CLUSTERING

Prof. Nielsen Rechia
nielsen.machado@uniritter.edu.br

Clustering

	Supervisionado	Não-supervisionado
Paradigmas		
Tarefas	Classificação	Análise associativa
	Regressão	Agrupamento (*clustering*)
	Outros	Redução de dimensionalidade
		Outros

7 tarefas comuns de aprendizado de máquina: http://vitalflux.com/7-common-machine-learning-tasks-related-methods/

Clustering

No aprendizado não supervisionado normalmente não temos a informação da classe das instâncias de treinamento.

Agrupamento é uma tarefa que visa agrupar um conjunto de objetos (instâncias) em diferentes classes (grupos) de objetos de acordo com sua similaridade.

Clustering

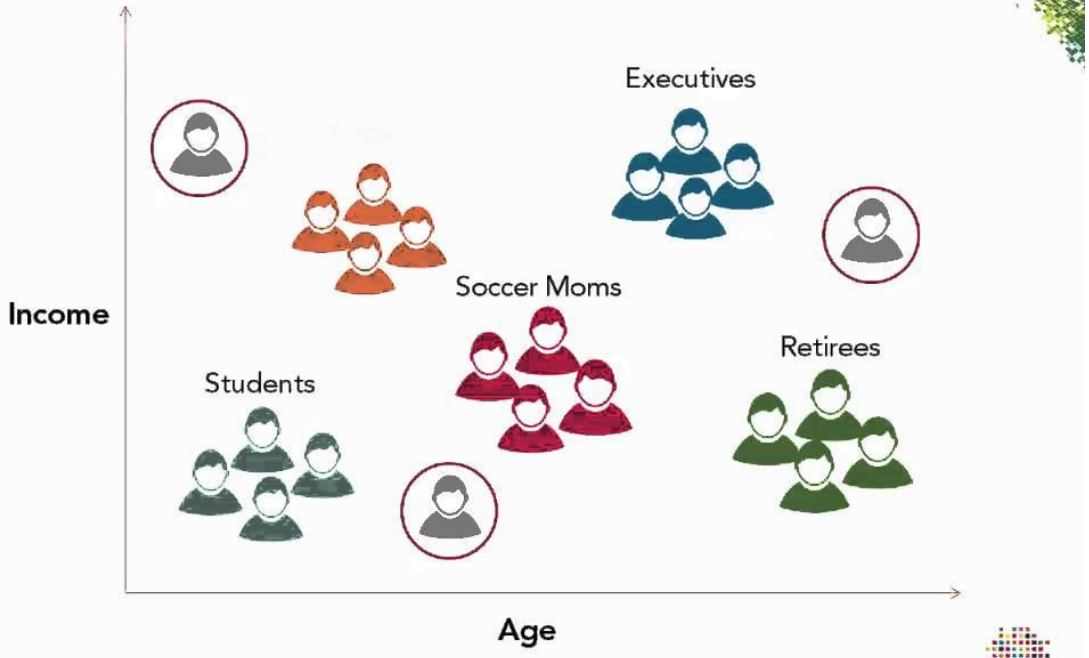
Problemas:

Os consumidores possuem perfis similares?

É possível agrupar tipos de câncer pelo comportamento que apresentam?

