

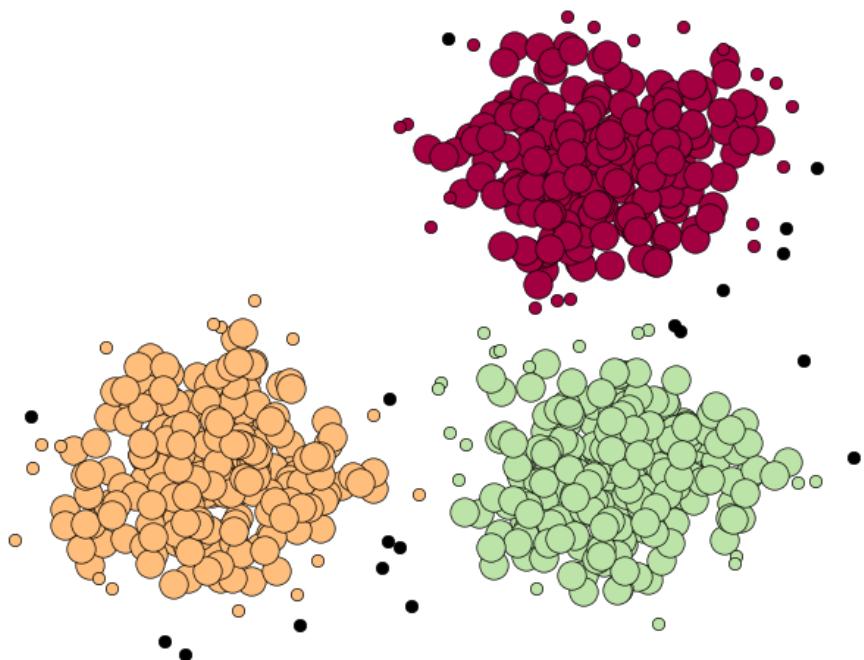


Universidade do Minho

13 de Abril de 2020

O Método de *Elbow*

Clusters particionais com dados numéricos



Bruno Jácome, A89515

Carolina Barros, A84950

Dinis Gomes, A87993

Joana Gouveia, A85650

João Silva, A84617

Jorge Gonçalves, A84133

Pedro Peixoto, A89602

Índice

1	Introdução	4
2	Clusters	5
2.1	O que são?	5
2.2	Clusters Particionais	5
2.2.1	O que são?	5
2.3	<i>Clustering</i>	5
3	Centróides	6
3.1	O que são?	6
3.2	Relação entre centróide e média	6
3.2.1	Semelhanças	6
3.2.2	Diferenças	6
4	Algoritmos de Clusters Particionais	7
4.1	K-means	7
4.1.1	O que é?	7
4.1.2	Restrições	7
4.1.3	Determinação do K-Means	8
4.2	K-medoids	9
4.2.1	O que é?	9
4.2.2	Medóide	9
4.3	Diferença entre K-Means e K-Medoids	9
4.3.1	A nível de sensibilidade	9
4.3.2	A nível de Centróide	9
4.3.3	A nível atributos	10
5	O Método de <i>Elbow</i>	11
5.1	O que é?	11
5.2	Pré-aplicação	12
5.3	WCSS	12
5.3.1	Ilusão de solução ótima	12
5.3.2	Solução ótima para o problema	13
5.4	Aplicação	13
6	Conclusões	15

Listas de Figuras

2.1	Agrupamento de clusters.	5
4.1	Convergência do <i>K-means</i>	7
4.2	<i>k-means clustering</i> com $\mathbf{k} = 2$	8
4.3	<i>PAM</i> com $\mathbf{k} = 3$	9
5.1	Número de clusters (3 vs 5) representando tamanhos de <i>t-shirts</i> .	11
5.2	WCSS mínimo	12
5.3	WCSS máximo	13
5.4	Solução ótima	13
5.5	Procedimentos do método de <i>method</i>	14
5.6	Valor de WCSS <i>versus</i> número de <i>clusters</i>	14

1 | Introdução

Este trabalho foi realizado no âmbito da Unidade Curricular de Matemática das Coisas e teve como principal objetivo o estudo de clusters particionais com dados numéricos (centróide) através do *The Elbow Method*.

