Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5

Machine Learning

Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5

Data Science Academy

Seja muito bem-vindo(a)!

Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5

Data Science Academy

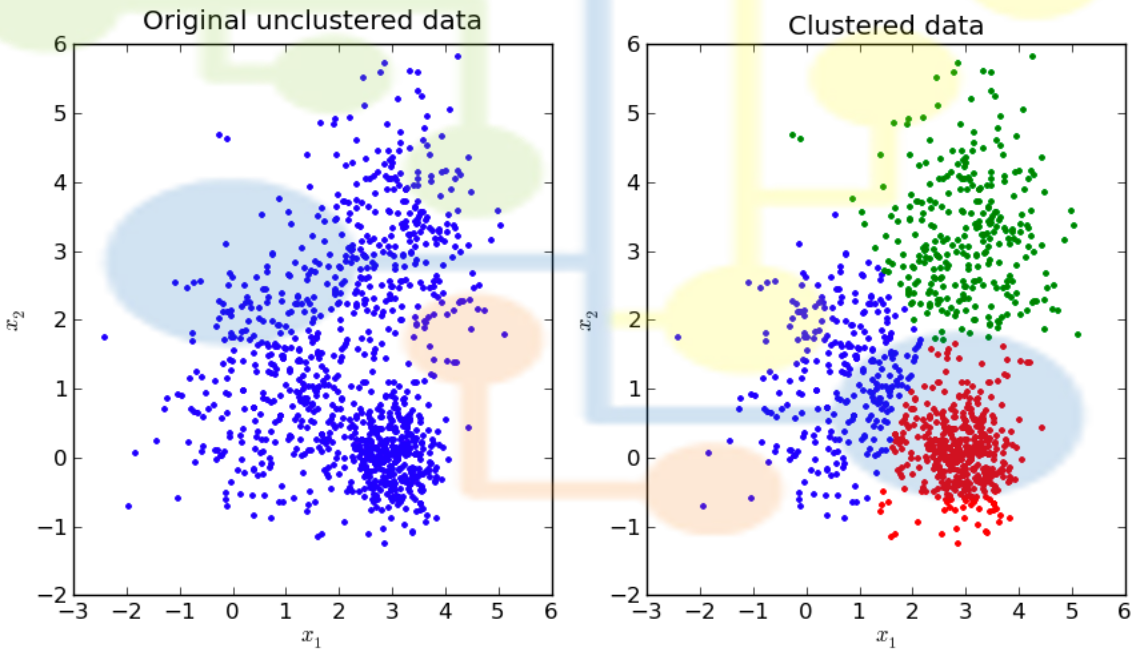
Clustering
(Clusterização ou Agrupamento)

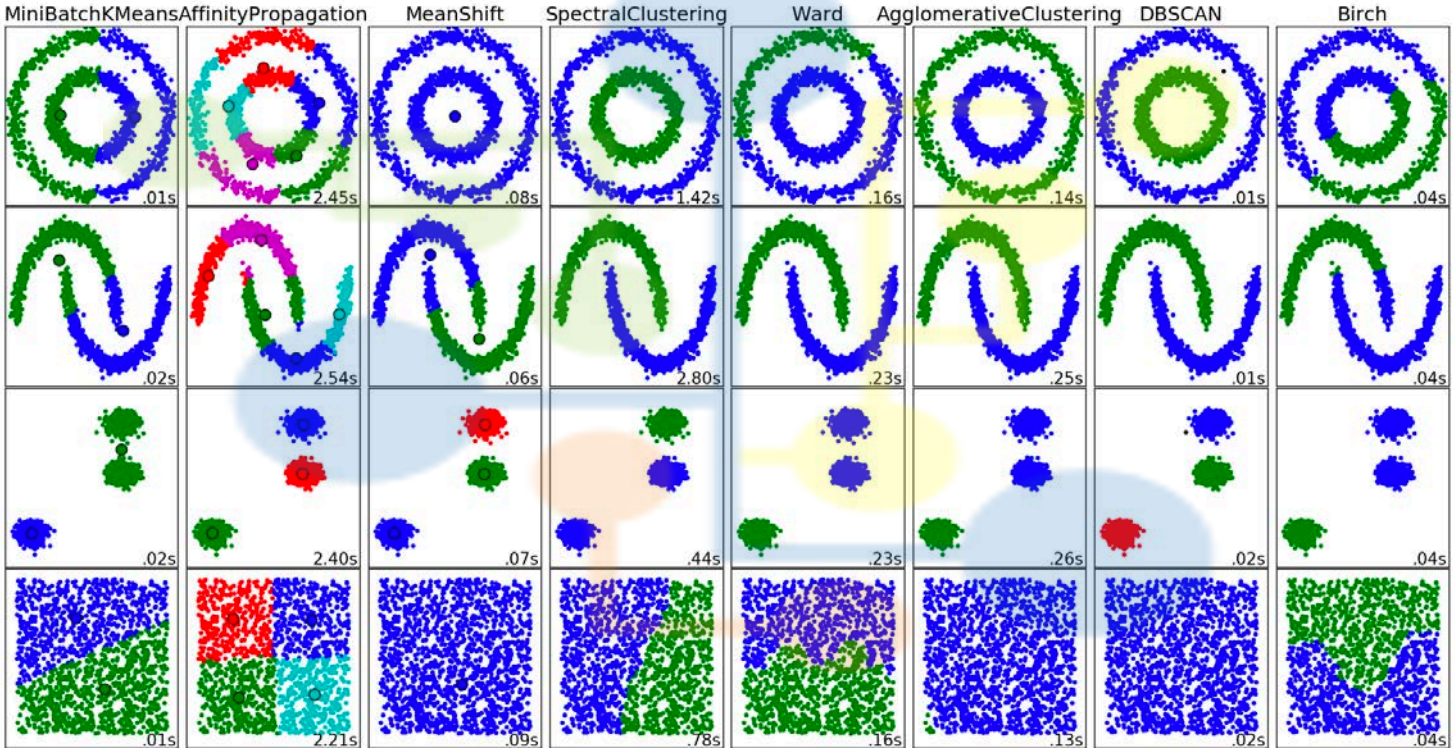
Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5

Clustering

Clustering ou Agrupamento é o processo de buscar padrões nos dados, não muito diferente do que é feito pelos seres humanos.

O principal objetivo da técnica de Clustering é encontrar grupos semelhantes ou homogêneos em dados, que são chamados clusters (ou grupos).


