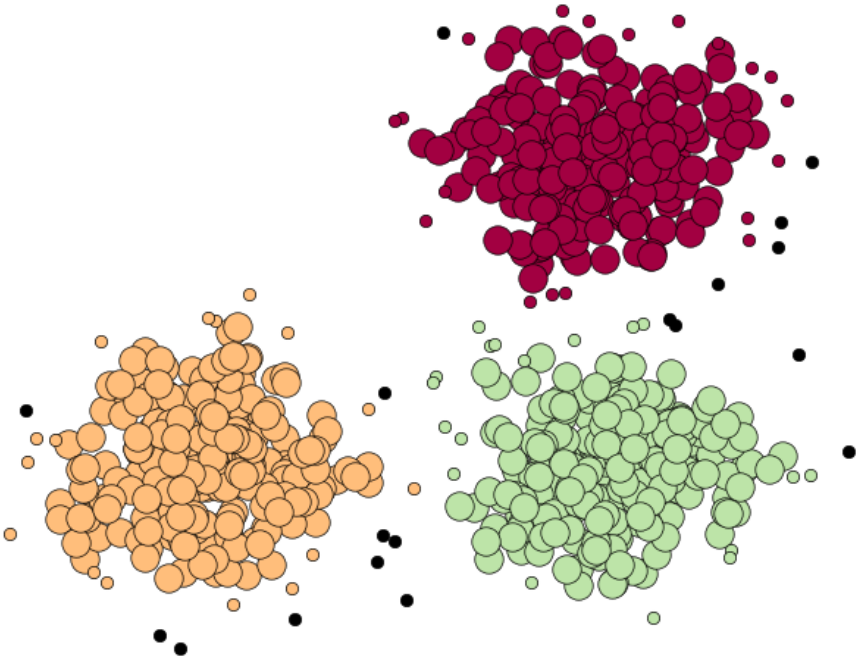
Universidade do Minho

31 de Março de 2020

The Elbow Method

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

Figuras/Gráficos

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 será feita uma contextualização abordando a definição de clusters no geral e, mais em concreto, de clusters particionais.

Seguidamente, é descrito o conceito de centróides, bem como outros aspetos relevantes relativos ao mesmo.

Depois é abordado o *The Elbow Method*, com a respetiva definição teórica e a sua aplicação mais prática.

Na parte seguinte, a contextualização teórica é aplicada em exemplos mais práticos, por forma a melhor entender a aplicação dos tópicos acima referidos.

Para finalizar, através das etapas atrás explicitadas expomos uma breve conclusão do trabalho apontando os aspetos mais enriquecedores para o nosso conhecimento.

2 | Clusters[6][7]

2.1 O que são?

Um cluster é um conjunto de objetos similares entre si dentro do mesmo cluster 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.

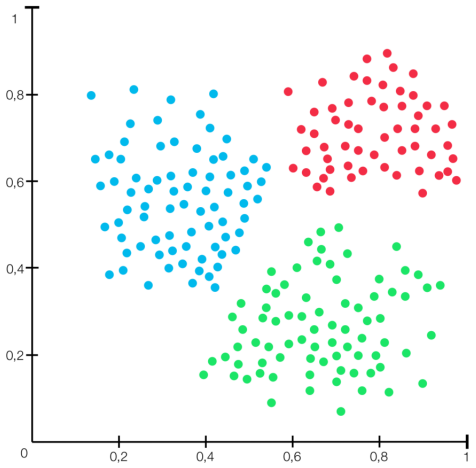


Figura 2.1: Figura ilustrativa do agrupamento de clusters.

Muito cedo nas escolas, nos primeiros anos de educação as crianças aprendem a classificar e agrupar, por exemplo distinguir entre gatos e cães, entre animais e plantas, progredindo num refinamento de classificação que tem subjacente teorias de clustering. A análise de clusters tem sido usada em inúmeras aplicações, tais como reconhecimento de padrões na análise de dados, processamento de imagem e pesquisa de mercado, entre outras.

