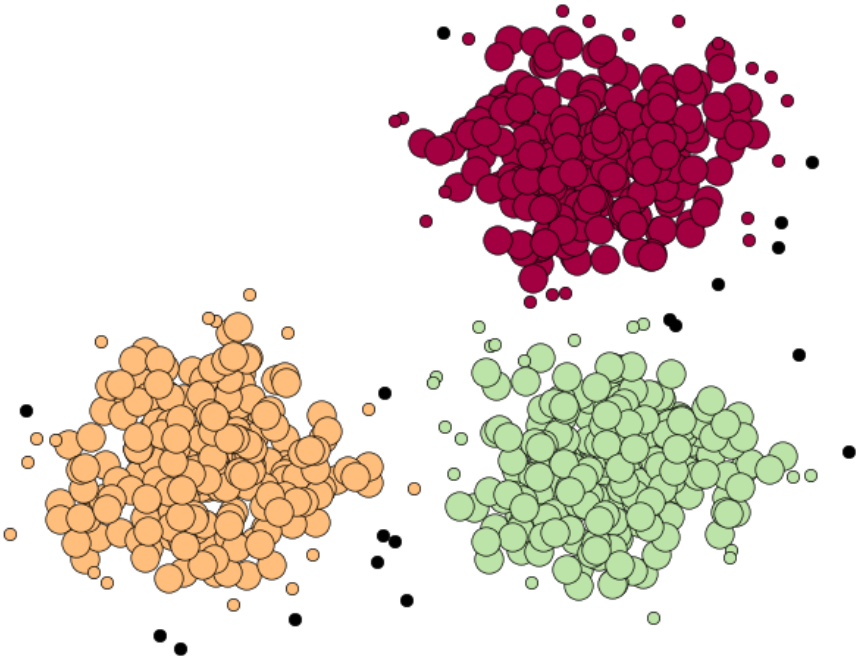
Universidade do Minho

1 de Abril de 2020

O Método *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

Lista de Ilustrações

Tabelas

1 | Introdução

Este trabalho foi realizado no âmbito da Unidade Curricular de Matemática das Coisas e tem como objetivo primordial o estudo do *Clusters* particionais com dados numéricos (centróide) atráves do *The Elbow Method*.

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

Seguidamente, no Capítulo 3, será descrito o conceito de centróides, bem como outros aspetos relevantes relativos.

Depois, no Capítulo 4, será abordado o *The Elbow Method*, com a apresentação da definição teórica e a sua aplicação mais prática.

No capítulo seguinte, a parte teórica será aplicada em exemplos mais práticos, de forma a melhor entendermos a aplicação dos tópicos referidos nos capítulos anteriores.

Para finalizar, expor-se-á uma breve conclusão do trabalho apontando-se os aspetos mais enriquecedores para o nosso conhecimento.

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. [6][7]

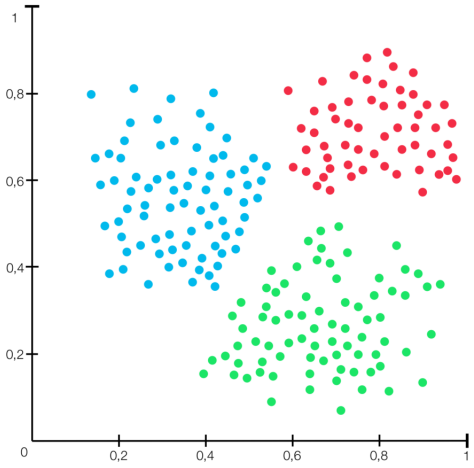


Figura 2.1: Figura ilustrativa do agrupamento de clusters.

Muito cedo nas escolas, os alunos aprendem a classificar e agrupar, por exemplo distinguir entre gatos e cães, entre animais e planta, progredindo num refinamento de classificação que tem subjacente teorias de *clustering*. A análise de clusters é usada em inúmeras aplicações, tais como no reconhecimento de padrões (*machine learning*), processamento de imagem e pesquisa de mercado.