Original unclustered data
Clustered data
x_1
x_2
−3 −2 −1 0 1 2 3 4 5 6
−2 −1 0 1 2 3 4 5 6



http://scikit-learn.org/stable/modules/clustering.html#clustering

Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5

Machine Learning

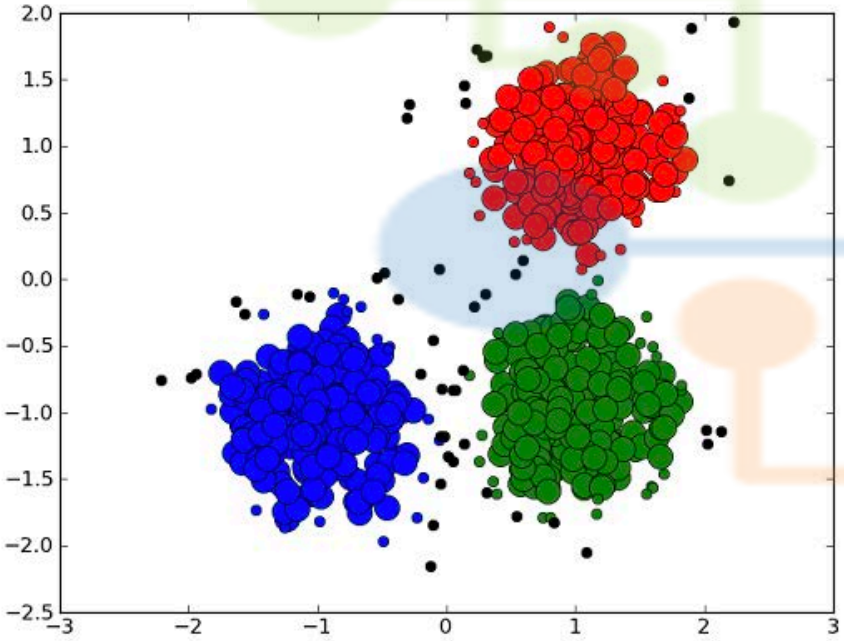
Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5

Data Science Academy

O Que é Clustering?

Clustering é uma atividade de aprendizagem não supervisionada que divide automaticamente os dados em clusters ou grupos de itens semelhantes.

O algoritmo faz essa classificação sem saber previamente como a classificação seria. Como nem sequer sabemos o que estamos procurando, o clustering é usado para a descoberta de conhecimento, ao invés de fazer previsões.



O agrupamento é guiado pelo princípio de que os itens dentro de um cluster devem ser muito semelhantes entre si, mas muito diferentes entre clusters.

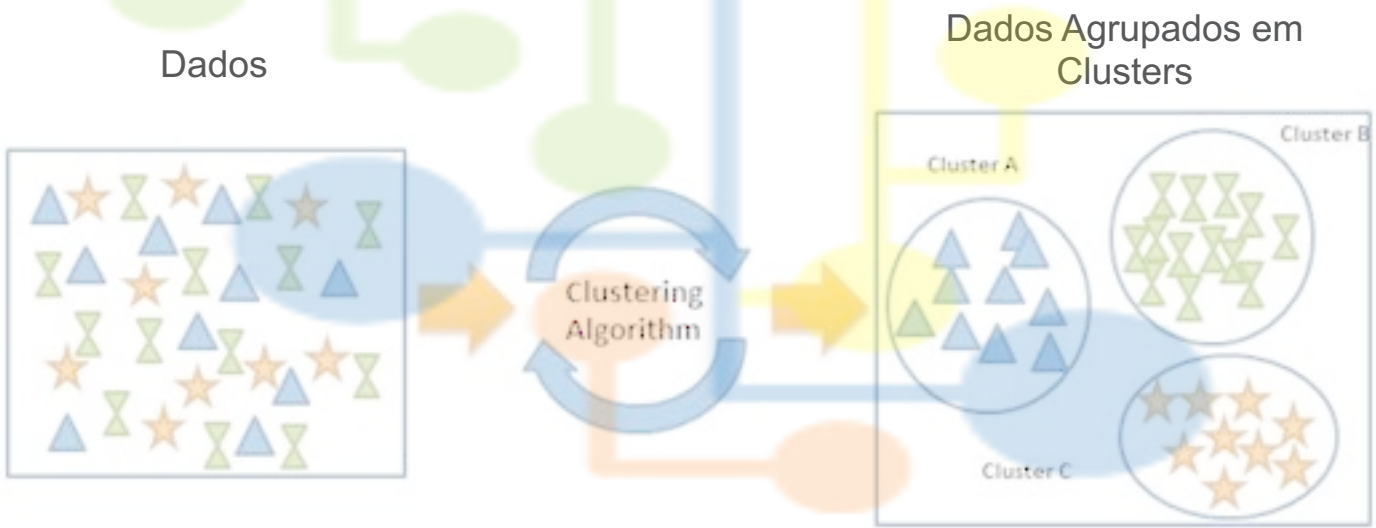
Quando Usamos Clustering?

- Segmentação de clientes em grupos com demografia semelhante ou padrões de compra para campanhas de marketing direcionadas.

- Detecção de comportamento anômalo, como intrusões de rede não autorizadas, identificando padrões de uso fora dos clusters conhecidos.

- Simplificação de grandes conjuntos de dados agrupando características com valores semelhantes em um número menor de categorias homogêneas.

O Processo de Agrupamento


Dados
Clustering Algorithm
Dados Agrupados em Clusters
Cluster A
Cluster B
Cluster C

Em uma k-clusterização (como o K-Means, por exemplo), o número total de diferentes formas de agrupamento de n elementos de um conjunto em k clusters, equivale à função N(n, k) abaixo:

$$N(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i \binom{k}{i}(k-i)^n$$

Outro aspecto a ser considerado em relação ao problema de clusterização é como medir o quanto um elemento é similar a outro e, assim, identificar se ambos devem estar contidos em um mesmo cluster ou não.

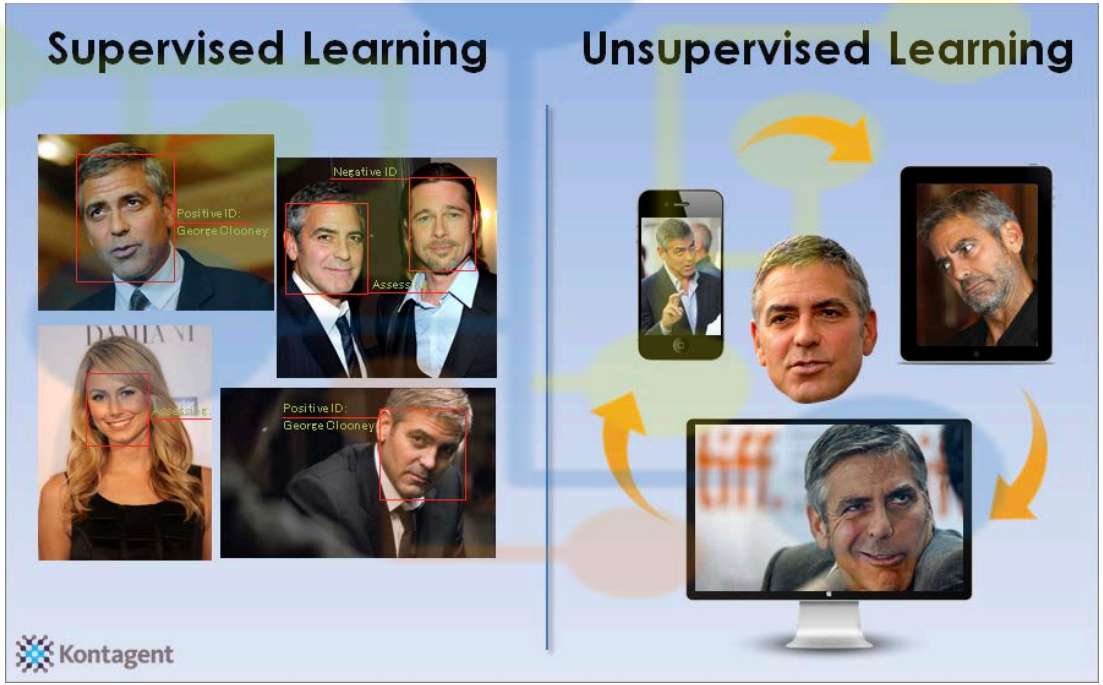
$$d(X_i, X_j) = \left[\sum_{l=1}^{p} (x_{il} - x_{jl})^2 \right]^{\frac{1}{2}}$$

Fórmula para cálculo da distância euclidiana

Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5

Quanto menor for a distância entre um par de elementos, maior é a similaridade entre eles.

