de avaliar a qualidade do *clustering* é utilizando o critério da **silhueta**, cujo valor pode ser determinado para o ponto i:

Mínimo das médias das distâncias do ponto i aos outros pontos dos outros clusters.

Média das distâncias do ponto i aos outros pontos do mesmo cluster.

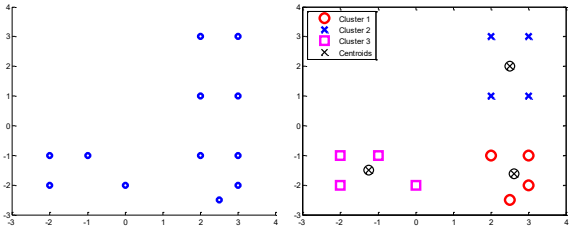
$$s_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

✓ Se:

- Os valores de s_i podem variar entre [-1 e 1];
- Se a maioria valores de s_i estiverem próximos de 1, indica que o ***clustering* é bom**;
- Se a muitos valores de s_i forem baixos ou próximos de -1, indica que o ***clustering* é mau** (precisa de mais ou menos clusters)

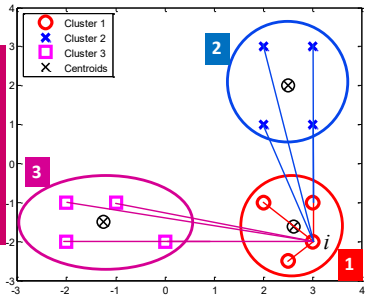
Silhueta- *Silhouette*

Exemplo: ✓ Considere-se a seguinte representação inicial de dados com o respetivo agrupamento com o *k-means*:



Silhueta - *Silhouette*

Exemplo:



Separação: Inter-cluster
d_{i1} =25
d_{i2} =17
d_{i3} =26
d_{i4} =9
b_{i3} =19.5

Separação: Inter-cluster
d_{i1} =10
d_{i2} =26
d_{i3} =9
d_{i4} =25
b_{i2} =17.5

Coesão: Intra-cluster
d_{i1} =1,
d_{i2} =2 ,
d_{i3} =0.5
a_i = 1.1667

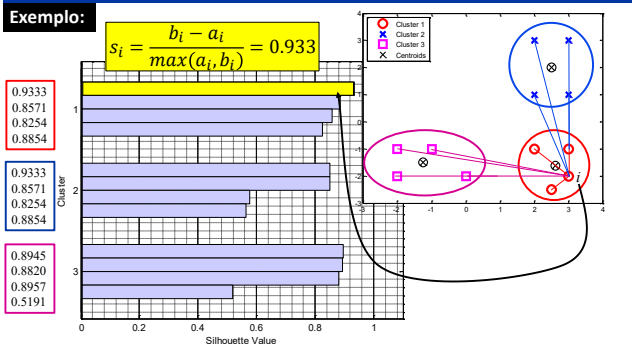
$b_i = \min(b_{12}, b_{13}) = 17.5$
$\min(a_i, b_i) = 17.5$

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} = 0.933$$

Silhueta - *Silhouette*

Exemplo:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} = 0.933$$



Silhueta- *Silhouette*

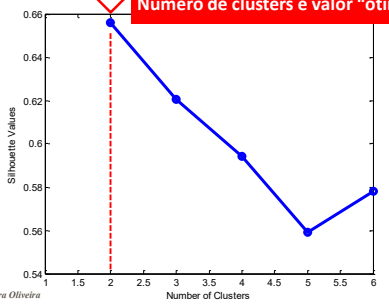
Exemplo 1

✓ Podemos tentar vário números de clusters e ver qual deles dá a **<u>média</u> dos valores da silhueta** menores:

Nota:
As **distâncias** foram calculadas utilizando o **quadrado da distância Euclidianas**.

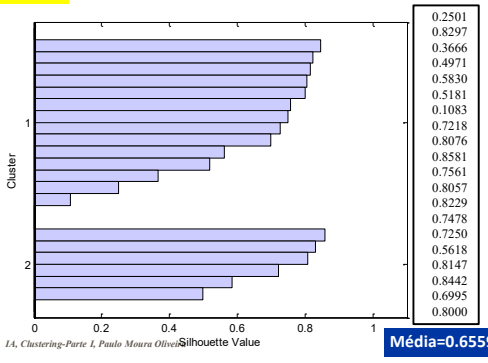
[1	2	3	4	5	6]
[NaN	0.6559	0.6204	0.5944	0.5591	0.5780]

Número de clusters e valor "ótimo"



Silhueta- *Silhouette*

Exemplo 1 ✓ Para o este caso o gráfico dos valores Silhueta é o seguinte:



Média=0.6559