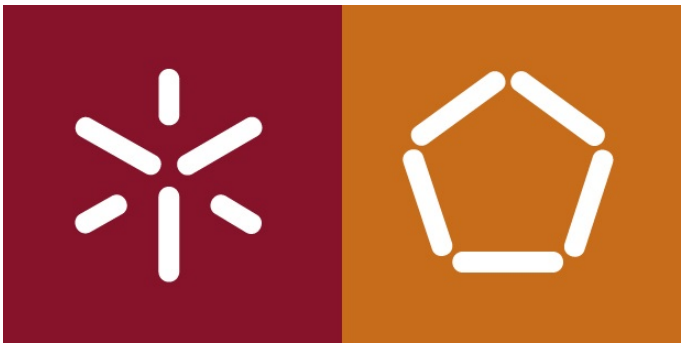

UNIVERSIDADE DO MINHO

DEPARTAMENTO DE INFORMÁTICA

DESCOBERTA DE CONHECIMENTO

Ficha 7

12 de Abril de 2019

Francisco Oliveira (A78416)

Conteúdo

1 Resolução de exercícios

1.1 Exercício 1

O ***k-Means*** é um algoritmo de *clustering*, isto é, analisa as entradas de um determinado dataset e agrupa esses mesmos registos em **k** *clusters*, sendo que os *clusters* são constituidos por registos com características semelhantes.

1.2 Exercício 2

Para isso, o k-Means recolhe uma amostra do *dataset* - um pequeno conjunto de observações - e calcula as médias para cada atributo de cada observação da amostra. De seguida, compara os atributos das restantes observações com as médias obtidas dos valores da amostra. Este processo repete-se iterativamente, e as observações próximas dos valores médios calculados (centróides) vão se agrupando, até que as médias calculadas não sofram uma variação significante entre iterações.

O resultado final são os *clusters* formados pelas observações cujos valores do atributos se assemelham entre si.

1.3 Exercício 3

A Centroid Table revela ao utilizador os valores médios para cada atributo de cada *cluster*.

Estes valores servem para entender qual *cluster* representa um certo grupo de pesquisa, bem como ter um ideia dos valores de dito grupo em comparação aos restantes

1.4 Exercício 4

O dataset a utilizar neste exercício será o ***Iris***. Este dataset contém 3 classes de 50 instâncias cada, em que cada classe se refere a um tipo de planta iris.

Data Set Characteristics:	Multivariate	**Number of Instances:**	150	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	4	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	2535358

Figura 1: Dataset Iris - http://archive.ics.uci.edu/ml/datasets/Iris

Informação dos atributos:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
    - Iris Setosa

- Iris Versicolour
- Iris Virginica

Para aplicar o algoritmo k-means, todos os atributos têm de ser numéricos. Visto que o seu objetivo é criar as suas próprias ”classes”(*clusters*), teremos de remover o atributo *class* do dataset, usando um Select Attributes.

Designamos tambem o atributo ID como identificador de cada observação, usando o operador Set Role.

Visto que o *dataset* já se encontra pronto para ser processado, isto é, todos os valores estão no tipo correto, encontram-se consistentes e não existem valores nulos, aplicou-se o operador ***Clustering k-Means***.

Sabendo, através de uma análise anterior ao dataset, que existem apenas 3 classes, iremos comfigurar o k, no operador k-Means, para que este seja 3:

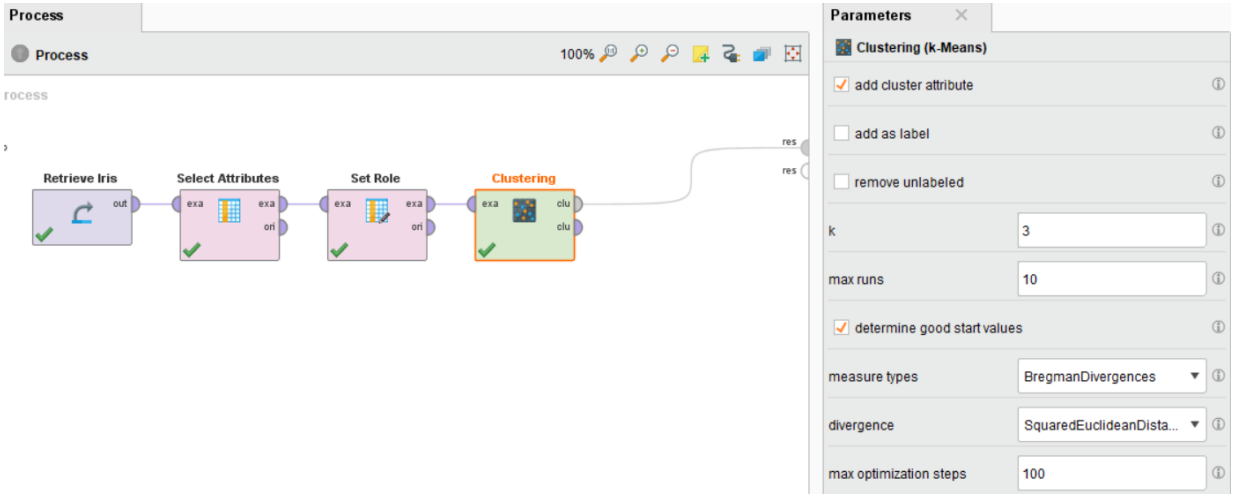

Figura 2: Modelo com algoritmo de *clustering k-Means*

O resultado deste modelo apresenta-se de seguida:

imagenscluster_result.PNG

Figura 3: Cluster Model

Sabendo que o dataset contém 50 instâncias de cada uma das classes/tipos de plantas, verificamos que, apesar de o algoritmo não ser 100% exato para este dataset, os valores encontram-se próximos à realidade. Porém, analisemos o quão bem as observações foram atribuídas a cada *cluster*.

Analisando a ***Centroid Table*** podemos verificar os valores dos centroides de cada *cluster* e assim ver quais os valores médios das observações de cada cluster/classe.

Attribute	cluster_0	cluster_1	cluster_2
a1	5.006	6.854	5.884
a2	3.418	3.077	2.741
a3	1.464	5.715	4.389
a4	0.244	2.054	1.434

Figura 4: Centroid Table

1.5 Exercício 5

Entre o k-means (default) e o k-means (fast), não se verificou nenhuma alteração. Todos os *clusters* existentes permaneceram iguais.

No entanto o k-means (kernel) é um pouco diferente e não permite retornar uma centroid table. Olhando para os resultados denota-se que os *clusters* do *kernel* são um pouco mais equilibrados em numero de elementos. Verifica-se também que os elementos de cada *cluster* estão um pouco mais dispersos