O presente relatório divide-se em 4 partes. Primeiramente, no Capítulo 2, será feita uma contextualização do assunto, apresentam-se a definição de clusters no geral e, mais em concreto, de clusters particionais.

Seguidamente, no Capítulo 3, será descrito o conceito de centróide, bem como outros aspectos relevantes associados a este conceito.

Depois, no Capítulo 4, serão abordados alguns algoritmos de clusters particionais, com a apresentação de algumas das suas aplicações mais práticas.

No capítulo seguinte, apresentar-se-á o **método de elbow**, o assunto principal deste trabalho.

Para finalizar, expor-se-á uma breve conclusão do trabalho apontando-se essencialmente as desvantagens do método de *elbow*.

2 | Clusters

2.1 O que são?

Um **cluster** é um conjunto de objetos similares entre si e dissimilares em relação a objetos noutros clusters. A análise de clusters ou o seu conceito, é um procedimento humano normal, muitas vezes usado de forma inconsciente.

A análise de clusters é usada em inúmeras aplicações, tais como no reconhecimento de padrões (*machine learning*), processamento de imagem e pesquisas de mercado.

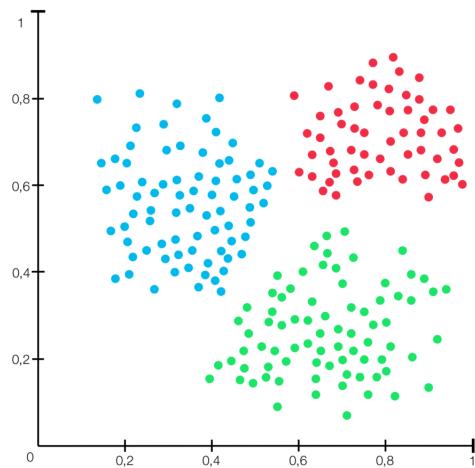


Figura 2.1: Agrupamento de clusters.

2.2 Clusters Particionais

2.2.1 O que são?

Cluster particional define-se, especificamente, quando se utiliza um agrupamento particional que consiste em dividir os dados em subconjuntos sem que haja intersecções, o que leva a que cada objeto esteja exatamente num subconjunto.

2.3 *Clustering*

O *clustering* é o conjunto de técnicas de prospeção de dados, isto é, exames minuciosos e metódicos, que fazem agrupamentos automáticos de dados segundo o seu grau de semelhança. Normalmente o usuário do sistema deve escolher a priori o número de grupos a serem detetados. Alguns algoritmos mais sofisticados pedem apenas o número mínimo e outros tem a capacidade de subdividir um grupo em dois. Existem vários tipos de agrupamentos, mas o que analisaremos com mais detalhe serão os agrupamentos **particionais**.

3 | Centróides

3.1 O que são?

Um **centróide** é o ponto que representa o *centro* de todos os pontos pertencentes a um cluster. No que diz respeito aos modelos centróides, a noção de similaridade deriva da proximidade dos pontos com o centróide do *cluster*.

Estes são obtidos através de operações algébricas e, em regra, não pertencem à base de dados. Logo, são uma mera interpretação de resultados e que dependem maioritariamente da definição de proximidade entre dois objetos de estudo.

3.2 Relação entre centróide e média

3.2.1 Semelhanças

A **média** de um cluster é o mesmo que o centróide, contudo o termo **centróide** é mais preciso quando se estuda *multivariate data*, isto é, dados multivariados.

Um centróide é às vezes denominado de **centro de massa** ou **barycenter** (centro de gravidade), baseado na sua interpretação física. Assim como a média, a localização do centróide **minimiza a soma dos quadrados das distâncias entre os pontos** ou **WCSS (With-in cluster sum-squared)**.

3.2.2 Diferenças

Há, no entanto, uma diferença entre **distância de centróide** e **distância média** quando se comparam clusters. A distância de centróide entre dois quaisquer clusters A e B é simplesmente a distância entre o centróide de A e o centróide de B. Já a distância média é calculada encontrando-se a distância média entre todos os pares de pontos de cada cluster.