Mas existem problemas de clusterização em que a distância não pode ser utilizada, ou não é conveniente que seja utilizada, como medida de similaridade, tendo em vista que os valores dos atributos não são escalares.


Supervised Learning
Unsupervised Learning
Positive ID:
George Clooney
Negative ID
Positive ID:
George Clooney
Kontagent

Os rótulos de classe obtidos de um classificador sem supervisão não possuem significado intrínseco.

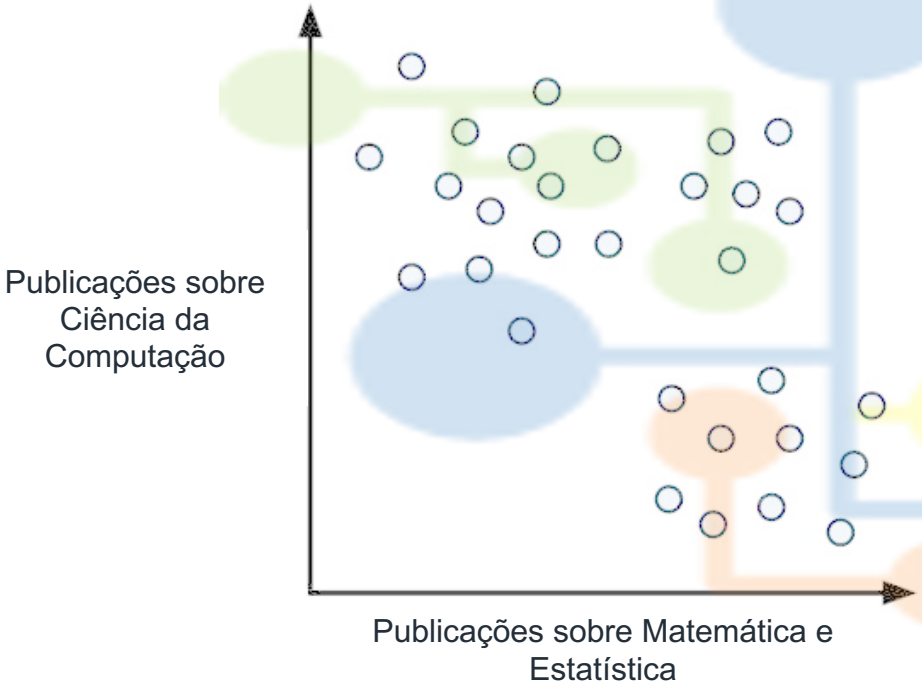
Exemplo de Agrupamento na Vida Real

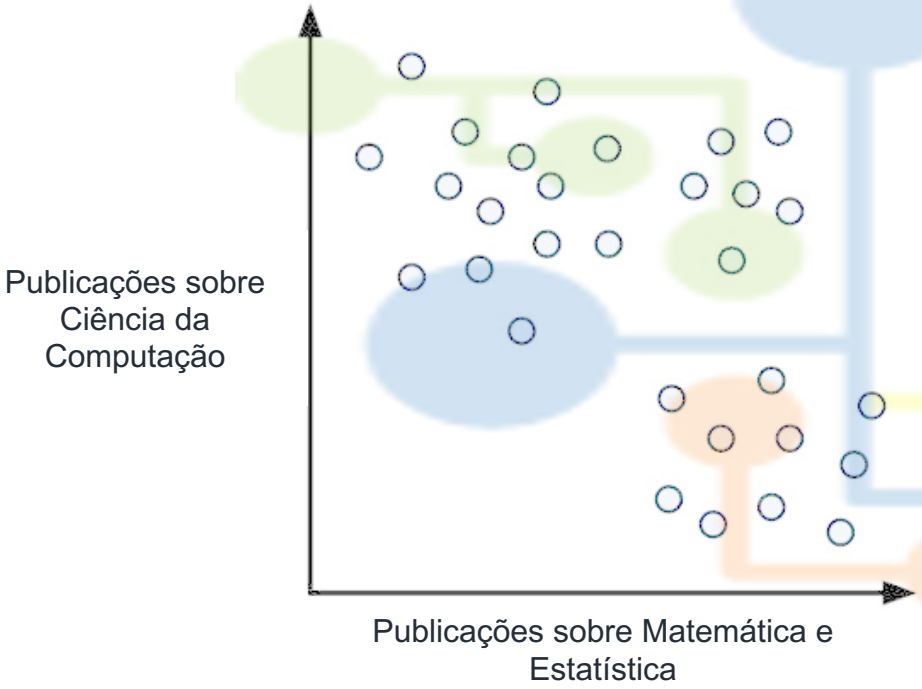
Ciência da Computação

Matemática e Estatística

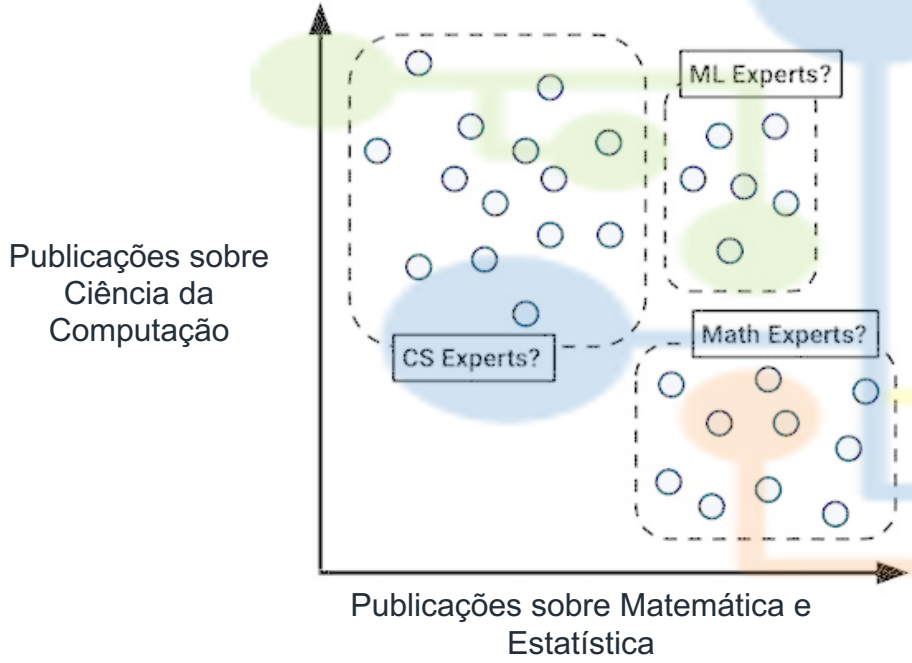
Machine Learning

Scatterplot




Publicações sobre Ciência da Computação
Publicações sobre Matemática e Estatística

Padrão



3 Grupos Criados a partir da similaridade encontrada nos dados

Se você começar com dados sem rótulo ou seja sem variável target, você pode usar o clustering para criar rótulos de classe. A partir daí, você pode aplicar um algoritmo supervisionado, como árvores de decisão para encontrar os preditores mais importantes dessas classes!

