

Agrupamento de dados com k-means

ESTAT0016 – Tópicos Especiais em Estatística (Introdução à Apredizagem de Máquina)

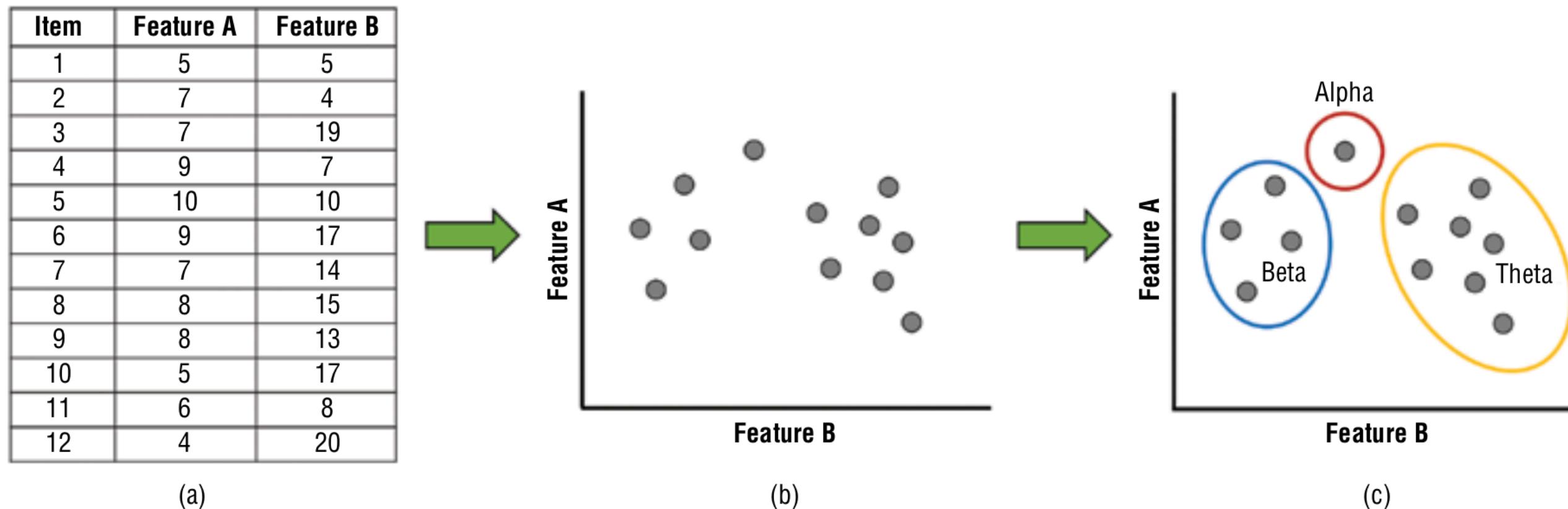
Prof. Dr. Sadraque E.F. Lucena

Agrupamento (*Clustering*)

- Agrupamento em aprendizagem de máquina consiste em métodos usados para partitionar dados não rotulados em *clusters* (subgrupos) baseados em similaridade.
- Os objetivos do agrupamento são:
 1. *Alta similaridade intraclasse*: itens de um mesmo cluster são o mais semelhantes quanto possível.
 2. *Baixa similaridade interclasse*: itens de classes diferentes são o mais diferentes quanto possível.
- O grau de similaridade entre itens se baseia em alguma medida de distância, como a distância euclidiana, que vimos no k-NN.

Agrupamento (*Clustering*)

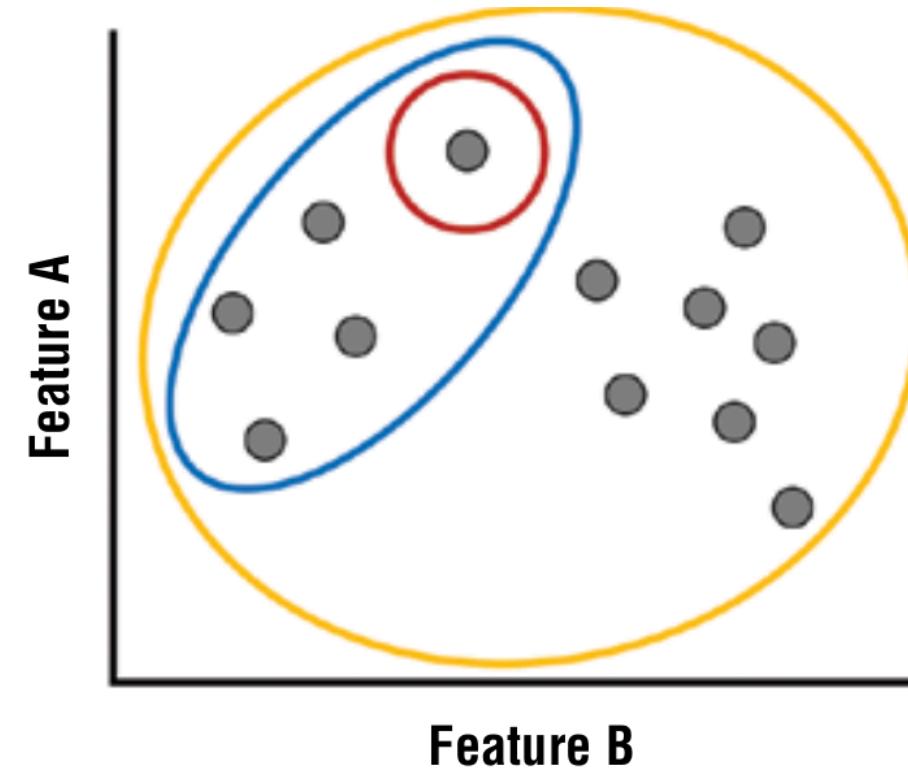
- *Clustering* é uma técnica não supervisionada que busca identificar padrões em dados não rotulados a partir de agrupamento.



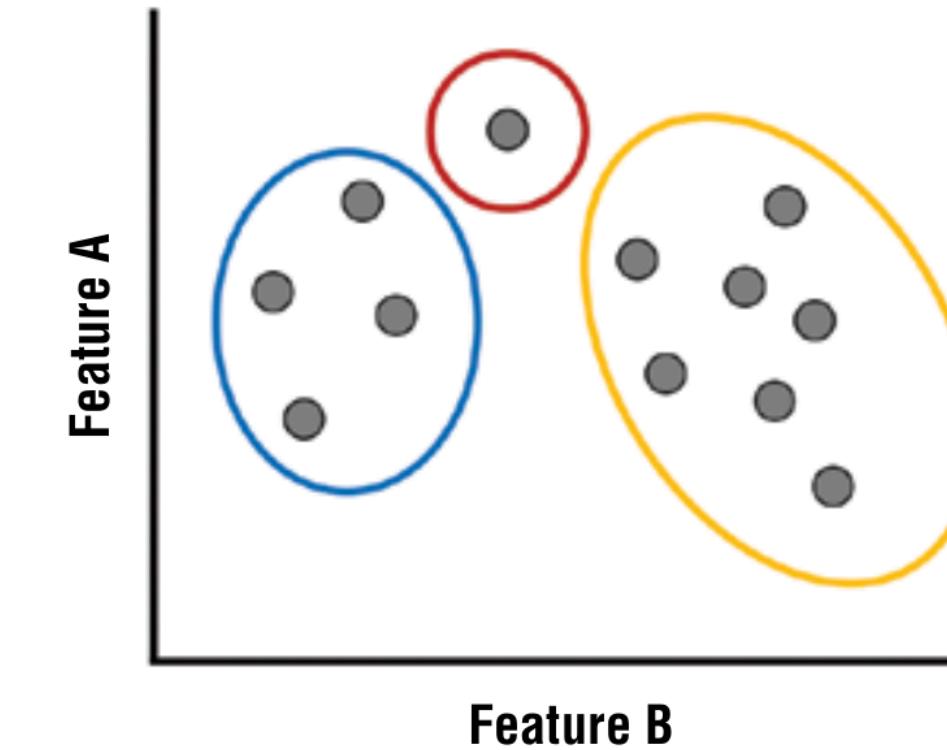
- Observação: no *clustering*, após criados os grupos, o usuário deve criar rótulos que descrevam o agrupamento gerado.