4 | Algoritmos de Clusters Particionais

O Cluster Particional tem dois algoritmos: o *K-means* e o *K-medoids*. Estes algoritmos tem as suas diferenças. Uma delas é o facto de no algoritmo *K-means* termos a soma máxima das distâncias e no algoritmo *K-medoids* termos a soma mínima nas distâncias. A execução destes métodos podem ser representados graficamente por diagramas de Voronoi criando **células de Voronoi**, que no caso do *k-means*, cria células distintas associadas a cada centróide, isto é, distingue cada cluster.

4.1 K-means

4.1.1 O que é?

O *k-means* é um algoritmo de *clustering* bastante comum e popular usado por numerosos investigadores em todo o mundo. Este tem por objetivo pôr em partes m observações dentro de k clusters, onde cada observação está dentro do cluster com o qual está mais próxima, usando o diagrama de *Voronoi*. Nestes modelos, os números de clusters necessários no final (k) têm de ser mencionados com antecedência, o que torna importante o conhecimento prévio do conjunto de dados. [13]

Figura 4.1: Convergência do *K-means*

4.1.2 Restrições

Uma restrição que este algoritmo tem é o facto de apenas funcionar com atributos quantitativos, necessita de fazer operações algébricas, como somas e multiplicações por escalar, que dará origem a uma matriz que é chamada "matriz da partição". A nível de pontos que se encontram fora da curva, temos de ter cuidado pelo facto de os mesmos poderem facilmente influenciar o valor da média e levar a mesma a alterar-se.

4.1.3 Determinação do K-Means

Algorithm 1 Pseudocódigo K-means

```
procedimento K-MEANS(K, Conjunto de dados = { $x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$ )
    Inicializar aleatoriamente K centróides  $\mu_1, \mu_2, \dots, \mu_K \in IR$ 
    repita
        para  $i = 1$  até  $m$  faz
             $c^{(i)} :=$  índice (de  $i$  até  $K$ ) do centróide mais próximo de  $x^{(i)}$    ▷ ( $d(x^{(i)}, \mu_i)$  mínima)
        para  $k = 1$  até  $K$  faz
             $\mu_k :=$  ponto médio dos pontos no cluster  $k$    ▷ Passo de reajustamento de centróides
        até nenhum centróide se reajustar
```

Ver mais em [11].

Figura 4.2: *k-means clustering* com $\mathbf{k} = 2$

4.2 K-medoids

4.2.1 O que é?

O *k-medoid* ou *partitioning around medoids (PAM)* são algoritmos de *clustering* reminiscentes do algoritmo de *k-means*, na medida em que ambos operam de modo particional e ambos tentam minimizar a distância entre os pontos e o centróide, dentro de um cluster. [10]

4.2.2 Medóide

Figura 4.3: PAM com $k = 3$

Uma ideia semelhante à de centróide é a de **medóide**, que é o *data point* que é *menos parecido* de todos os outros pontos de dados.

Ao contrário do centróide, a medóide tem de ser um dos pontos originais.

4.3 Diferença entre K-Means e K-Medoids

4.3.1 A nível de sensibilidade

O *K-medoid* lida melhor com os *outliers* (pontos fora da curva) do que o *K-means*. É menos sensível a eles, porque minimiza a soma das diferenças, contrariamente a *k-means*, que maximiza.

4.3.2 A nível de Centróide

O centro de *k-medoids* não é o ponto médio mas sim um ponto real, porque é o objeto mais centralmente localizado do cluster, que como já referimos, tem somas mínimas de distância.

4.3.3 A nível atributos

Os atributos de *K-medoids* podem ser atributos quantitativos, tal como *k-mean*, mas também podem ser atributos qualitativos, o que leva a que não exista uma necessidade e obrigação do uso de operações algébricas neste algoritmo. Estes atributos encontram-se representados na base de dados.