2.2 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 nú-

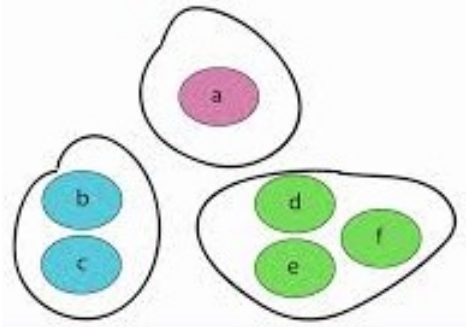


Figura 2.2: Clustering.

mero mínimo e outros tem a capacidade de subdividir um grupo em dois. Existem vários tipos de agrupamentos, mas o que será analisado com mais detalhe serão os **particionais**.

2.2.1 Clustering na História

O primeiro registo publicado sobre um método de clustering foi feito em 1948, com o trabalho de *SORENSEN* (1948) sobre o Método Hierárquico de Ligação Completa. Desde então mais de uma centena de algoritmos distintos de clustering já foram definidos.

2.3 Clusters 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*.

Além disso, os centróides são obtidos através de operações algébricas (somas e multiplicações por escalares) e, em regra, estes 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 dada*, isto é, dados multivariados.

Um centróide é ás 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 *sum-squared distance* entre os outros pontos**.

3.2.2 Medóide

Uma ideia semelhante é a de **medóide**, que é o ponto de dado que é *menos parecido* de todos os outros pontos de dados.

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

3.2.3 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.

$$\text{dist}(A, B) = \frac{\sum_{ij} \text{dist}(a_i, b_j)}{\#A \times \#B}, \quad \forall a_i \in A, b_j \in B$$

Métrica de Clusters: Distância média

$$\text{dist}(A, B) = \text{dist}\left(\frac{\sum_i a_i}{\#A}, \frac{\sum_i b_i}{\#B}\right), \quad \forall a_i \in A, b_i \in B$$

Métrica de Clusters: Distância entre centróides

Estes dois cálculos são duas métricas possíveis para calcular a distância entre dois clusters, mas existem mais métodos.[8]

3.2.4 Exemplo

3.3 Como se determinam?

4 | O Método Elbow

4.1 Conceitos Relevantes

Uma etapa fundamental para qualquer aprendizagem não-supervisionada é determinar o número ideal de *clusters* segundo os quais os dados podem ser agrupados. Neste sentido, o *The Elbow Method* é um dos métodos mais populares para determinar esse valor ótimo de *K*, sendo *K* o número de *clusters* que o utilizador da informação decide agrupar.

Desta forma, o método pode ser considerado heurístico, ou seja, corresponde a um método ou processo criado com o objetivo de encontrar soluções para um problema de interpretação e validação de consistência dentro análise de agrupamento concebido para ajudar a encontrar o número apropriado de aglomerados num conjunto de dados. Para que tal seja possível, são definidas estratégias que ignoram parte da informação com o objetivo de tornar a escolha mais fácil e rápida.

Apesar das características mais positivas relativas a este método, em algumas situações, pode ser considerado ambíguo e pouco confiável, e, portanto, podem ser utilizadas outras abordagens para determinar o número de *clusters*, são preferíveis.

Assim sendo, o *The Elbow Method* é utilizado 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*.

No entanto, não há uma resposta definitiva para esta pergunta. O número ideal de *clusters* é de alguma forma subjetivo e depende do método usado para medir as similaridades e os parâmetros usados para particionar.

Uma solução simples e popular consiste em, 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.

Após esta primeira fase, é fundamental 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 métodos de teste estatístico:

• Métodos diretos: consistem em otimizar um critério, como a somas de erros quadrados dentro do *cluster* ou a média silhouette. Os métodos correspondentes são denominados métodos de *Elbow* e silhouette, respetivamente.

