

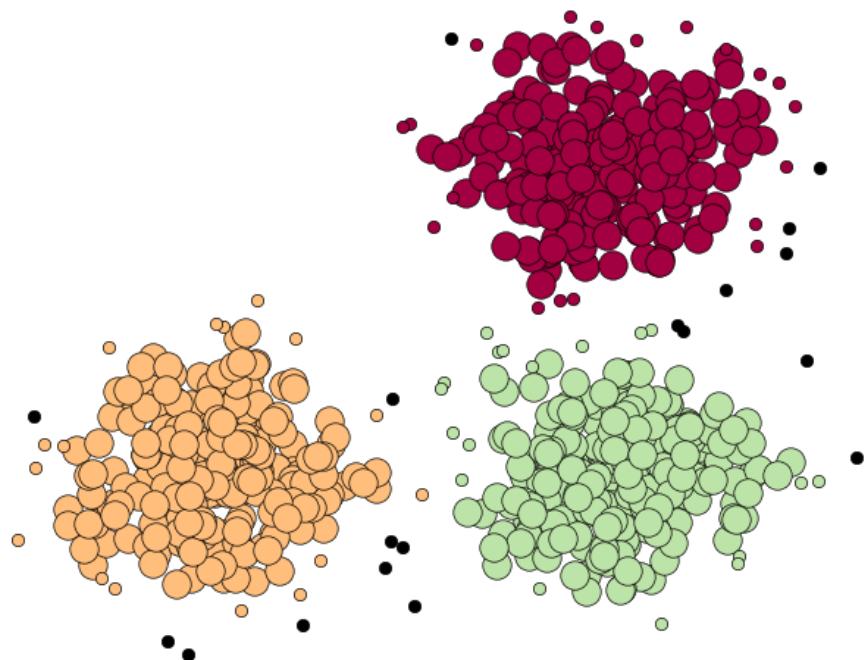


Universidade do Minho

14 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	Algoritmo <i>K-means</i>	8
4.2	K-medoids	9
4.2.1	O que é?	9
4.2.2	Medóide	10
4.3	Diferença entre K-means e K-medoids	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 $k = 2$	9
4.3	<i>PAM</i> com $k = 3$	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>elbow</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 [5, 6].

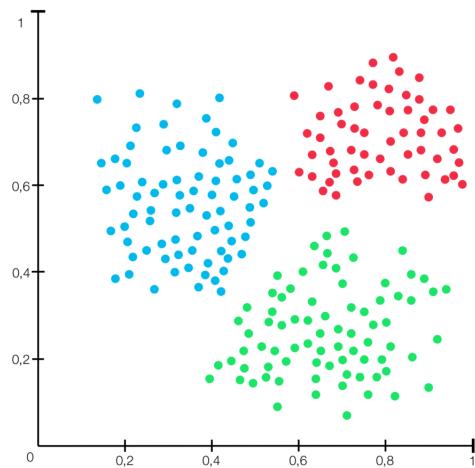


Figura 2.1: Agrupamento de clusters.

2.2 Clusters Particionais

2.2.1 O que são?

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

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 detectados. Alguns algoritmos mais sofisticados pedem apenas o número mínimo e outros tem a capacidade de subdividir um grupo em dois.

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 está associada à proximidade dos pontos ao centróide do *cluster*.

Os centróides são obtidos através de operações algébricas e, em regra, não pertencem à base de dados. Logo, são uma mera interpretação de resultados e 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 pode ser 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 para cada cluster e é a média das distâncias entre todos os pontos do cluster e o centróide.

4 | Algoritmos de Clusters Particionais

O Cluster Particional tem dois algoritmos: o *K-means* e o *K-medoids*. Estes algoritmos 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 iterativo de *clustering* bastante comum e popular usado por numerosos investigadores em todo o mundo. Este tem por objetivo distribuir m observações (totalidade dos *data points*) por 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, uma vez que requer operações algébricas, como somas e multiplicações por um escalar, o que dará origem a uma matriz que é chamada "matriz da partição". A nível de pontos que se encontram fora da curva representativa dos dados, terá de haver preocupação com os mesmo por puderem facilmente alterar o valor da média.