5 | O Método de *Elbow*

5.1 O que é?

Uma etapa fundamental para qualquer aprendizagem não-supervisionada é determinar o número ideal de clusters nos quais os dados podem ser agrupados: K .

O **Método de Elbow** é uma heurística, uma vez que é um método criado para encontrar soluções sobre um problema complexo, como uma medida que preserva e conserva energia e os recursos mentais. Neste caso, para determinar o número ideal de clusters no *k-means clustering*, este método parcela o valor da função custo produzida pelos diferentes valores de K . Ora, isto só é possível ignorando parte da informação com o objetivo de tornar a escolha mais fácil e rápida.

Sendo assim, não há uma resposta universal para este problema já que o número ideal de clusters é de alguma forma subjetivo e depende não só do propósito do *clustering*, mas também do método usado para medir as similaridades e os parâmetros usados para particionar.

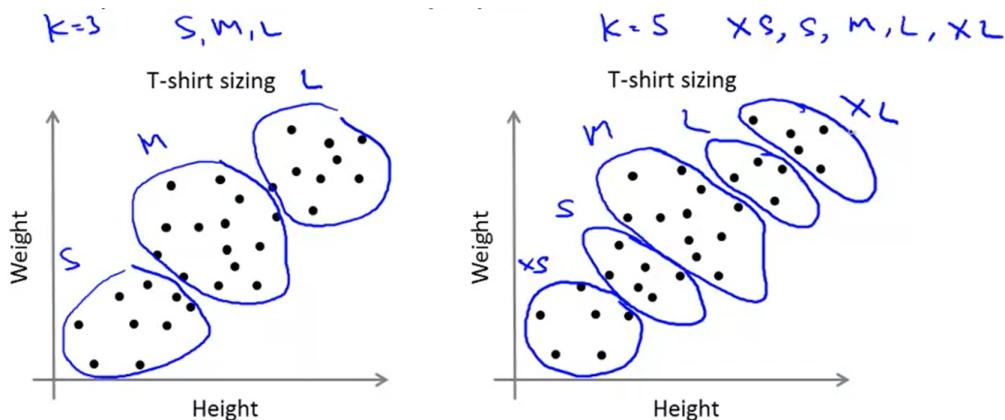


Figura 5.1: Número de clusters (3 vs 5) representando tamanhos de *t-shirts*.

Ver em [12].

5.2 Pré-aplicação

Numa fase inicial, criar um dendrograma, ou seja, um diagrama que organize as variáveis, agrupando-as de forma hierárquica ascendente - o que em termos gráficos se assemelha aos ramos de uma árvore.

Seguidamente, inspecionar o dendrograma produzido usando o cluster hierárquico para verificar se ele sugere um número específico de clusters. (Todavia, esta abordagem também é subjetiva.)

Estes métodos, apresentados a seguir, incluem métodos diretos e teste estatístico.

- **Métodos diretos:** consistem em otimizar um critério, como a somas dos quadrados das distâncias *intra-cluster* (*Within Cluster Sum of Squares*) ou a média *silhouette*. Os métodos correspondentes são denominados métodos de *Elbow* e *silhouette*, respectivamente.
- **Métodos de teste estatístico:** consistem em comparar evidências contra hipóteses nulas. Um exemplo é a estatística de gap.

5.3 WCSS

5.3.1 Ilusão de solução ótima

Comummente, considera-se que para se obter a solução ótima de número de clusters, deve-se calcular o mínimo de **WCSS**. Porque será então isto um erro?

Para um valor mínimo de **WCSS**, a solução ótima de número de clusters é igual ao número total de *data points*. A noção de cluster acaba por perder o seu propósito, acabando por ser uma solução trivial para o problema.