Agrupamento (*Clustering*)

- A clusterização pode ser do tipo:
 - *Hierárquica*: os *clusters* podem estar contidos dentro de outros *clusters*.
 - *Particionada*: o limite de cada cluster é independente do outro.



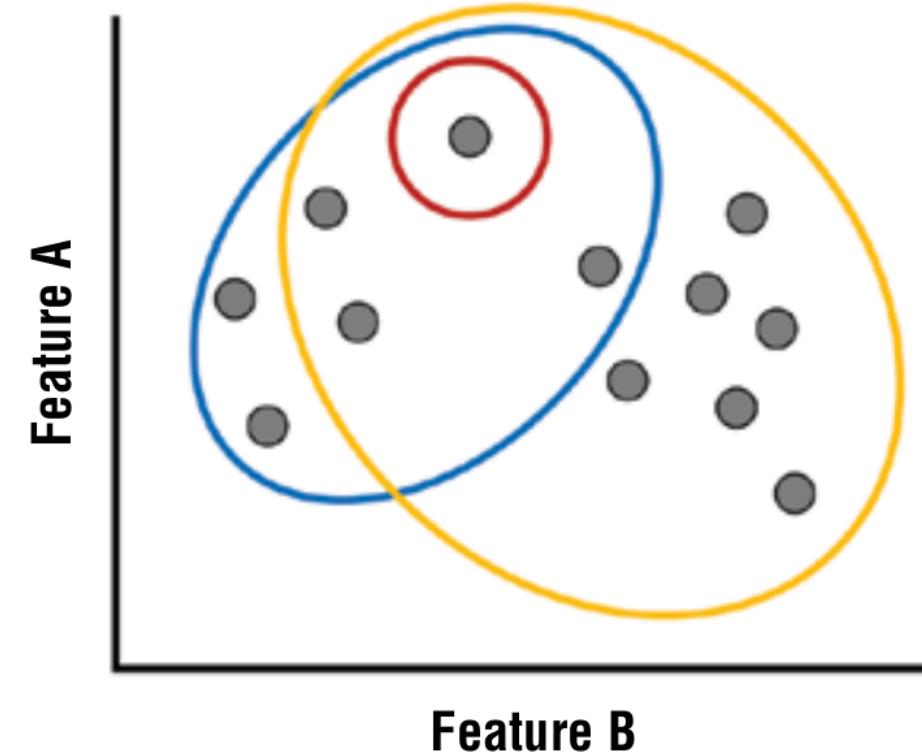
Hierarchical



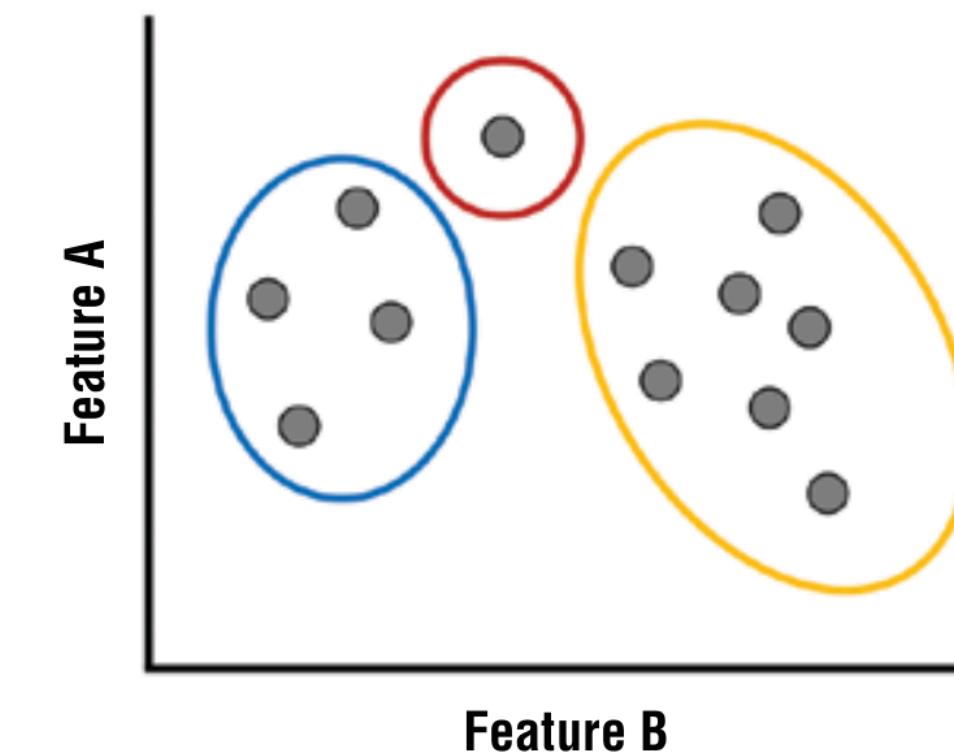
Partitional

Agrupamento (*Clustering*)

- A clusterização também pode ser classificada em:
 - *Sobreposta*: cada item pode pertencer a um ou mais *clusters*.
 - *Exclusiva*: um item pertence apenas a um único *cluster*.