4.1.3 Algoritmo 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  $\mu_1, \mu_2, \dots, \mu_K \in \mathbb{R}^n$ 
    repita
        para  $i = 1$  até  $m$  faz                                 $\triangleright$  Passo de associação de centróides
             $c^{(i)} :=$  índice (de 1 até  $K$ ) do centróide mais próximo de  $x^{(i)}$   $\triangleright$   $\text{dist}(x^{(i)}, \mu_k)$  mínima
        para  $k = 1$  até  $K$  faz                                 $\triangleright$  Passo de reajuste de centróides
             $\mu_k :=$  ponto médio dos pontos no cluster  $k$ 
        até nenhum centróide se reajustar
    
```

Nota: Uma outra possibilidade para a inicialização dos K centróides é escolhermos os primeiros K elementos do conjunto dos dados $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$.

Para $x^{(i)} = (x_1^{(i)}, \dots, x_n^{(i)})$ e $\mu_k = (\mu_1^{(k)}, \dots, \mu_n^{(k)})$, temos

$$\begin{aligned}\text{dist}(x^{(i)}, \mu_k) &= \|x^{(i)} - \mu_k\| \\ &= \sqrt{\left(x_1^{(i)} - \mu_1^{(k)}\right)^2 + \dots + \left(x_n^{(i)} - \mu_n^{(k)}\right)^2}.\end{aligned}$$

Minimizar $\|x^{(i)} - \mu_k\|$ equivale a minimizar $\|x^{(i)} - \mu_k\|^2$, ou seja, a determinar o mínimo da soma dos quadrados,

$$\|x^{(i)} - \mu_k\|^2 = \left(x_1^{(i)} - \mu_1^{(k)}\right)^2 + \dots + \left(x_n^{(i)} - \mu_n^{(k)}\right)^2.$$

Portanto, em vez de determinarmos o mínimo das distâncias $\text{dist}(x^{(i)}, \mu_k)$, para $k = 1, \dots, K$, determinamos o mínimo dos quadrados das distâncias. Assim, se $c^{(i)}$ é o índice do centróide mais próximo de $x^{(i)}$, temos

$$\text{dist}^2(x^{(i)}, \mu_{c^{(i)}}) = \min_{k=1, \dots, K} \|x^{(i)} - \mu_k\|^2.$$

No final de cada iteração, os centróides são atualizados e cada novo centróide passa a ser o ponto médio dos pontos do respetivo cluster. O processo repete-se até que todos os clusters se mantenham inalterados. Ver pormenores 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].