2.2 Clustering

O clustering é o conjunto de técnicas de prospeção de dados que faz 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

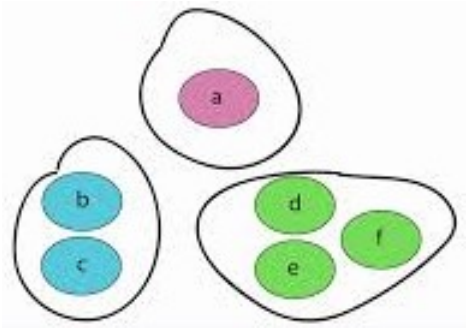


Figura 2.2: Clustering.

mais sofisticados pedem apenas o número
mínimo, outros tem a capacidade de subdividir um grupo em dois. Existem vários tipos de agrupamentos, mas os que nos interessam para este trabalho são 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

Contextualizando, o **centróide** diz respeito ao ponto que representa o centro de todos os pontos pertencentes a um **cluster**.

Os modelos centróides, por seu lado, são algoritmos iterativos de *clustering* nos quais a noção de similaridade é derivada pela proximidade de um ponto de dados com o centróide dos *clusters* <- MUITO CONFUSO.

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 os ótimos locais.

Além disso, para a obtenção dos centróides é necessário fazer operações algébricas (somas e multiplicações por escalares), sendo que, em regra, esses mesmos centróides que são obtidos não pertencem à base de dados.

4 | *The Elbow Method*

4.1 qualquer coisa

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* (O QUE É K?).

Desta forma, o método pode ser considerado heurístico (DEFINIR HEURÍSTICA À PARTE), 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, são preferíveis outras abordagens para determinar o número de *clusters*.

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*.

Determinar o número ótimo de *clusters* num conjunto de dados é uma questão fundamental (REPETICÃO DO MESMO) no *clustering* por particionamento, como o *cluster de k-means*, que requer que o usuário especifique o número de *clusters k* a serem gerados. (OUTRA VEZ REPETIÇÃO)

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

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

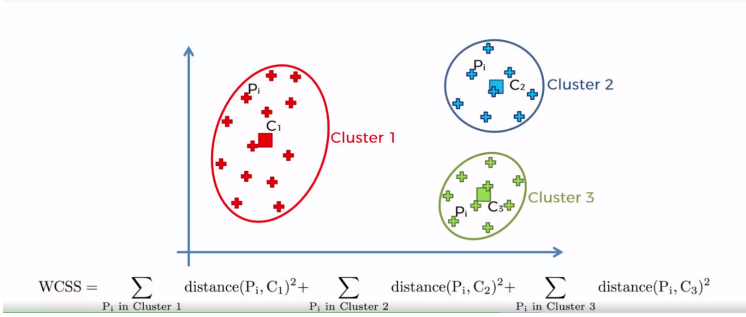


Figura 4.2: Cálculo do valor wcss para três conjuntos de dados de *cluster*

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.

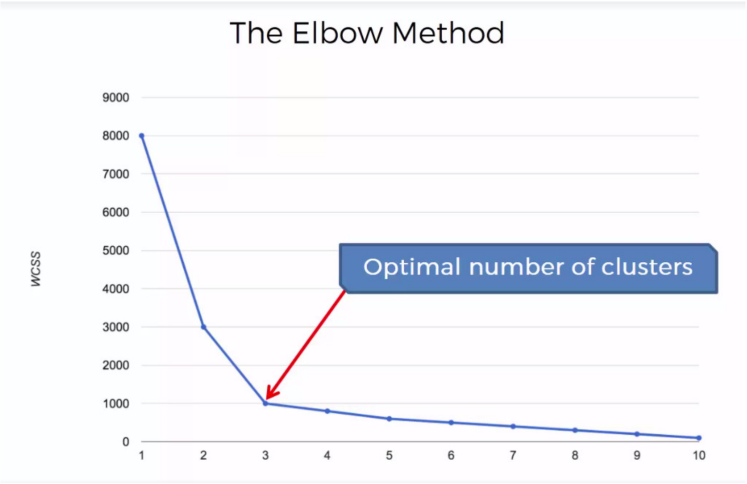


Figura 4.3: Valor de wcss *versus* número de *clusters*

5 | Aplicações práticas

5.1 Exemplo 1

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