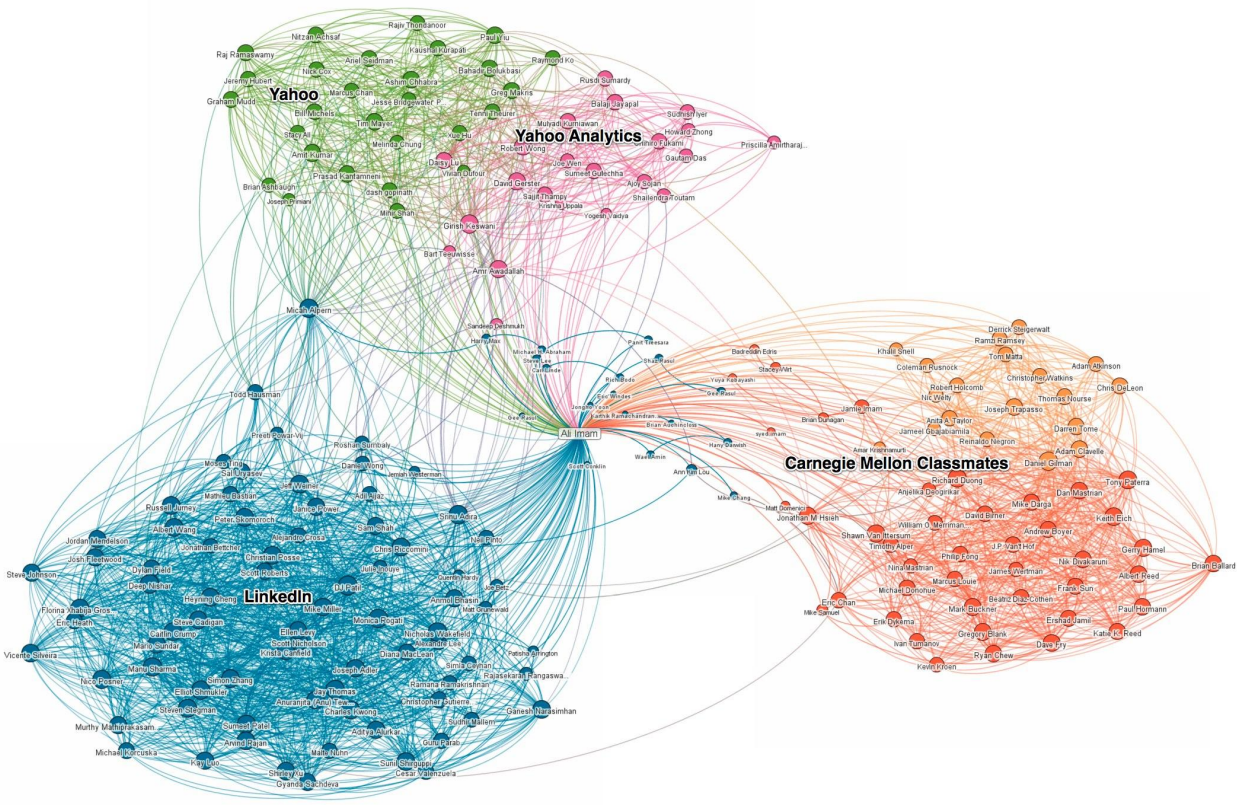
Quais as regiões onde determinado tipo de crime acontece mais frequentemente?

Clustering



Angoss
Predict. Act. Perform.

Clustering


Yahoo
Yahoo Analytics
Carnegie Mellon Classmates
LinkedIn
Ali Imam

Clustering

Existem tipos de algoritmos de agrupamento

Algoritmos particionais:

K-Means

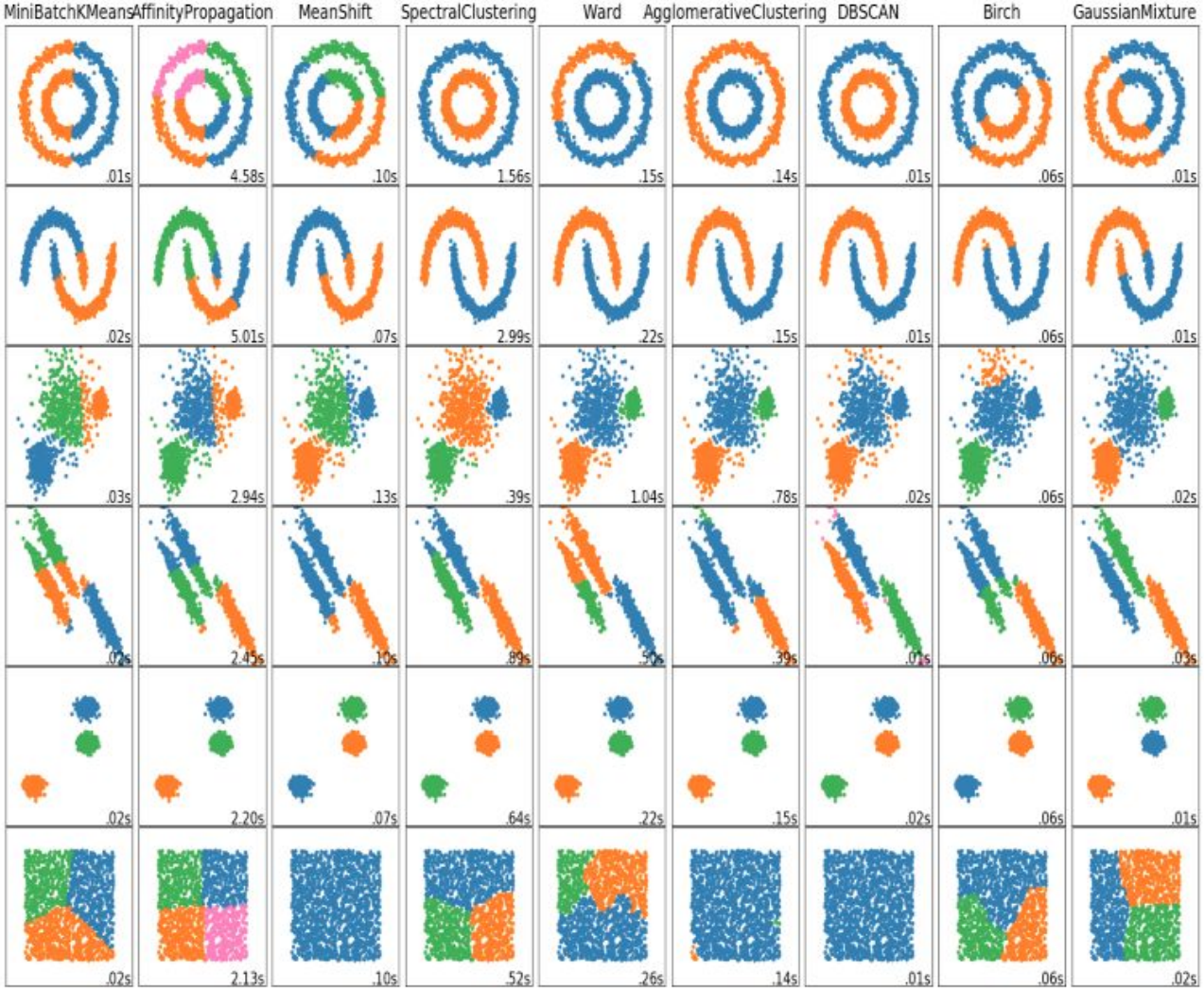
Mais popular, rápido e com bom desempenho

Algoritmos de densidade:

DBSCAN

Algoritmos hierárquicos:

Complete Linkage, Single Linkage, WARD, Average (UPGMA), PAM e ROCK


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
GaussianMixture
.01s
4.58s
.10s
1.56s
.15s
.14s
.01s
.06s
.01s
.02s
5.01s
.07s
2.99s
.22s
.15s
.01s
.06s
.01s
.03s
2.94s
.13s
.39s
1.04s
.78s
.02s
.06s
.02s
.02s
2.45s
.10s
.89s
.50s
.39s
.01s
.06s
.03s
.02s
2.20s
.07s
.64s
.22s
.15s
.02s
.06s
.01s
.02s
2.13s
.10s
.52s
.26s
.14s
.01s
.06s
.02s