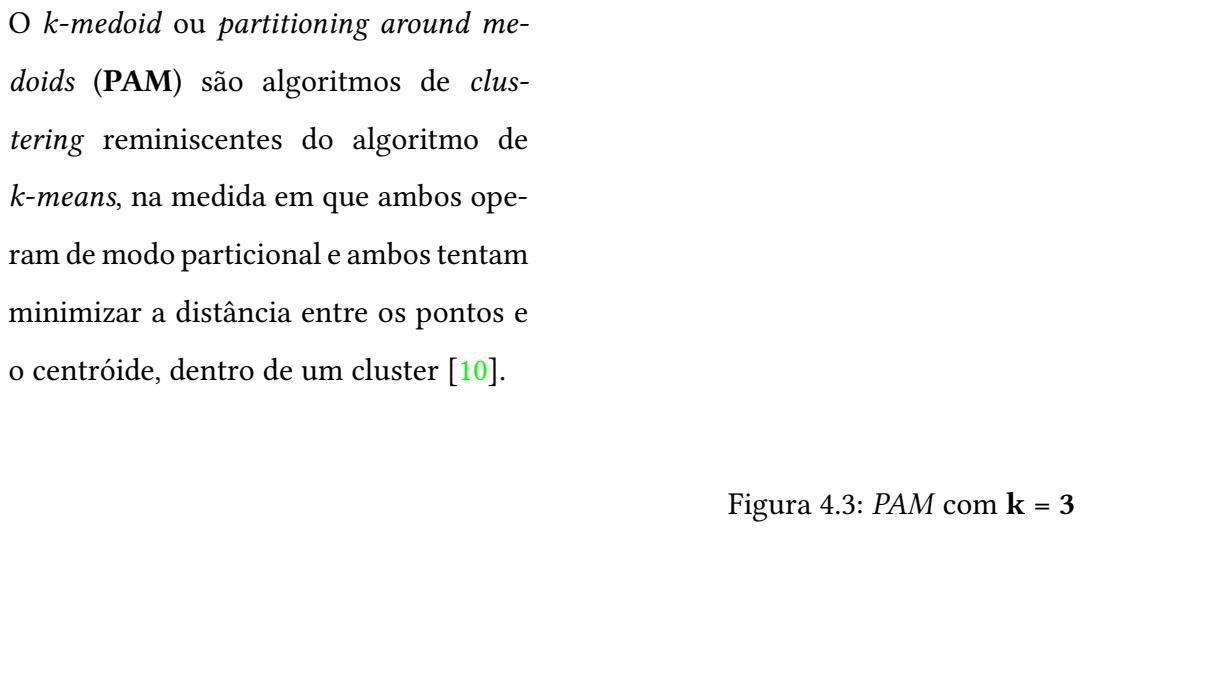


Figura 4.3: *PAM* com $\mathbf{k} = 3$

4.2.2 Medóide

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

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

4.3 Diferença entre K-means e K-medoids

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.

No que diz respeito ao 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.

Quanto aos atributos, os atributos de *K-medoids* podem ser quantitativos, tal como *k-mean*, mas também podem ser 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.

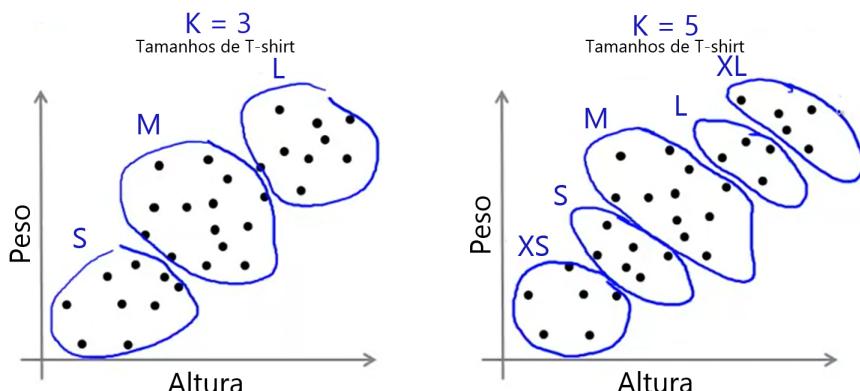


Figura 5.1: Número de clusters (3 vs 5) representando tamanhos de *t-shirts*.

Ver mais detalhes em [12].

5.2 Pré-aplicação

Numa fase inicial, é útil criar um dendrograma, isto é, 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, deve-se inspecionar o dendrograma produzido usando o cluster hierárquico para verificar se este sugere um número específico de clusters. Todavia, esta abordagem também é subjetiva.

Estes métodos incluem métodos diretos e de teste estatístico.

- **Métodos diretos:** consistem em otimizar um critério, como a somas dos quadrados das distâncias dentro dos clusterss. (*WCSS - With-in 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, pois temos uma solução trivial para o problema.

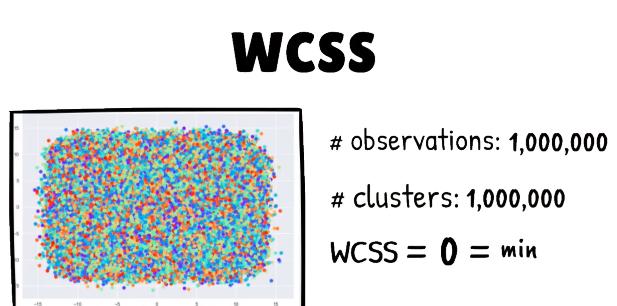
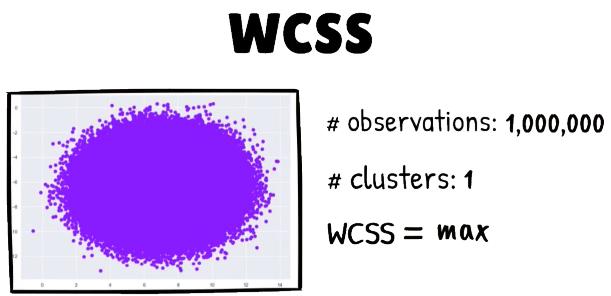


Figura 5.2: WCSS mínimo



Por oposição, para um valor máximo de WCSS, a solução ótima seria apenas um cluster. Isto é óbvio uma vez que a soma dos quadrados 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 uma solução ótima. 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 de 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. Ver [1, 2].