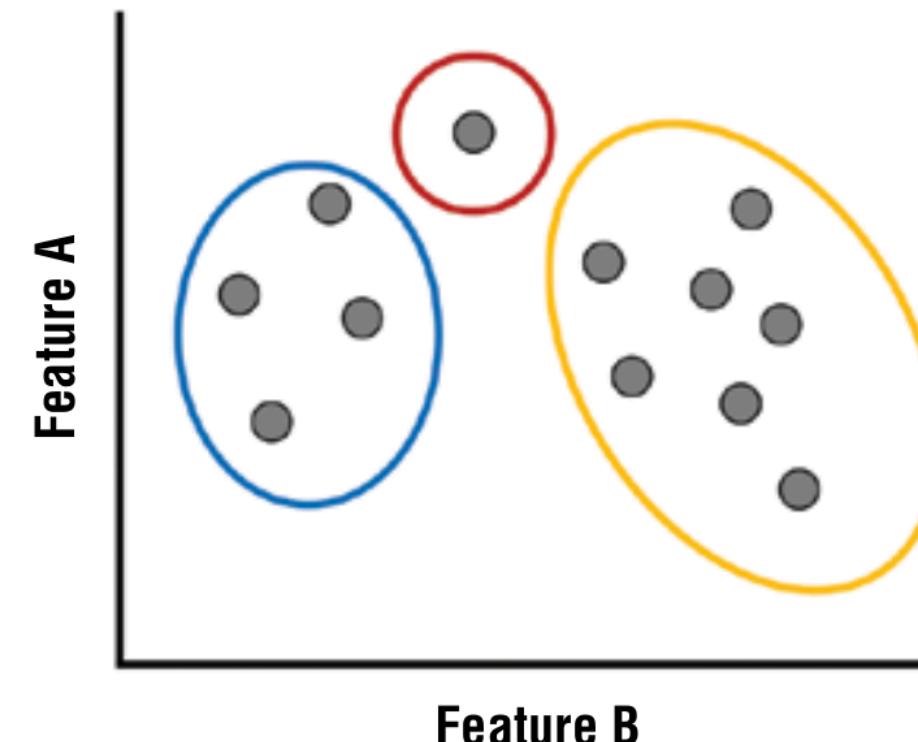
Overlapping



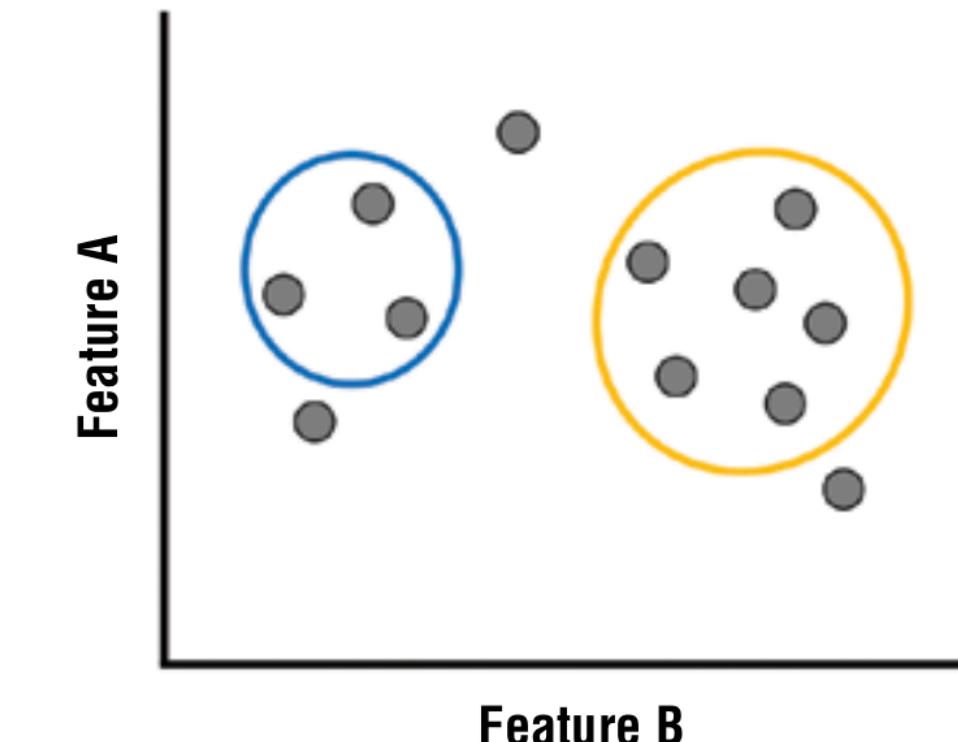
Exclusive

Agrupamento (*Clustering*)

- A clusterização também pode ser classificada em:
 - *Completa*: todos os itens devem pertencer a pelo menos um *cluster*.
 - *Parcial*: nem todos os itens devem pertencer a pelo menos um *cluster*.
 - Nesse caso, itens que não têm similaridade suficiente com outros (geralmente *outliers*) não são atribuídos a anenhum *cluster*.



Complete



Partial

Agrupamento k-Means

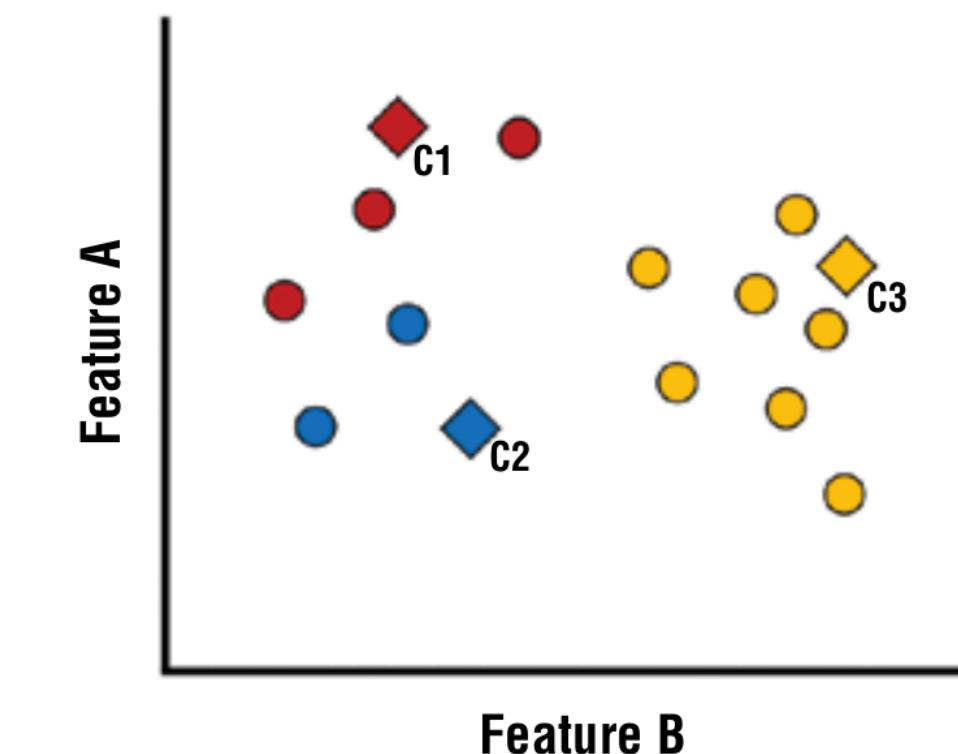
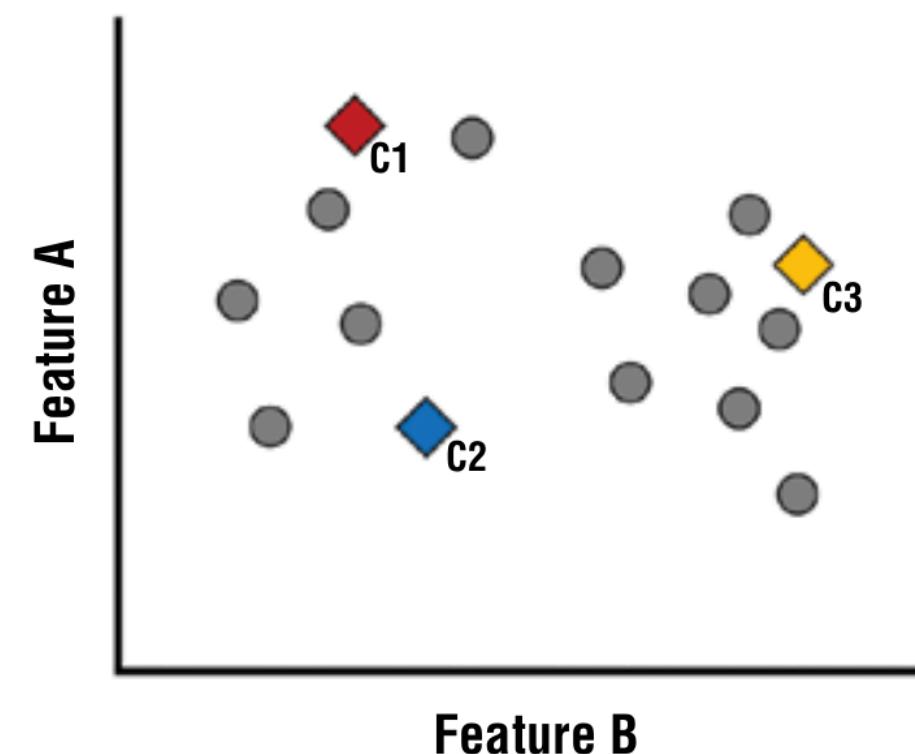
- Agrupamento k-means é uma das técnicas de agrupamento mais usadas.
- Esta abordagem é classificada como
 - Particionada, pois os limites dos *clusters* são independentes;
 - Exclusiva, pois cada item pertence a apenas um *cluster*;
 - Completa, pois todos os itens são atribuídos a um *cluster*.
- No agrupamento k-means o usuário define o número de *clusters* que o conjunto de dados terá. A partir daí, o algoritmo atribui cada item a um *cluster* de acordo com a similaridade entre os itens.