Isso é chamado de aprendizagem semi-supervisionada.

Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5

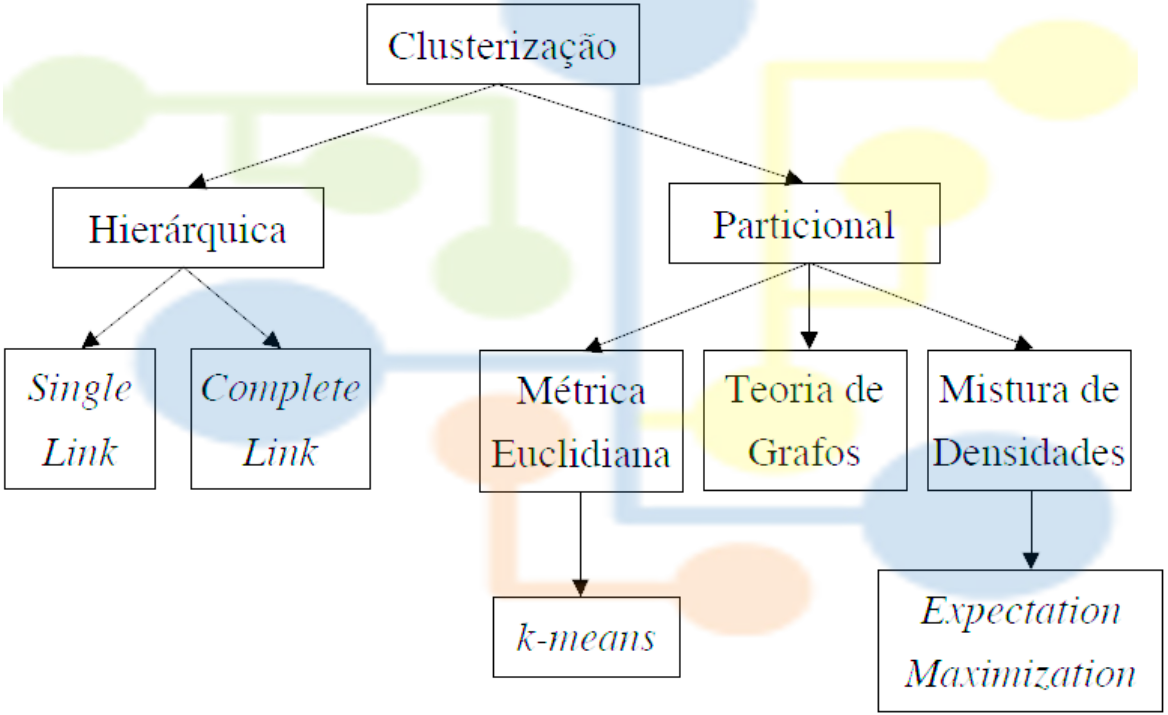
Machine Learning

Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5

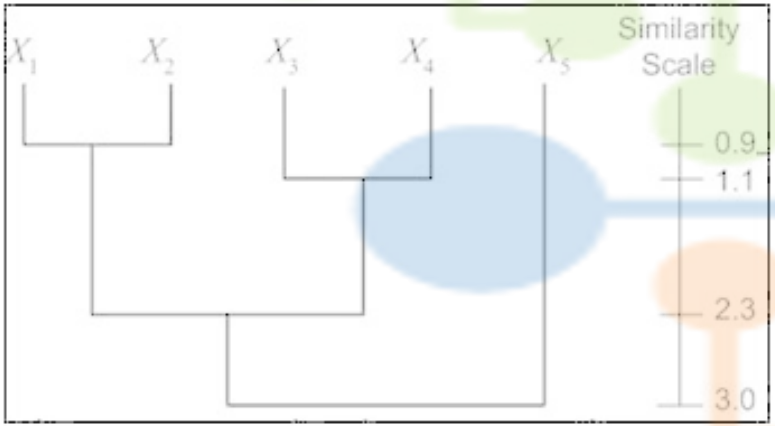
Data Science Academy

Tipos de Clustering

Clustering
Clustering Hierárquico
Clustering Particional
Aglomerativo
Divisivo


Clusterização
Hierárquica
Particional
Single Link
Complete Link
Métrica Euclidiana
Teoria de Grafos
Mistura de Densidades
k-means
Expectation Maximization

Clustering Hierárquico



Dendograma

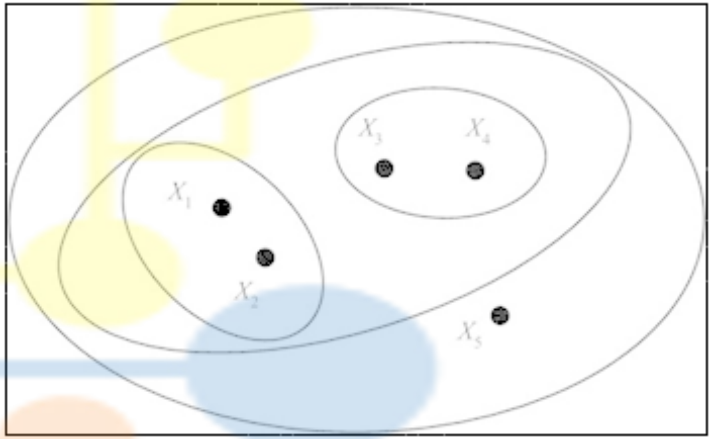


Diagrama de Venn

Clustering Hierárquico

Agglomerative Clustering	Divisive Clustering
bottom-up	top-down

Para o agrupamento hierárquico, os pontos de dados reais não são necessários. Somente a matriz de medidas de distância é suficiente, pois o agrupamento é feito com base nas distâncias.