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 aparência de **cotovelo (elbow)**, geralmente considerado o indicador do número ótimo de *clusters*.

Figura 5.5: Procedimentos do método de *elbow*

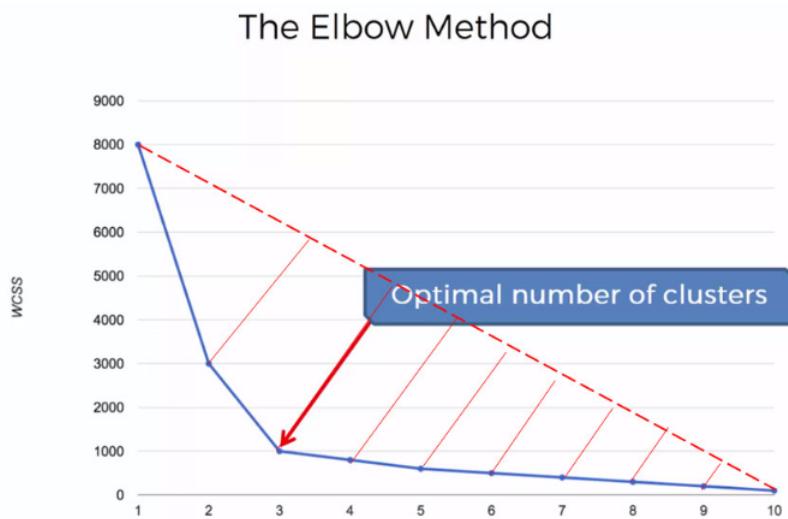


Figura 5.6: Valor de WCSS *versus* número de *clusters*

ASANKA PERERA[3] 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".

6 | Conclusões

O método de *Elbow*, apesar de bastante prático por ser um método visual, tem uma interpretação subjetiva e nem sempre muito clara. Mais concretamente, o verdadeiro *elbow*(cotovelo), nem sempre é identificado sem ambiguidade, já que pode acontecer não haver nenhum *elbow* ou haver mais do que um.

Além disso, verifica-se, em geral, que ocorrem descidas abruptas no *WCSS* até $k=3$ e a partir deste valor a curva desce lentamente. De facto, como é explicado na figura 5.2, à medida que K cresce, o número de *data points* por cluster diminui, e o *WCSS* também. Quando K é igual ao número total de *data points*, o *WCSS* é 0. Logo, torna-se óbvio que o número ideal de clusters nunca deverá ser elevado, a menos que o utilizador o queira, mas nesse caso nem fará sentido à aplicação método.

Outros problemas associados ao método de *elbow*, devem-se à aplicação do algoritmo de *clustering k-means*. O procedimento 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 devido ao cálculo das distâncias de cada ponto aos K centróides. A maior parte dos pontos ficam associados aos mesmos centróides após algumas iterações e, por isso, a maior parte deste trabalho é inútil, tornando esta implementação bastante ineficiente.

Apesar das dificuldades referidas, que aparecem sobretudo quando o número de *data points* é elevado, a aplicação do método de *elbow* em combinação com o *k-means* produz resultados bastante satisfatórios em diversas aplicações.

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)