Agrupamento k-Means

Funcionamento

Suponha que vamos testar o agrupamento do conjunto de dados em três *clusters*. Isto é, $k = 3$.

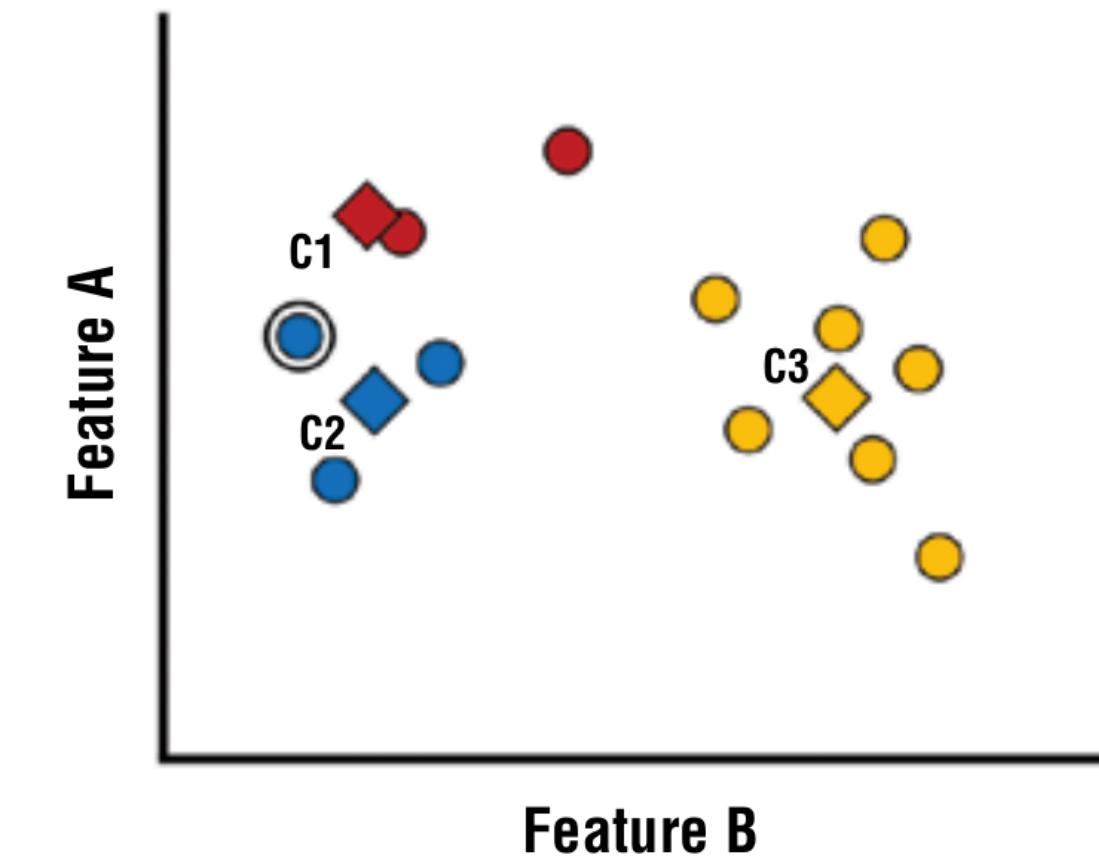
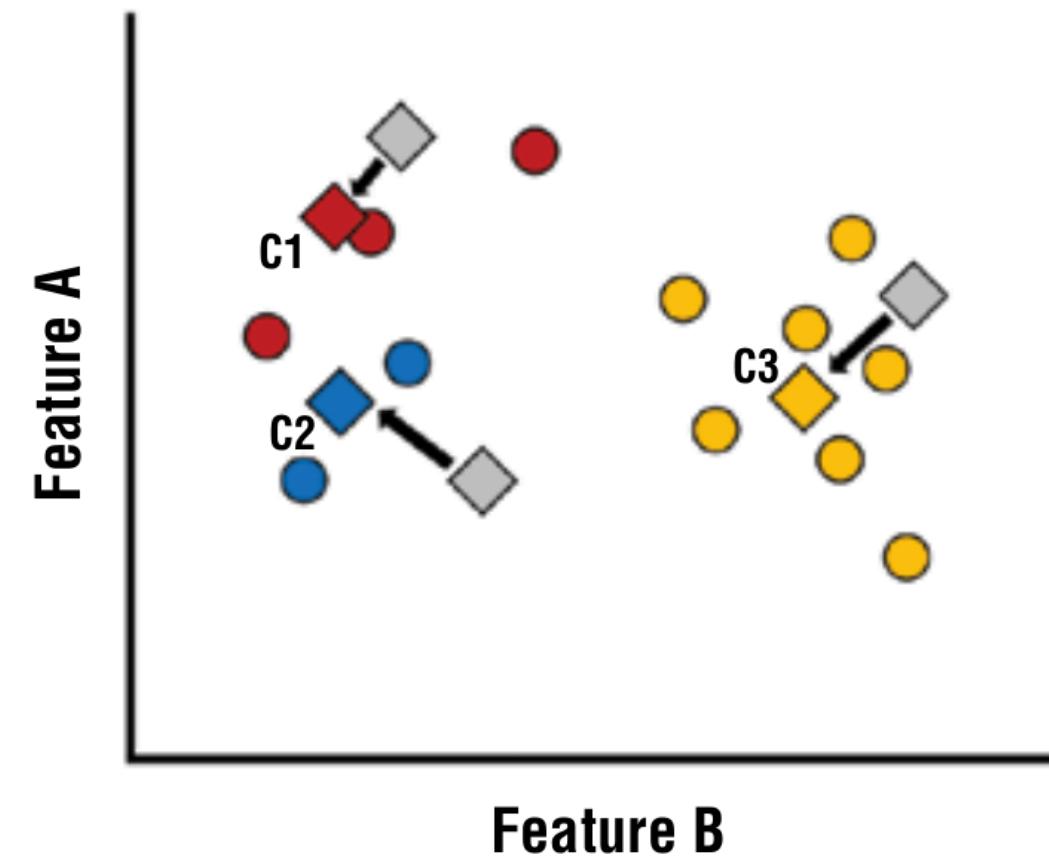
1. O algoritmo escolhe k pontos aleatórios que servem como centro dos *clusters* iniciais.
2. O algoritmo calcula a distância de cada item no conjunto de dados aos centros dos *clusters* e atribui o item ao *cluster* cujo centro está mais próximo. A medida de distância utilizada geralmente é a distância euclidiana.