Clustering Hierárquico

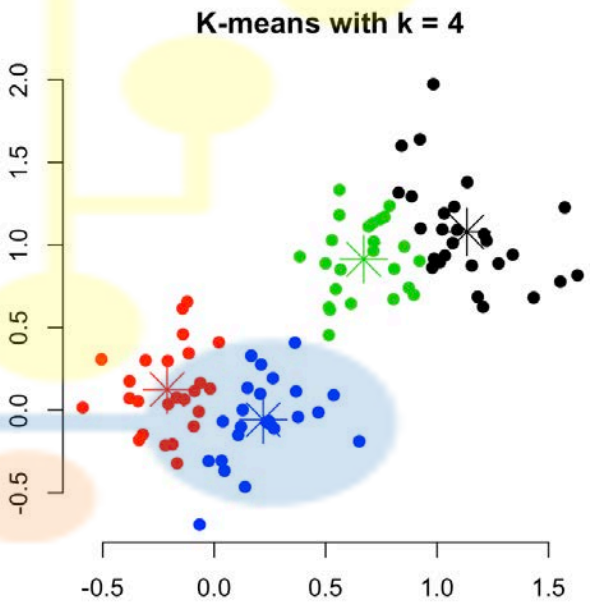
1- Comece com clusters como S1 = {X1}, S2 = {X2} ... Sm = {Xm}.

2- Encontre um conjunto dos clusters mais próximos e mescle-os em um único cluster.

3- Repita o passo 2 até que o número de aglomerados formados seja igual a um número definido.

Clustering Particional

Uma vez que cada cluster formado é mutuamente exclusivo, nunca pode haver uma relação hierárquica entre os clusters.



Clustering Hierárquico

X

Clustering Particional

- Suposições Básicas
- Velocidade
- Processamento
- Parâmetros de Entrada
- Definição e Uso

A qualidade do agrupamento depende do algoritmo escolhido, da função de distância e da aplicação. Diz-se que a qualidade de um modelo de cluster é superior, quando a distância **inter-cluster** é maximizada e a distância **intra-cluster** é minimizada.

Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5

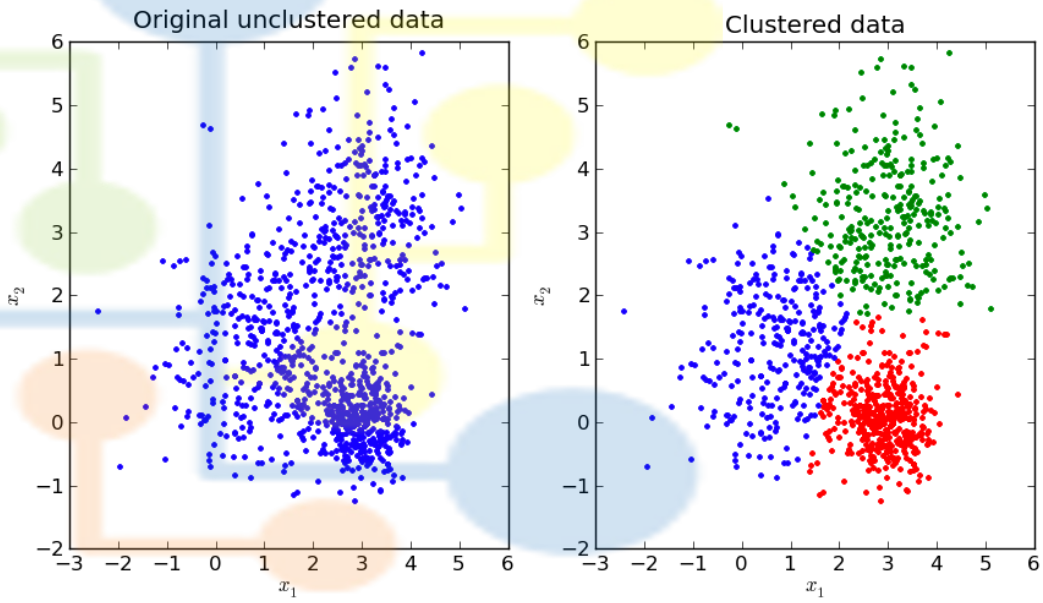
Machine Learning

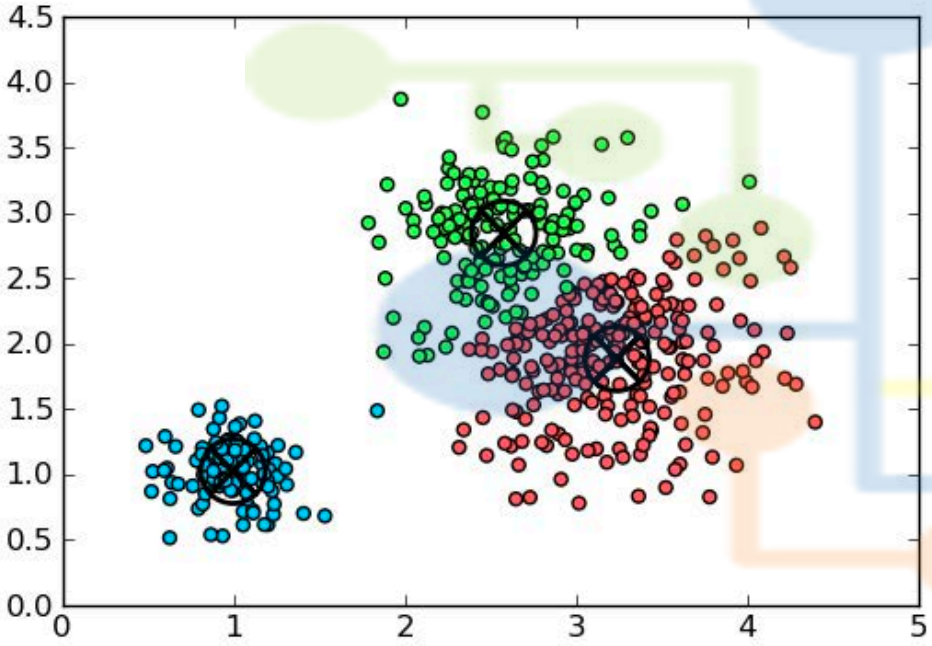
Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5

Data Science Academy

Algoritmo K-Means

O algoritmo K-Means atribui cada um dos n exemplos de dados a um dos k clusters, onde k é um número que foi determinado previamente.





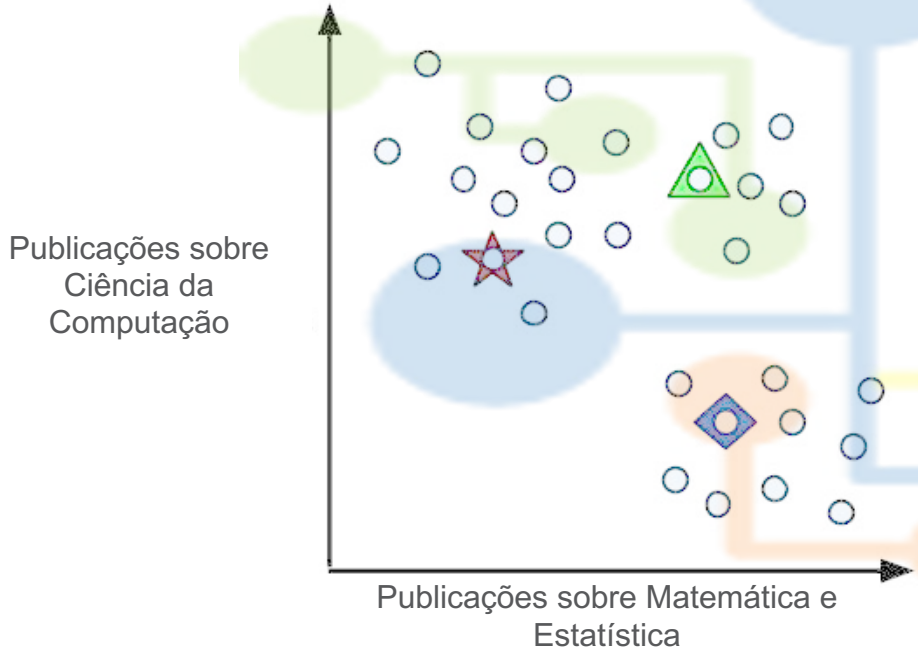
1- O algoritmo **atribui** exemplos a um conjunto inicial de k clusters.