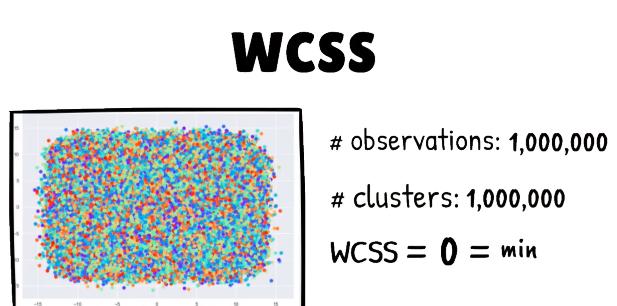
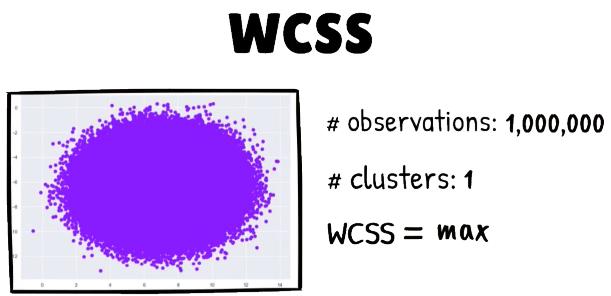


Figura 5.2: WCSS mínimo



Por oposição, para um valor máximo de WCSS, a solução ótima seria apenas um cluster, isto é óbvio uma vez que a soma do quadrado das distâncias *intra-cluster* só poderia ser máxima se este contivesse todos os pontos.

Figura 5.3: WCSS máximo

5.3.2 Solução ótima para o problema

Em boa verdade, não há. Como foi dito anteriormente, o número de clusters ótimo depende de vários fatores, inclusivé do objetivo de cada "clusterização" específica.

Contudo, com a ajuda do **método Elbow**, é possível obter um resultado ótimo de **equilíbrio** entre o **número de clusters** e **WCSS**.

WCSS
MIDDLE GROUND?

observations: **N**
clusters: **SMALL**
WCSS = **LOW**

Figura 5.4: Solução ótima

5.4 Aplicação

O método de Elbow considera o WCSS total como uma função do número de clusters: deve-se escolher um número de clusters para que a adição de outro cluster não melhore muito mais o WCSS total.

O número ótimo de clusters pode ser obtido da seguinte forma:

1. Aplicar o algoritmo de *clustering*, por exemplo, *k-means clustering*, para diferentes valores de *k*.
2. Para cada *k*, calcular o WCSS;
3. Representar, graficamente, o WCSS em função do *k*;
4. Localizar, no gráfico, a curva com uma aparênciade **cotovelo (elbow)**, geralmente considerado o indicador do número ótimo de *clusters*.

Ver em [1, 2].

Figura 5.5: Procedimentos do método de *method*

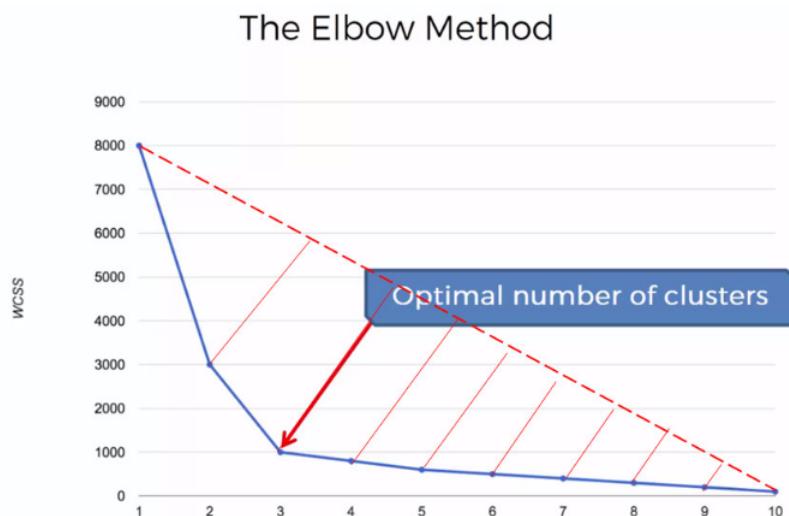


Figura 5.6: Valor de WCSS *versus* número de *clusters*

ASANKA PERERA afirma que, "após ter lido alguns artigos, descobriu que se desenharmos uma linha reta entre o ponto 1 e 10, e se calcularmos a distância de cada ponto até à linha, o ponto com a maior distância será o ponto que contém o cotovelo". Ver em [3].