Agrupamento k-Means

Funcionamento

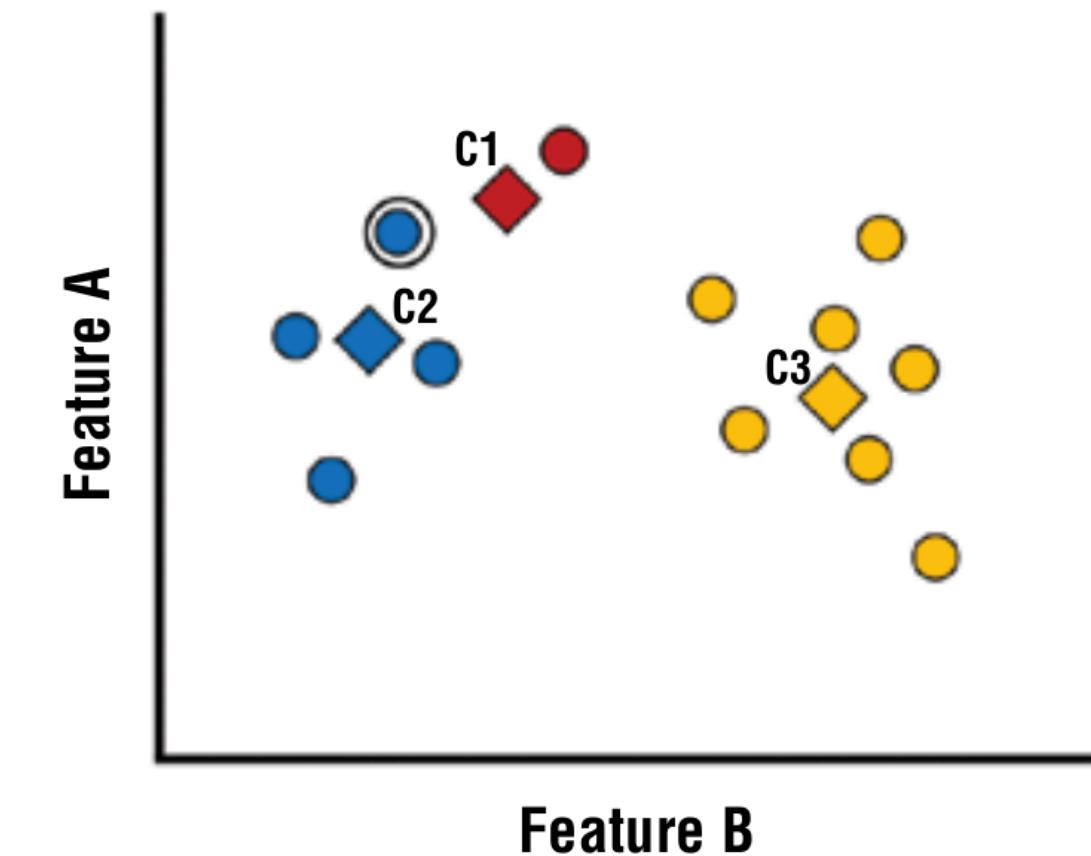
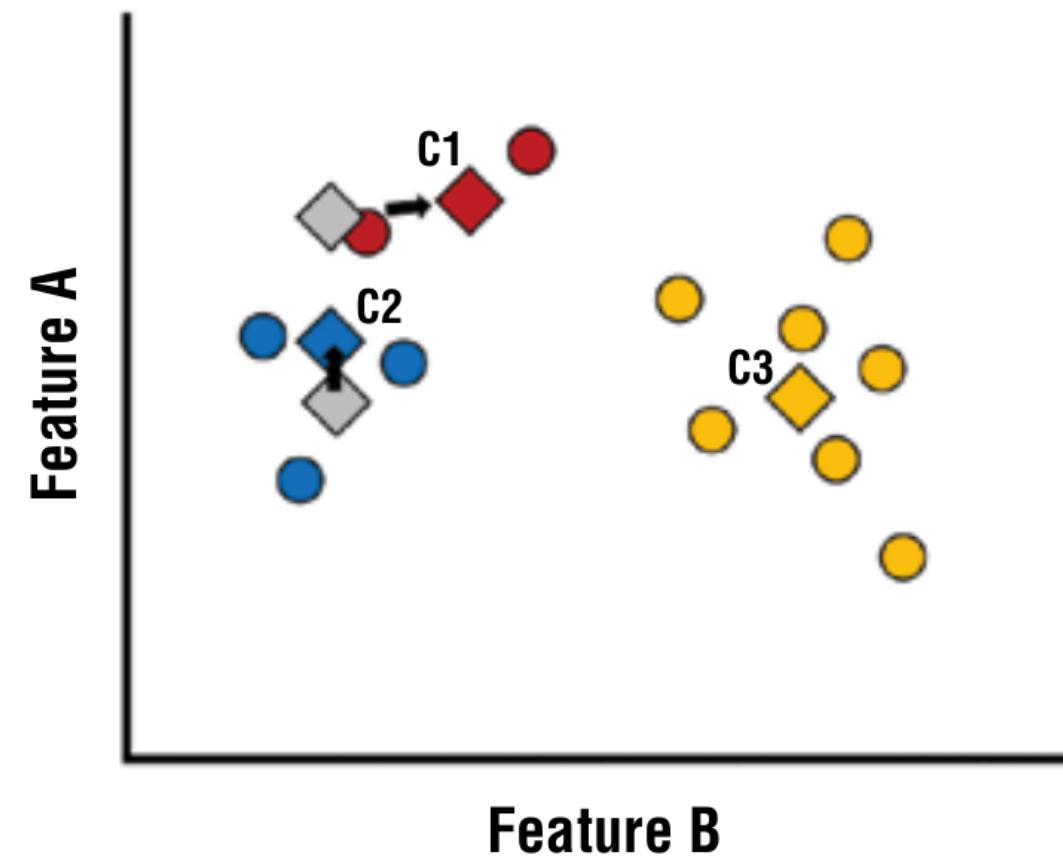
3. Após atribuir cada item a um *cluster*, o algoritmo calcula o centroide de cada *cluster* formado. O centroide é o ponto médio dos itens.
4. O algoritmo recalcula a distância de cada item do conjunto de dados a cada centroide e o reatribui ao cluster representado pelo centroide mais próximo.