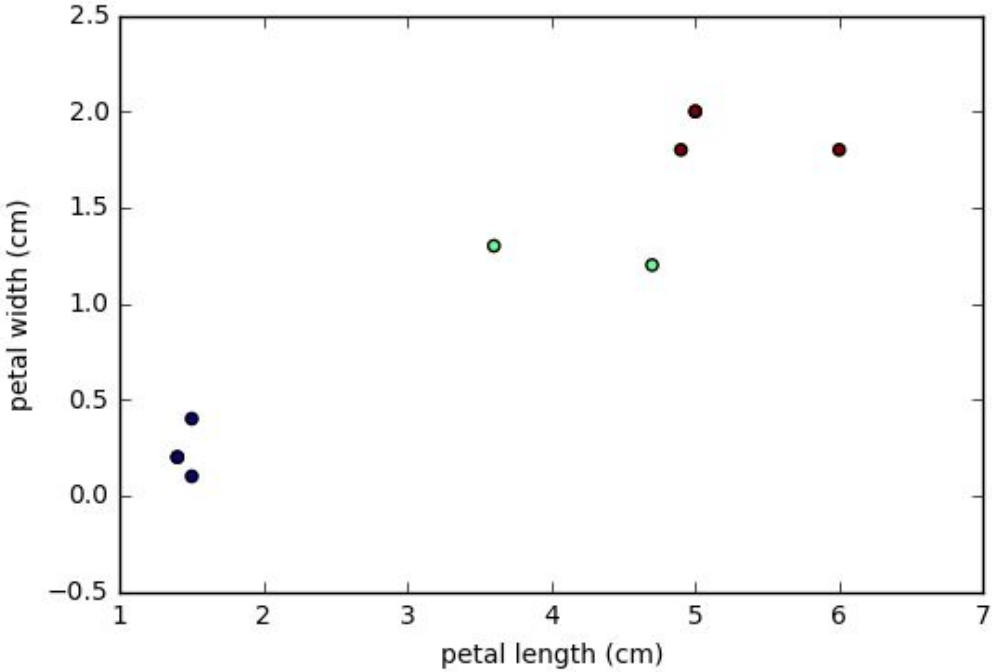
Clustering

Dado parte do conjunto de dados iris, como agrupar as instâncias similares?

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	class
9	4.9	3.1	1.5	0.1	0.0
125	7.2	3.2	6.0	1.8	2.0
15	5.7	4.4	1.5	0.4	0.0
64	5.6	2.9	3.6	1.3	1.0
113	5.7	2.5	5.0	2.0	2.0
123	6.3	2.7	4.9	1.8	2.0
113	5.7	2.5	5.0	2.0	2.0
8	4.4	2.9	1.4	0.2	0.0
73	6.1	2.8	4.7	1.2	1.0
0	5.1	3.5	1.4	0.2	0.0

Clustering

Dados originais no plano.



Clustering – K-Means

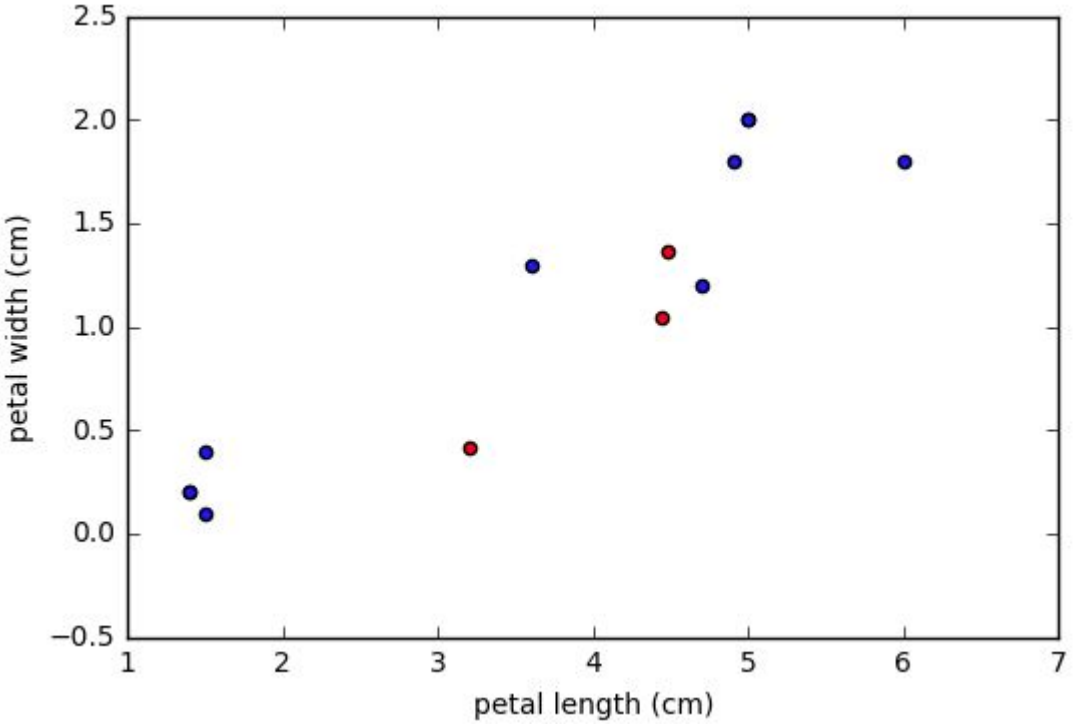
Escolhemos aleatoriamente K centróides para os clusters (grupos).