6 | Conclusões

Apesar de bastante prático, devido ao método de *Elbow* ser um método visual, este tem uma interpretação subjetiva e nem sempre clara. Mais concretamente, o verdadeiro *elbow*, ou cotovelo, nem sempre é identificado sem ambiguidade já que, pode ou nem haver nenhum, ou nem haver um só único.

Além disso, verifica-se, em geral, que ocorrem abruptas distorções descendentes até $k=3$, a partir do qual a curva descende lentamente. De facto, como é explicado na figura 5.2, à medida que K cresce, o número de *data points* por cluster diminui, até que K seja esse número e, nesse caso, *WCSS* é 0. Logo, torna-se óbvio que o número ideal de clusters nunca será elevado, a menos que o utilizador o queira, mas nesse caso nem fará sentido auxiliar-se neste método.

Outros problemas associados ao método, devem-se à aplicação do *k-means*. O algoritmo mais comum usa uma técnica de refinamento iterativo e é muitas vezes chamado de algoritmo de *Lloyd*. Embora existam alternativas bem mais eficientes, este algoritmo tem um longo tempo de execução, particularmente a calcular as distâncias de cada nodo em relação aos K centróides. Já que, a maior parte dos pontos ficam associados aos mesmos centróides após algumas iterações, a maior parte deste trabalho é inútil, tornando esta implementação bastante ineficiente.

Bibliografia

- [1] THE SCIKIT-YB DEVELOPERS, *Elbow Method*,
<https://www.scikit-yb.org/en/latest/api/cluster/elbow.html>
(13/04/2020)
- [2] *Determining The Optimal Number Of Clusters: 3 Must Know Methods*,
<https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/#elbow-method>
(13/04/2020)
- [3] ASANKA PERERA, *Finding the optimal number of clusters for K-Means through Elbow method using a mathematical approach compared to graphical approach*,
<https://www.linkedin.com/pulse/finding-optimal-number-clusters-k-means-through-elbow-asanka-perera>
(13/04/2020)
- [4] LACHI, RICARDO LUÍS & ROCHA, HELOÍSA VIEIRA DA. FEVEREIRO 2005, *Aspectos básicos de clustering: conceitos e técnicas(Brasil)*
(13/04/2020)
- [5] RODRIGO CEZAR MENEZES, *Clusterização de Dados*,
https://www.maxwell.vrac.puc-rio.br/24787/24787_5.PDF
(13/04/2020)
- [6] MANUEL ALTINO TORRES ANICETO CASTRO, *Agrupamento – “Clustering”*,
<http://www.dei.isep.ipp.pt/~paf/proj/Julho2003/Clustering.pdf>
(13/04/2020)

- [7] VICTOR LAVRENKO, *Hierarchical Clustering 3: single-link vs. complete-link*,
<https://www.youtube.com/watch?v=VMyXc3SiEqs>
(13/04/2020)
- [8] VICTOR LAVRENKO *K Means Clustering: Pros and Cons of K Means Clustering*,
<https://www.youtube.com/watch?v=YIGta1P1mv0>
(13/04/2020)
- [9] TEST MY CHATBOT, *How to Choose the Number of Clusters | Advanced Statistical Methods - K-Means Clustering*,
<https://www.youtube.com/watch?v=SCA07-7Xe6Q>
(13/04/2020)
- [10] *k-medoids*,
<https://en.wikipedia.org/wiki/K-medoids>
(13/04/2020)
- [11] UNIVERSIDADE DE STANFORD, *K-Means Algorithm*,
<https://pt.coursera.org/lecture/machine-learning/k-means-algorithm-93VPG>
(13/04/2020)
- [12] UNIVERSIDADE DE STANFORD, *K-Means Algorithm*,
<https://pt.coursera.org/lecture/machine-learning/choosing-the-number-of-clusters-Ks0E9>
(13/04/2020)
- [13] *k-means clustering*,
https://en.wikipedia.org/wiki/K-means_clustering
(13/04/2020)