Agrupamento k-Means

Funcionamento

5. O processo de atribuição e avaliação é repetido, com novos centroides calculados para cada *cluster* e cada item é atribuído ao cluster mais próximo baseado na distância ao centroide.

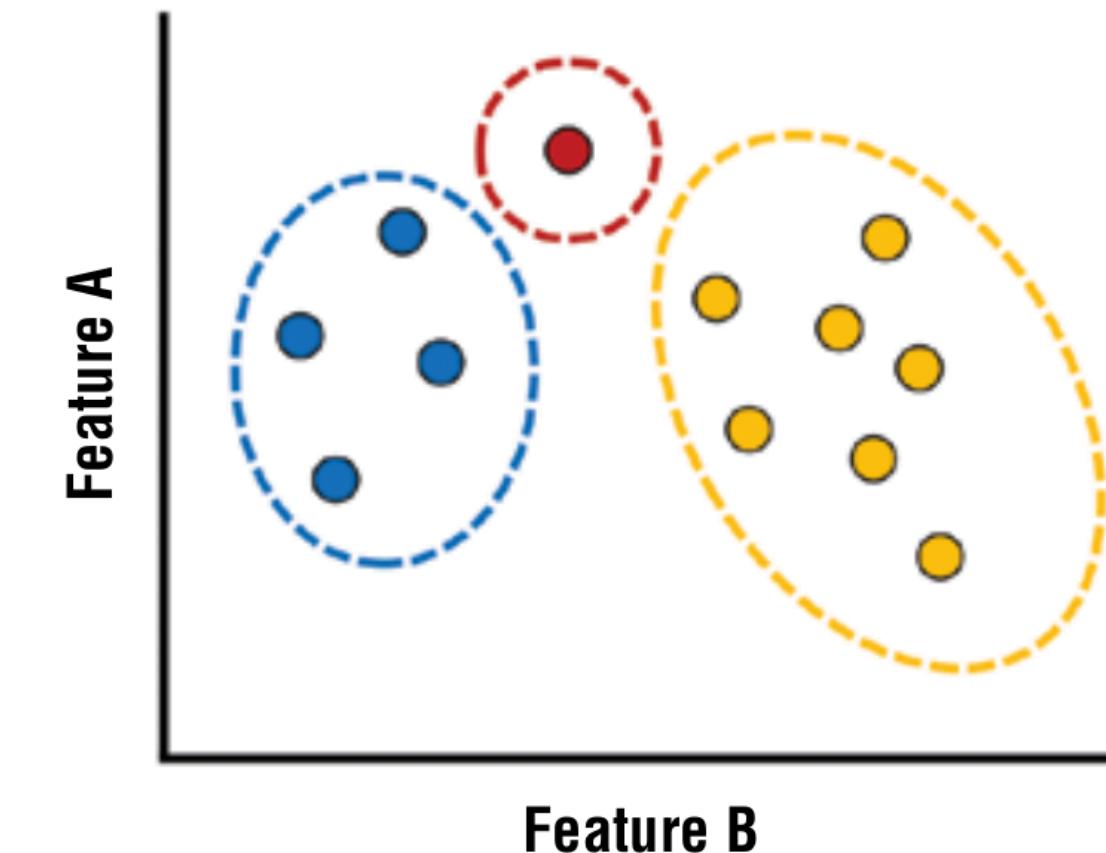
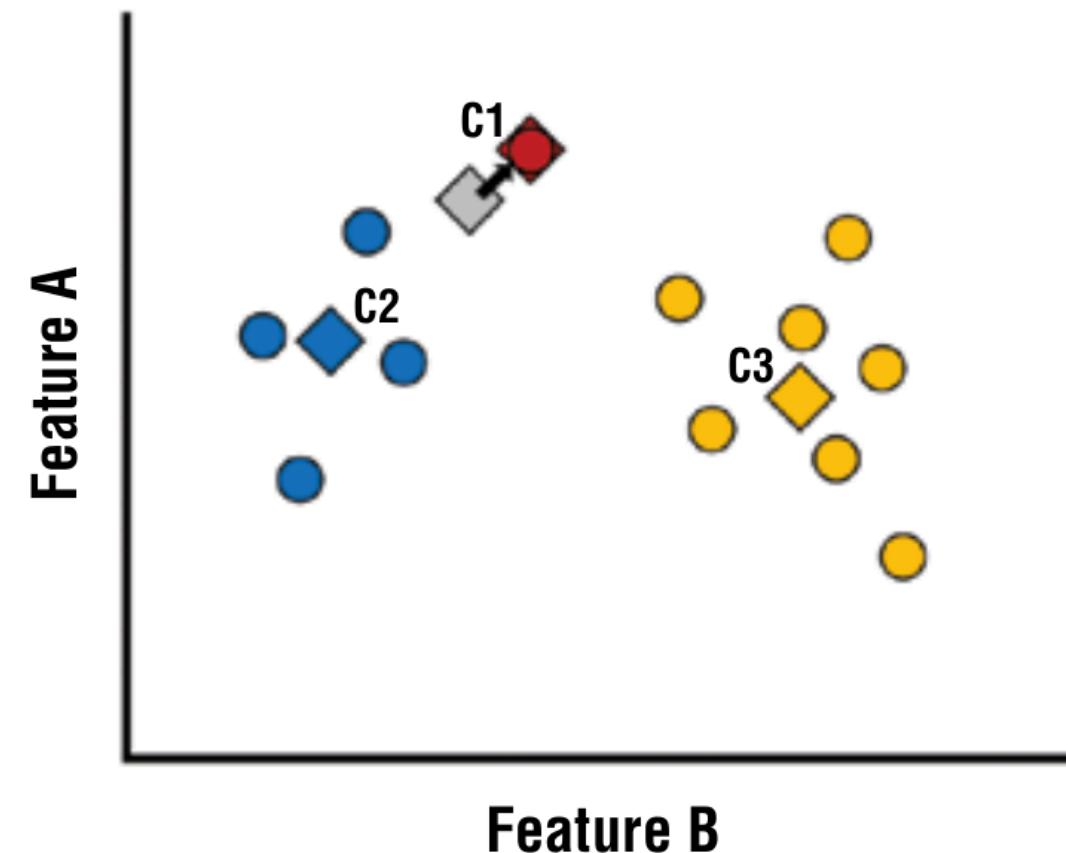


Agrupamento k-Means

Funcionamento

6. Em algum momento os centroides não mudarão muito de lugar e não resultarão em mudanças nas atribuições dos itens aos *clusters*.