• Métodos de teste estatístico: consiste em comparar evidências contra hipóteses nulas. Um exemplo é a estatística de gap.

É importante referir ainda que, a ideia básica por detrás dos métodos de particionamento, como o *k-means clustering*, é definir *clusters* de forma que a variação total intra-*cluster*, ou a soma total quadrada dentro do *cluster* (WSS), seja minimizada.

O WSS, *Within-cluster sum of square* total, soma de quadrados dentro do *cluster*, mede a compactação do cluster e queremos que esta seja o menor possível.

4.2 Soluções

Uma solução simples e popular consiste em inspecionar o dendrograma (O QUE É UM 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 métodos de teste estatístico:

• Métodos diretos: consistem em otimizar um critério, como a somas de erros quadrados dentro do *cluster* ou a média silhouette. Os métodos correspondentes são denominados métodos de *Elbow* e silhouette, respetivamente.

• Métodos de teste estatístico: consiste em comparar evidências contra hipóteses nulas. Um exemplo é a estatística de gap.

É importante referir ainda que, a ideia básica por detrás dos métodos de particionamento, como o *k-means clustering*, é definir *clusters* de forma que a variação total

intra-*cluster*, ou a soma total quadrada dentro do *cluster* (WSS), seja minimizada.

O WSS, *Within-cluster sum of square*(METER TRADUÇÃO) total mede a compactação do cluster e queremos que esta seja o menor possível.

4.3 The Elbow Method

O método de Elbow considera o WSS total como uma função do número de *clusters*: deve-se escolher um número de *clusters* para que a adição de outro *clusters* não melhore muito mais o WSS total.
O número ótimo de *clusters* pode ser definido da seguinte forma:

1. Calcular o algoritmo de *clustering*, por exemplo, *k-means clustering*, para diferentes valores de *k*. Por exemplo, variando *k* de 1 a 10 clusters;

2. Para cada *k*, calcular a soma total quadrada (WSS) dentro do *clusters*;

3. Fazer o gráfico (curva) de wws de acordo com o número de *clusters k*;

4. A localização de uma curva, curva joelho, provavelmente, é uma curva com uma dobra acentuada, é geralmente considerada um indicador do número apropriado de *clusters*

Note-se que, às vezes, o método de *Elbow* é ambíguo. Uma alternativa é o método de *silhouette média* (*Kaufman e Rousseeuw* [1990]), que também pode ser usado com qualquer abordagem de *clustering*.

4.3.1 K-Means

O *K-Means* é um algoritmo de *clustering* muito comum e popular usado por muitos investigadores em todo o mundo. Ao usar o algoritmo *K-Means*, distintamente de algoritmos como o DBSCAN (O QUE É O DBSCAN??), deve-se sempre especificar o número de *clusters* nos quais é necessário um conjunto de dados em *clusters*. Portanto, a maneira mais fácil de fazer isto, é usando o método de *Elbow*.

Na maioria das vezes, o método *Elbow* é usado ou com a soma de erros quadrados (sse) ou com a soma dos erros do *cluster* (wcss) (EXPLICAR O QUE CADA UM É!). Neste exemplo, irá ser usado o wcss para encontrar o número ideal de *clusters*.

```
from sklearn.cluster import KMeans
        for i in range(1, 11):
                kmeans = KMeans(n_clusters = i, init = 'k-means++',
                                                random_state = 42)
                kmeans.fit(X)
                wcss.append(kmeans.inertia_)
```