2- **Atualiza** as atribuições ajustando os limites de cluster de acordo com os exemplos que estão no cluster.

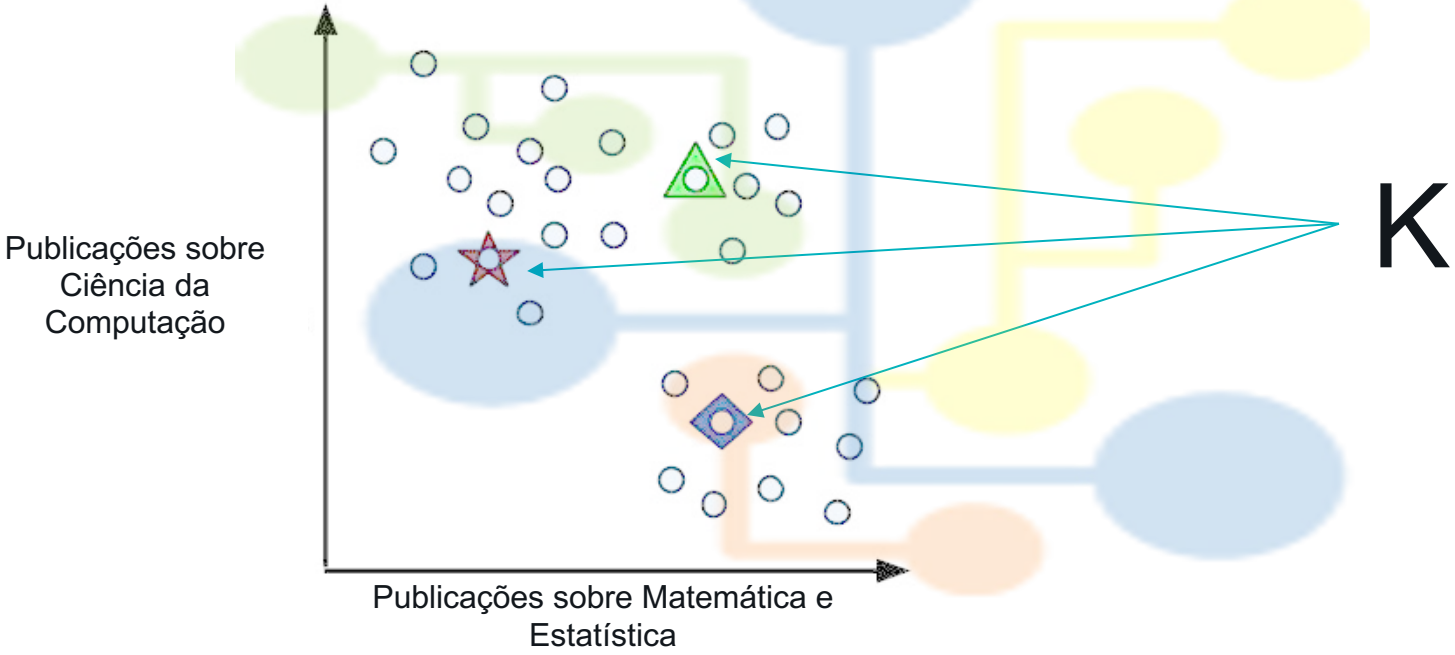
Algoritmo *K*-means

K-MEANS$(\{\vec{x}_1, \ldots, \vec{x}_N\}, K)$

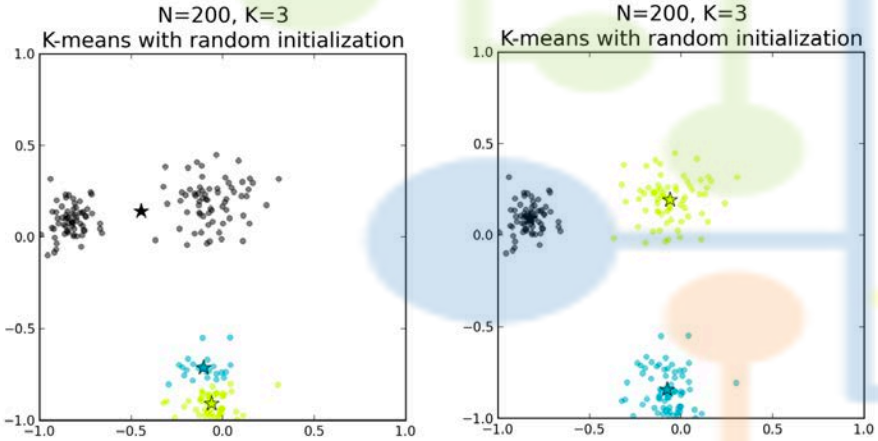
$(\vec{s}_1, \vec{s}_2, \ldots, \vec{s}_K) \leftarrow$ SELECTRANDOMSEEDS$(\{\vec{x}_1, \ldots, \vec{x}_N\}, K)$
for $k \leftarrow 1$ **to** K
do $\vec{\mu}_k \leftarrow \vec{s}_k$
while stopping criterion has not been met
do for $k \leftarrow 1$ **to** K
  do $\omega_k \leftarrow \{\}$
  for $n \leftarrow 1$ **to** N
  do $j \leftarrow \arg\min_{j'} |\vec{\mu}_{j'} - \vec{x}_n|$
    $\omega_j \leftarrow \omega_j \cup \{\vec{x}_n\}$ *(reassignment of vectors)*
  for $k \leftarrow 1$ **to** K
  do $\vec{\mu}_k \leftarrow \frac{1}{|\omega_k|} \sum_{\vec{x} \in \omega_k} \vec{x}$ *(recomputation of centroids)*
return $\{\vec{\mu}_1, \ldots, \vec{\mu}_K\}$



O algoritmo k-means começa por escolher k pontos no espaço de atributos para servir como centros de cluster.