- Nesse ponto dizemos que o algoritmo convergiu e o processo é interrompido.



Agrupamento k-Means

Distância euclidiana

- Para dois pontos p e q , com respectivos atributos p_1, p_2, q_1 e q_2 , a distância euclidiana é dada por

$$dist(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$

Centroide

- Assumindo que temos um cluster composto por três itens x, y e z , representados pelos pontos x_1, x_2, y_1, y_2 e z_1, z_2 , respectivamente, o centroide do cluster é calculado da seguinte forma:

$$centroide(x, y, z) = \left(\frac{x_1 + y_1 + z_1}{3}, \frac{x_2 + y_2 + z_2}{3} \right)$$

Agrupamento k-Means

Nota

A medida de distância utilizada para o agrupamento pode influenciar os *clusters* obtidos. Outras medidas que podem ser usadas incluem:

- distância de Manhattan;
- distância de correlação de Pearson;
- distância de correlação de Spearman;
- distância de correlação de Kendall.

Importante

- Os pontos centroides iniciais do algoritmo não precisam representar pontos que existem no conjunto de dados.
- Como os pontos iniciais são aleatórios, eles podem influenciar os *clusters* finais.
 - Se rodarmos k-means várias vezes, podemos obter clusters diferentes a cada vez.
 - Uma abordagem que tenta superar isso é conhecida como k-means++. A ideia é sempre escolher centros de *clusters* iniciais que sejam o mais distante quanto possível um dos outros.

Escolhendo o número certo de clusters

- Uma regra que pode ser utilizada para definir o número k de *clusters* a serem utilizados é a raiz quadrada do número de observações nos dados.
- O conhecimento sobre requisitos ou restrições de negócios pode determinar previamente o valor de k . Exemplos:
 - Se uma loja *on-line* tem três equipes de *marketing* especializadas em diferentes estratégias, pode ser benéfico ter apenas três clusters ($k = 3$).
 - Se a loja tem um orçamento restrito para segmentação e campanhas de *marketing*, pode ser preferível limitar o número de clusters para otimizar os recursos disponíveis.
- Outra abordagem é utilizar métodos estatísticos. Veremos três deles:
 - Método Cotovelo;
 - Método da Silhueta Média;
 - Estatística Gap.

Método Cotovelo (*Elbow Method*)

- A ideia do método é testar vários valores de k e dizer qual deles é o número ótimo de *clusters*.
 - Quanto maior o número de *clusters*, menor a diferença entre eles e maior a diferença entre as observações intra-cluster.
 - Então é preciso encontrar um equilíbrio entre a diferença entre os *clusters* e a homogeneidade dentro de cada *cluster*.
- Para encontrar o k ótimo, usamos a soma dos quadrados intra-clusters (*within-clusters sum-of-squares*), que é a soma das distâncias entre os itens do *cluster* e seu centroide. Para $k = 3$, temos:

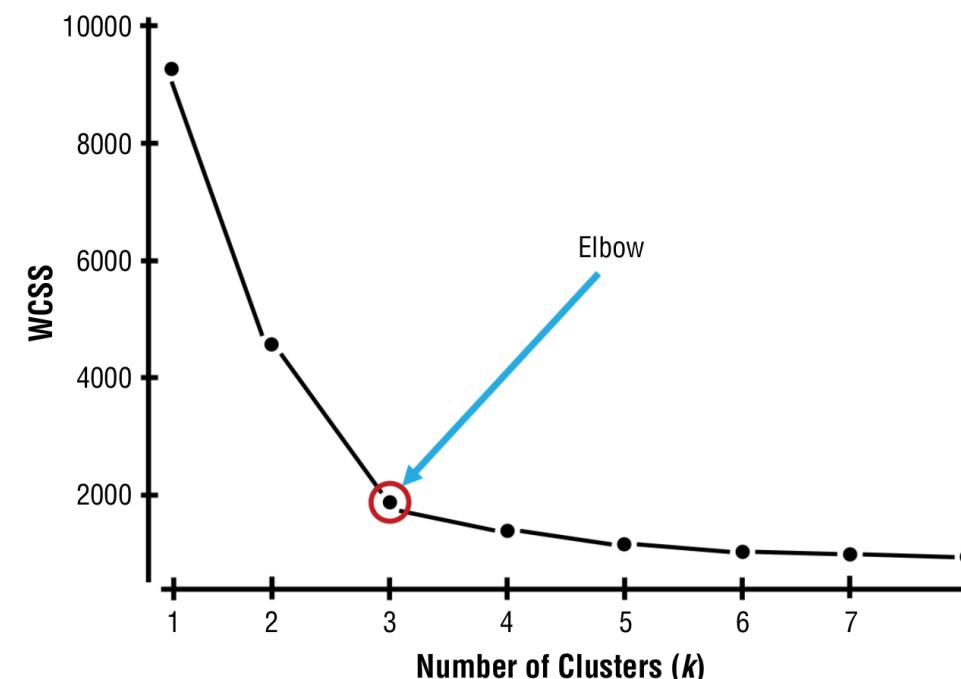
$$WCSS = \sum distancia(P_1, C_1)^2 + \sum distancia(P_2, C_2)^2 + \sum distancia(P_3, C_3)^2,$$

em que

- C_1, C_2 e C_3 representam os centroides dos *clusters* 1, 2 e 3, respectivamente;
- P_1, P_2 e P_3 representam os itens nos *clusters* 1, 2 e 3, respectivamente.

Método Cotovelo (*Elbow Method*)

- Quanto mais próximos os itens do *cluster* estiverem do centroide, menor o valor de $WCSS$.
 - Quanto menor o valor de $WCSS$, mais similares são os itens de um *cluster*.
- A medida que aumentamos o valor de k , mais próximo os itens de um cluster ficam de seu centroide e menor o valor de $WCSS$.
- Em algum ponto, o aumento no valor de k não reduz significativamente o valor de $WCSS$. Este ponto é chamado de cotovelo e o valor respectivo de k é utilizado.



Método da Silhueta Média (*Average Silhouette Method*)

- A silhueta é uma medida de quão semelhante um item é em relação aos itens do mesmo *cluster* comparado com os *clusters* vizinhos.
- A silhueta varia no intervalo $(-1, 1)$ de forma que
 - Quanto mais próximo de 1, mais o item está bem ajustado ao seu *cluster* (situação desejável).
 - Se o valor estiver próximo de 0, o objeto está próximo do limite entre dois *clusters*, indicando sobreposição ou ambiguidade na classificação.
 - Quanto mais próximo de -1 , o item pode ter sido atribuído ao cluster errado.
- Se a maioria dos itens têm valor alto, então a configuração do agrupamento é apropriada.
- Se muitos itens têm um valor baixo ou negativo, então a configuração de agrupamento pode ter muitos ou poucos *clusters*.