Figura 4.1: Código para obtenção de valor wcss

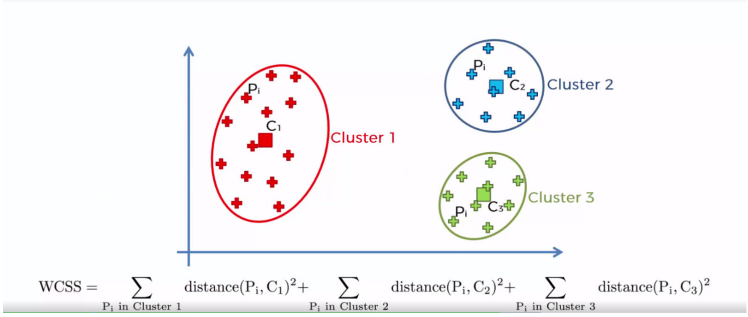


Figura 4.2: Cálculo do valor wcss para três conjuntos de dados de *cluster*

O algoritmo de *clustering K-Means* é um algoritmo popular que se enquadra nesta categoria. Nestes modelos, os números de *clusters* necessários no final têm de ser mencionados com antecedência, o que torna importante o conhecimento prévio do conjunto de dados. Estes modelos são executados iterativamente para encontrar o local ótimo.

4.3.2 WCSS

O código abaixo é uma maneira fácil de obter o valor wcss para diferentes números de *clusters*.

Assim como o nome sugere, wcss é o somatório da distância de cada *cluster* entre esses *clusters* específicos e cada um dos pontos contra o centróide do *cluster*.

Na imagem abaixo, é possível entender como calcular o valor wcss para três conjuntos de dados de *cluster*.

Portanto, se confrontarmos o valor de wcss com o número de *clusters* que tentamos obter, esse valor, normalmente obtemos um gráfico semelhante ao abaixo.