Intervalos de valores:

	petal length (cm)	petal width (cm)
min	1.4	0.1
max	6.0	2.0

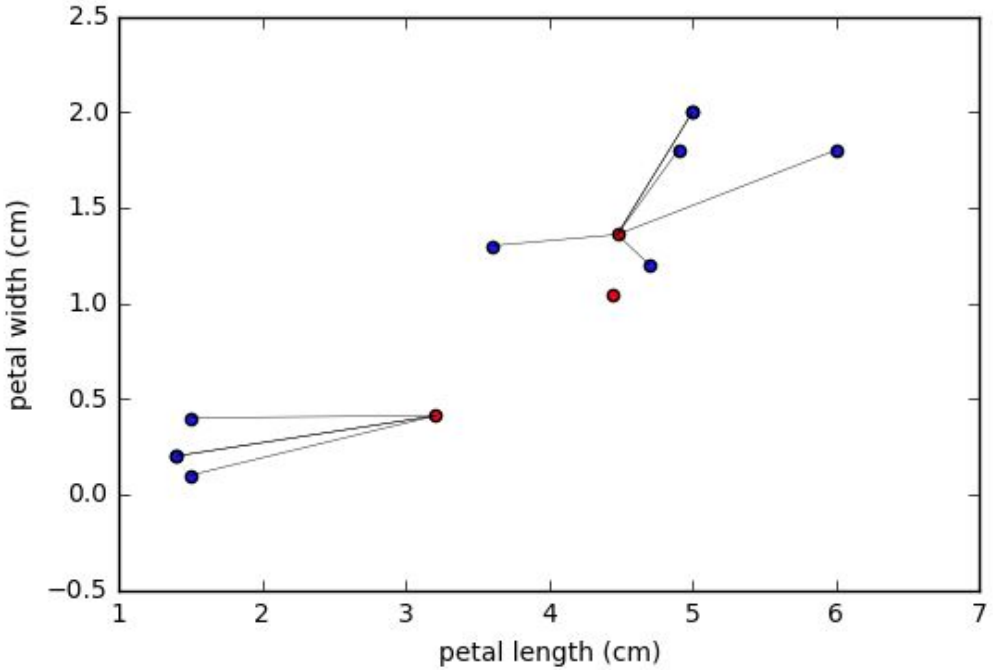
centróides:

	petal length (cm)	petal width (cm)
0	4.448337	1.039741
1	4.474348	1.358150
2	3.209553	0.410570



Clustering – K-Means

Atribuímos cada uma das instâncias ao centróide mais próximo.

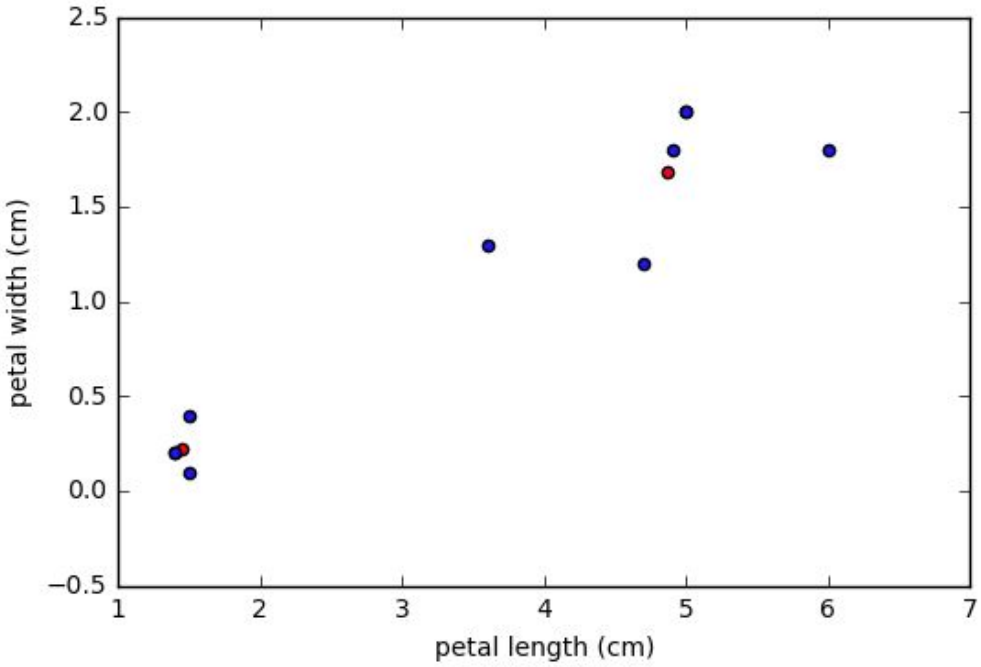


Clustering – K-Means

Movemos cada centróides para a média dos objetos do cluster correspondente

novos centróides:

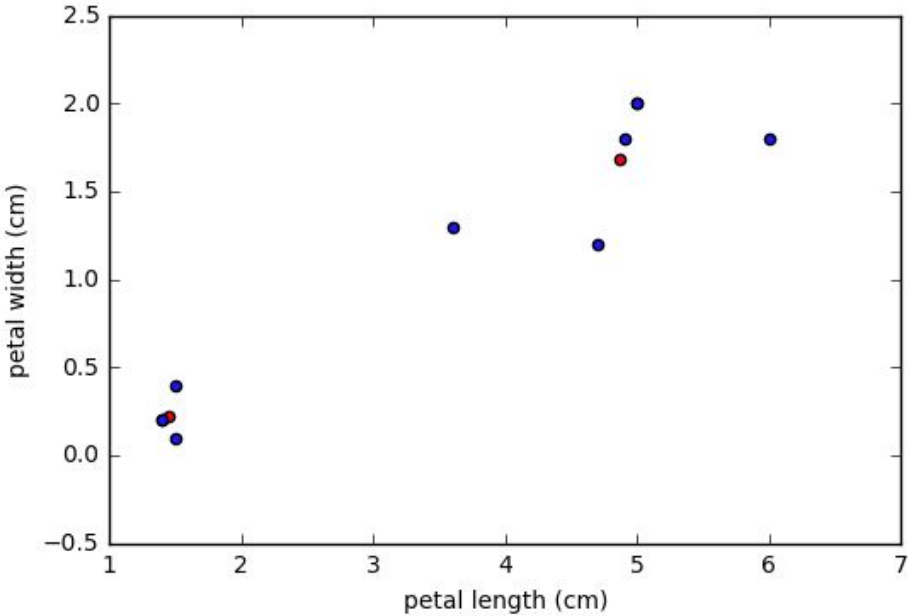
	petal length (cm)	petal width (cm)
0	NaN	NaN
1	4.866667	1.683333
2	1.450000	0.225000



Clustering – K-Means

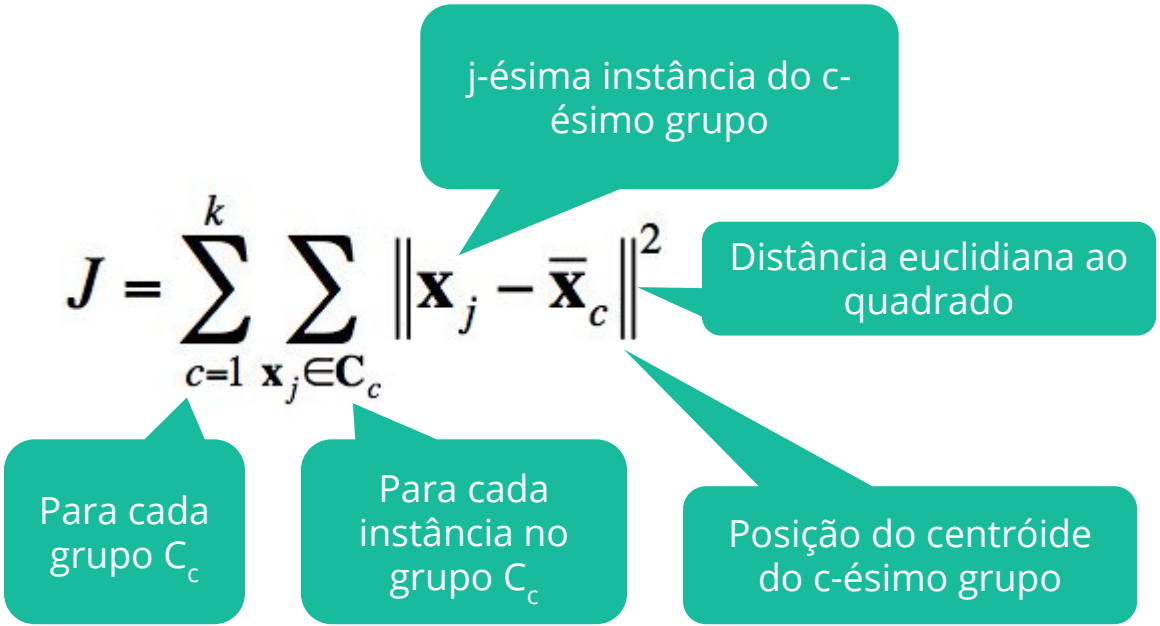
Iteramos entre os passos anteriores até que algum critério de convergência seja satisfeito