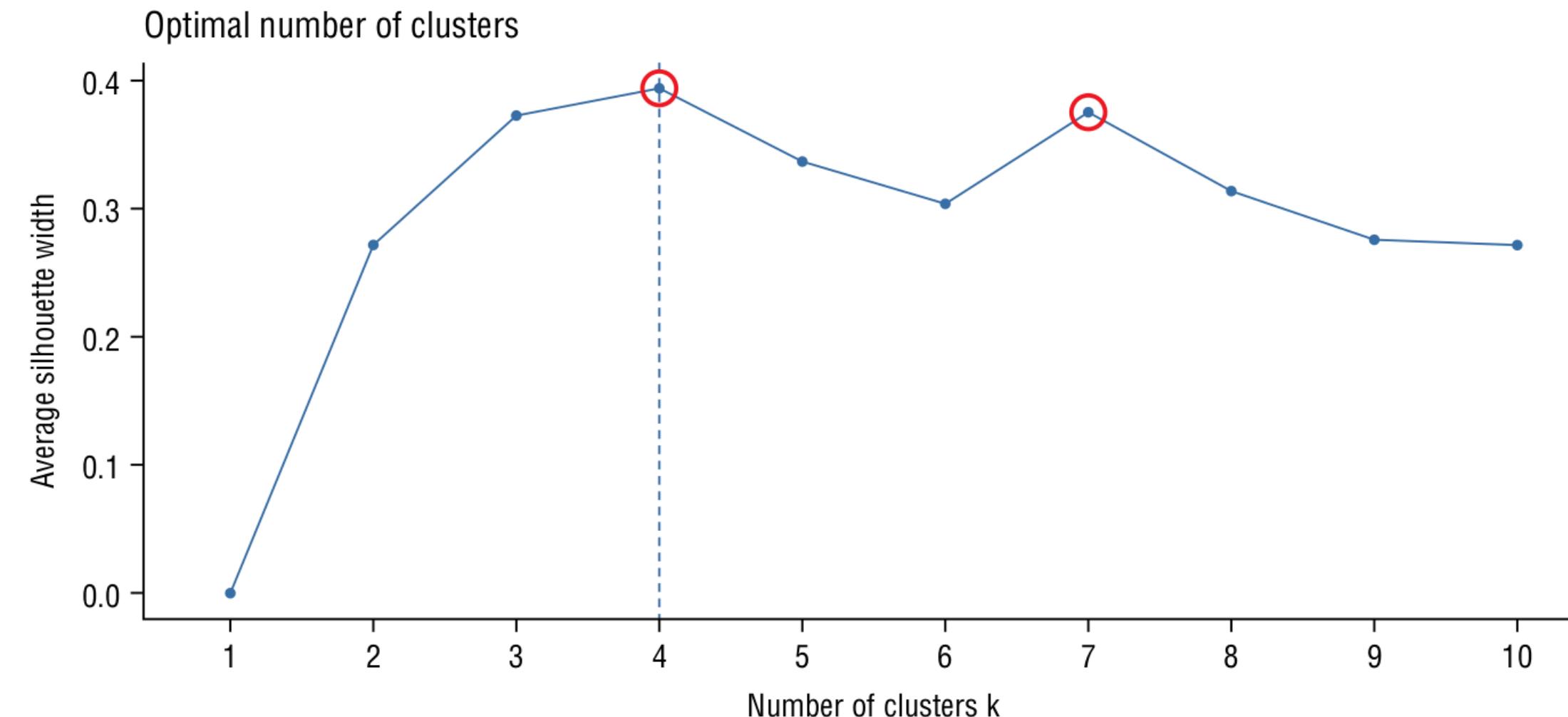
Método da Silhueta Média (*Average Silhouette Method*)

- Para o cálculo da silhueta:
 1. Calculamos a distância entre todos os itens usando alguma medida (por exemplo, a distância euclidiana).
 2. Calculamos:
 - $a(i)$: a distância média entre item i e outros itens do mesmo *cluster*;
 - $b(i)$: a distância média entre o item i e os itens dos demais *clusters*;
 3. A silhueta do item i então é

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

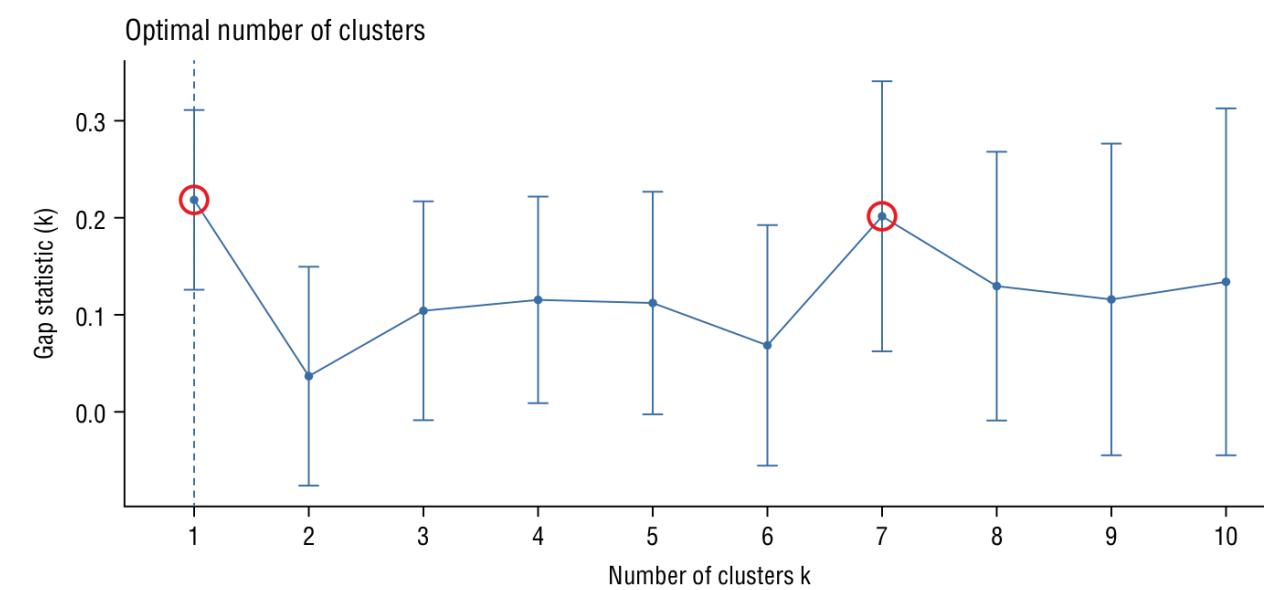
Método da Silhueta Média (*Average Silhouette Method*)

- Para diferentes valores de k é calculada a média da silhueta de todos os itens.
 - Se a média é alta, a configuração do *cluster* é adequada.
 - Se a média é baixa, a configuração do *cluster* é inadequada.



Estatística Gap

- Esta abordagem compara a dispersão dos itens dentro dos *clusters* com a dispersão esperada de um conjunto de dados aleatório (uniforme) com a mesma quantidade de itens, mas sem estrutura de *cluster*.
- Este conjunto de dados gerado aleatoriamente é chamado de *conjunto de dados de referência*.
- Para um dado k , a estatística gap é a diferença entre $WCSS$ para os dados observados e para os dados de referência.
 - O k ótimo é aquele que fornece maior distância entre os $WCSS$.



FIM