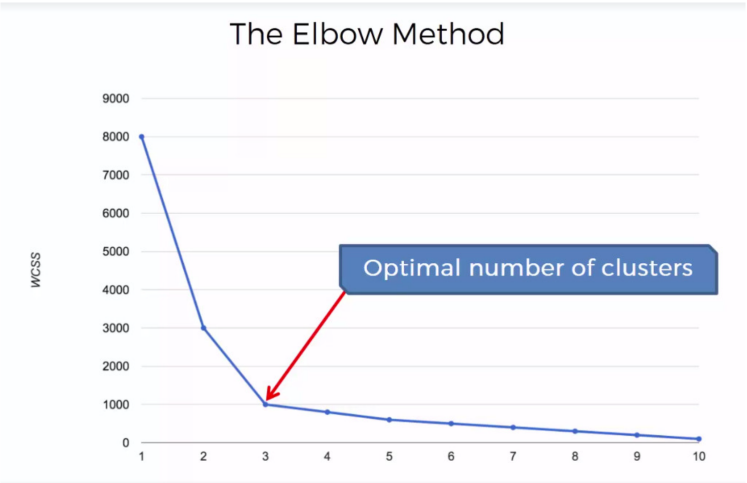


Figura 4.3: Valor de wcss *versus* número de *clusters*

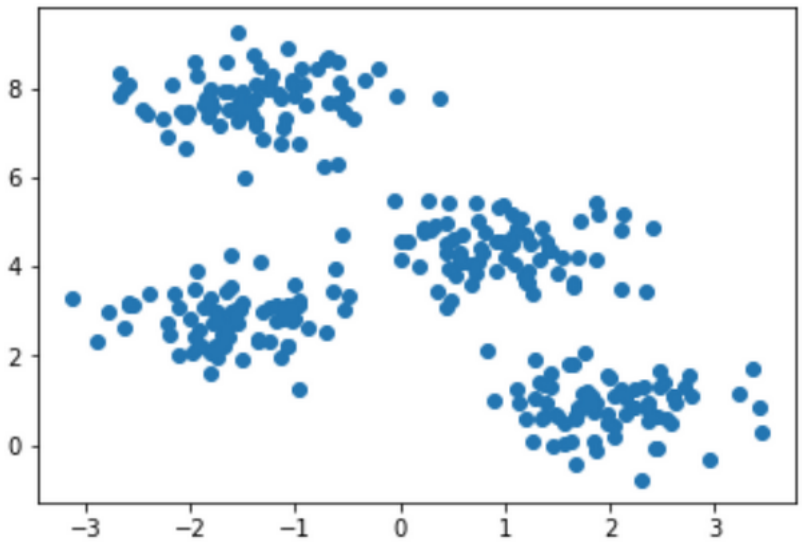
5 | Aplicações práticas

5.1 Exemplo 1

5.1.1 Como aplicar Clustering:

Para aplicar o algoritmo, precisamos de primeiro criar alguns conjuntos aleatórios de pontos e distribui-los com algum espaçamento.

```
points = make_blobs(n_samples=300, centers=4, cluster_std=0.60, random_s
points.scatter(distance=1.5);
```



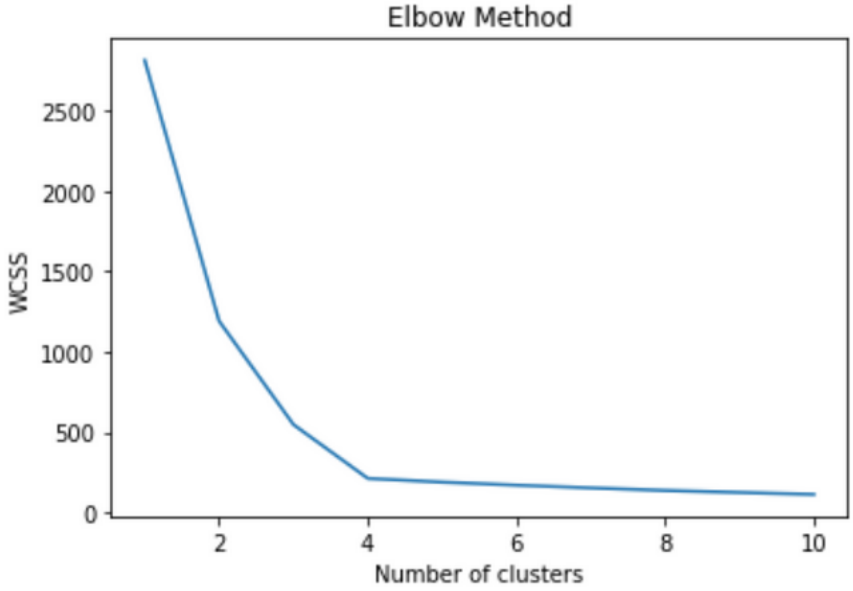
De seguida vamos aplicar kmeans aos nossos pontos. Vamos aplicar a função várias vezes, para numeros de clusters desde 1 até 9 e vamos guardar o valor de WCSS de cada resultado.

```
int wcss[10];

for(int i=1; i<10; i+=1) {
        kmeans = points.KMeans(n_clusters=i, init="k-means++", max_iter=3
```

```
        wcss[i] = kmeans.getWCSS();
}
```

Com os falores de WCSS obtidos podemos gerar um gráfico que os relaciona com o respetivo numero de clusters.



6 | Conlusões

Bibliografia

[1] *What is "Within cluster sum of squares by cluster" in K-means*

https://discuss.analyticsvidhya.com/t/what-is-within-cluster-sum-of-squares-by-cluster-in-k-means/2706

[2] *Elbow Method*,

https://www.scikit-yb.org/en/latest/api/cluster/elbow.html

[3] *Determining the optimal number of clusters*,

https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/#elbow-method

[4] *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

[5] Lachi, Ricardo Luís & Rocha, Heloísa Vieira da. Fevereiro 2005. *Aspectos básicos de clustering: conceitos e técnicas* . (Brasil).

[6] https://www.maxwell.vrac.puc-rio.br/24787/24787_5.PDF

[7] http://www.dei.isep.ipp.pt/~paf/proj/Julho2003/Clustering.pdf

[8] Hierarchical Clustering 3: single-link vs. complete-link https://www.youtube.com/watch?v=VMyXc3SiEqs