Publicações sobre Ciência da Computação
K
Publicações sobre Matemática e Estatística

K-Means++

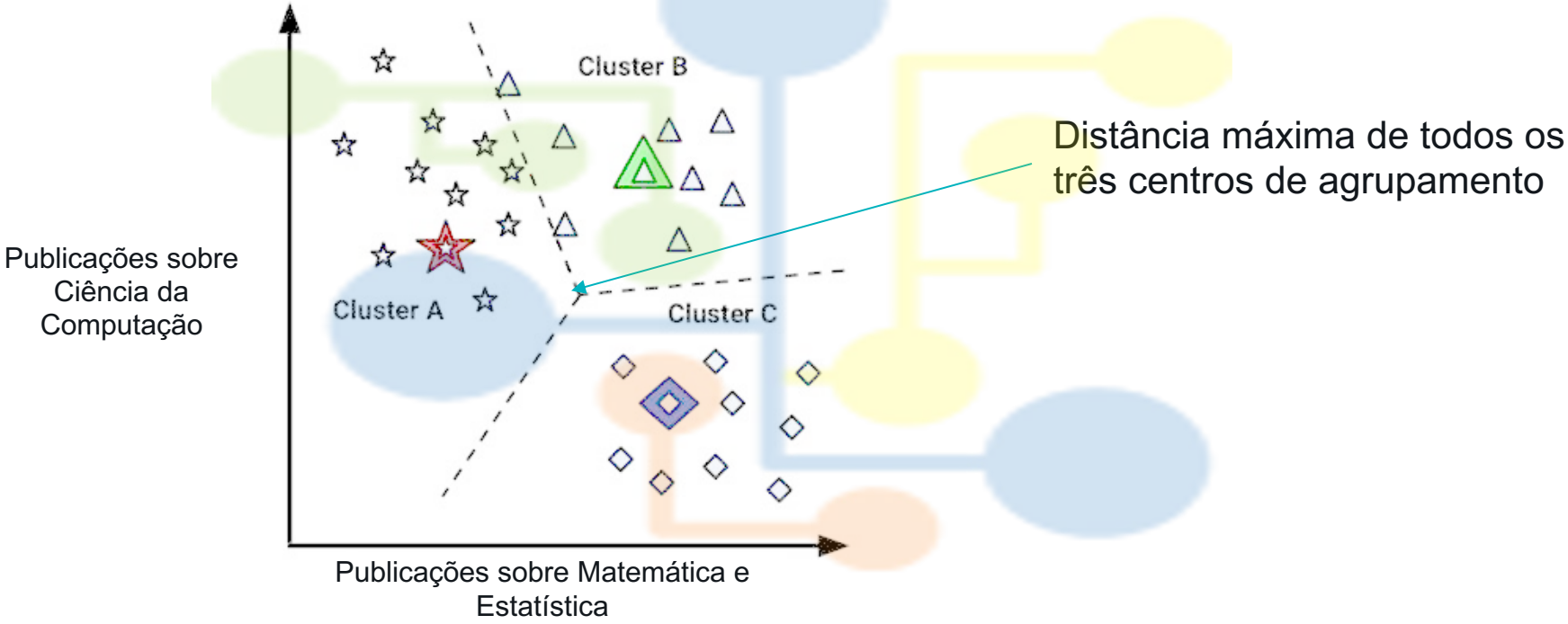


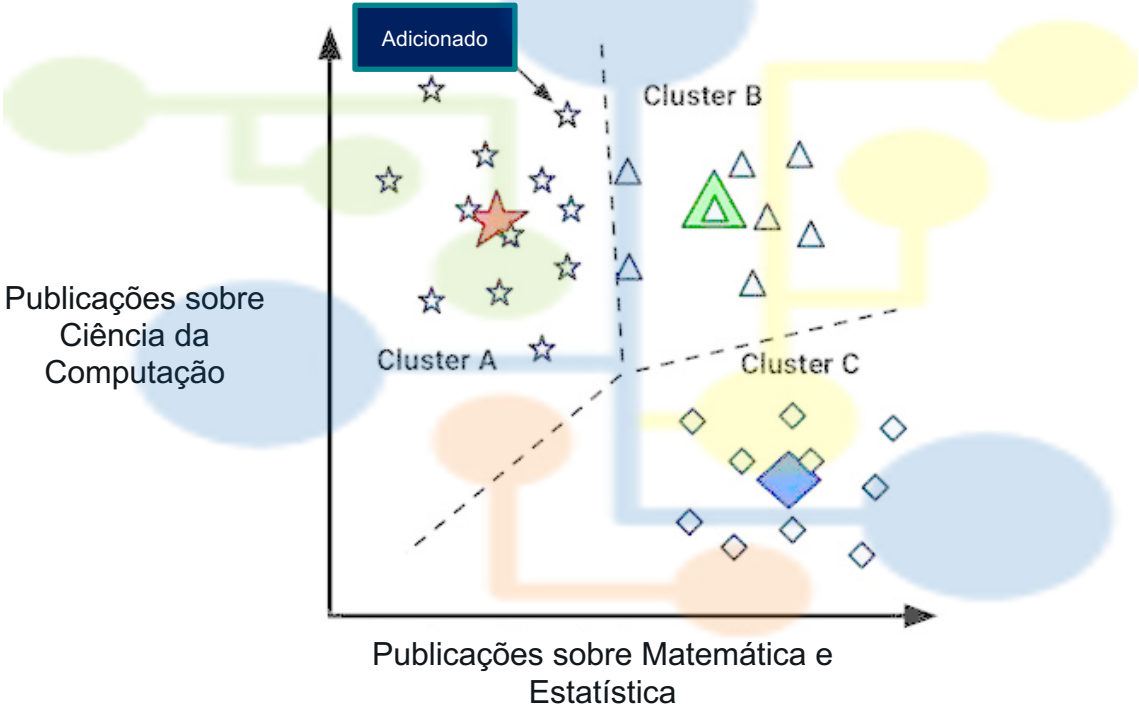
Em 2007, foi introduzido um algoritmo chamado k-means++, que propõe um método alternativo para a seleção dos centros de cluster iniciais.

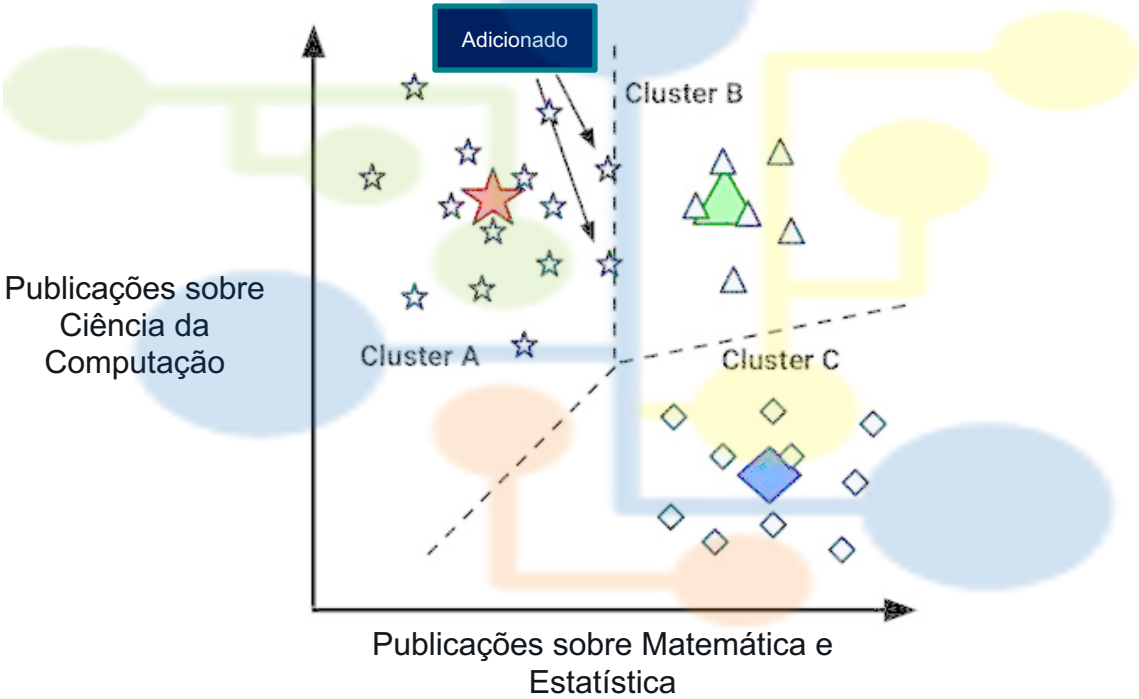
Distância Euclidiana

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

```
> sqrt(( 5 - 0) ^ 2 + (1 - 2) ^ 2)
[1] 5.09902
```


Cluster B
Cluster A
Cluster C
Distância máxima de todos os três centros de agrupamento
Publicações sobre Ciência da Computação
Publicações sobre Matemática e Estatística


Adicionado
Cluster B
Cluster A
Cluster C
Publicações sobre Ciência da Computação
Publicações sobre Matemática e Estatística


Adicionado
Cluster B
Cluster A
Cluster C
Publicações sobre Ciência da Computação
Publicações sobre Matemática e Estatística

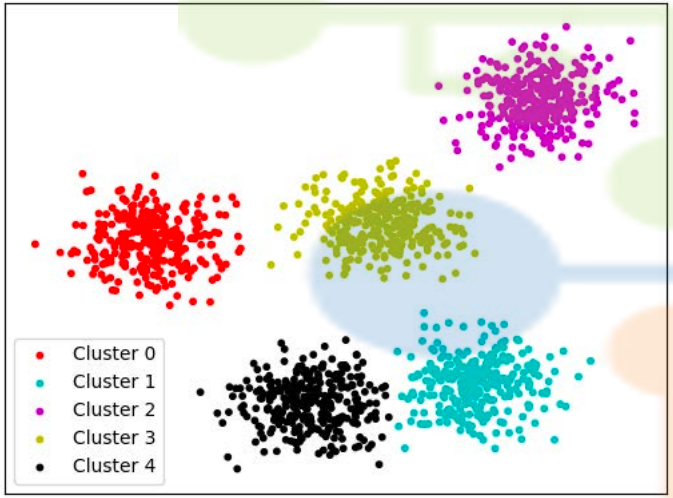
Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5

Machine Learning

Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5

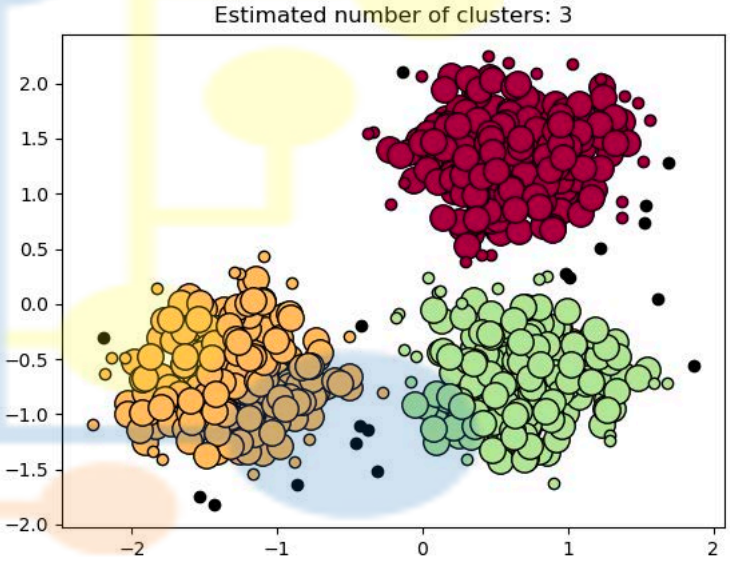
Data Science Academy

Escolhendo o Número Apropriado de Clusters



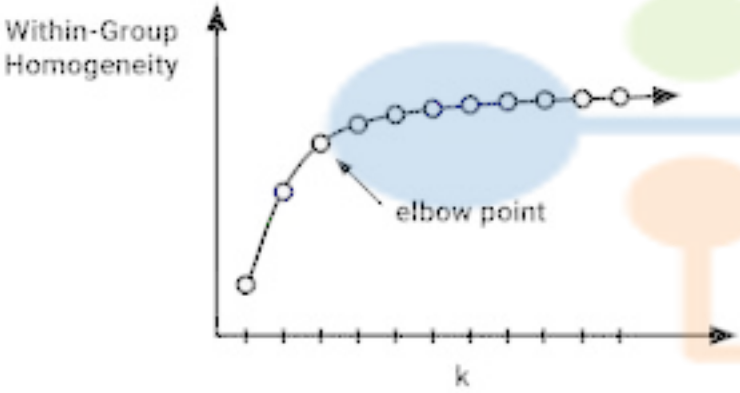
Muito provavelmente, você terá um conhecimento a priori sobre os verdadeiros agrupamentos e você pode aplicar essa informação para escolher o número de clusters.

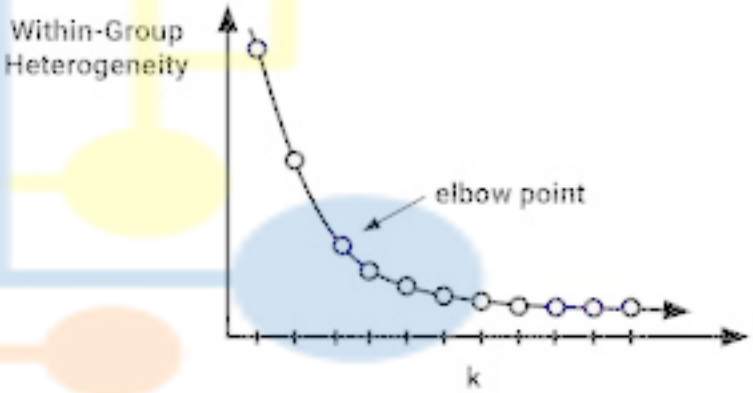
Às vezes, o número de clusters é definido por requisitos de negócios ou a motivação para a análise.



Sem qualquer conhecimento prévio, uma regra geral sugere que k seja igual à raiz quadrada de (n / 2), onde n é o número de exemplos no conjunto de dados.

Método de Elbow





Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5