- Número máximo de iterações seja atingido
- Limiar mínimo de mudanças no centróides



Clustering - K-Means

Minimizar a seguinte função de custo:



Clustering - K-Means

K-Means pode sofre problemas ao lidar com grupos de diferentes

- Tamanhos
- Densidades
- Formas (principalmente não-esféricas)
- É sensível a outliers

Existem variações como **K-medianas**

Hierarchical Clustering

Algoritmos hierárquicos

Ex: single linkage, ward, complete linkage

Permitem verificar a relação que os dados possuem

Relação é visual em uma estrutura chamada dendrograma

Existem duas formas de agrupamento hierárquico:

Aglomerativo: Começa com N grupos de uma instância, e combina tais grupos até chegar a um único grupo

Divisivo: Começa com apenas um grupo, e separa sucessivamente até terminar em grupos de uma instância

Na prática, algoritmos aglomerativos são mais populares

Agglomerative Clustering

Dendrograma

É uma ferramenta útil para sumarizar medidas de (dis)similaridade e mostrar as hierarquias de partições resultantes de algoritmos hierárquicos aglomerativos

Pode-se examinar o dendrograma inclusive para estimar o número correto de grupos.

Partições são obtidas vias cortes no dendrograma

- Cortes horizontais
- No de grupos da partição = Número de intersecções

Agglomerative Clustering

Algoritmos hierárquicos

Ex: single linkage

$$d(u, v) = \min(dist(u[i], v[j]))$$

j-ésima instância do cluster v

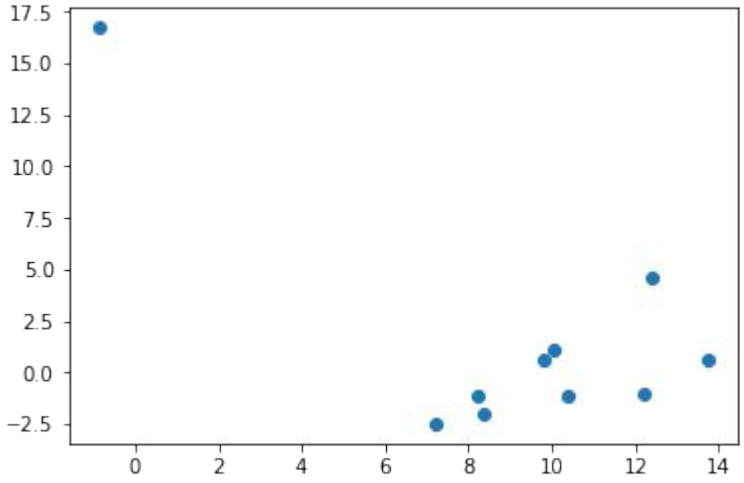
Distância entre os clusters u, v

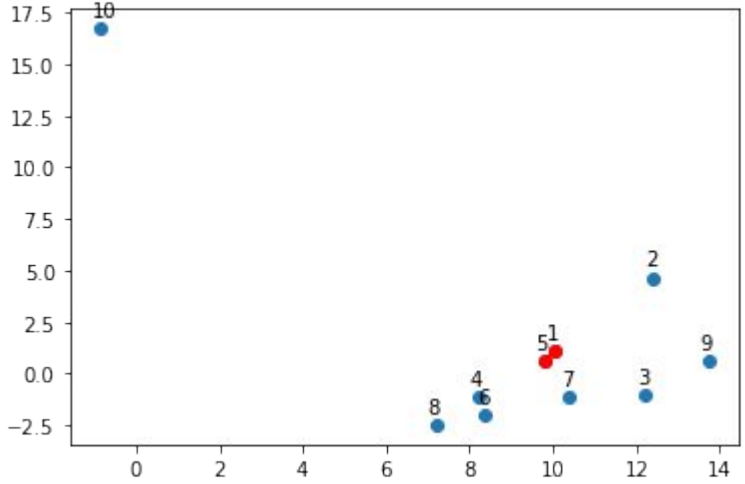
i-ésima instância do cluster u

Agglomerative Clustering

Single Linkage

Encontrar a menor distância entre os objetos do conjunto de dados

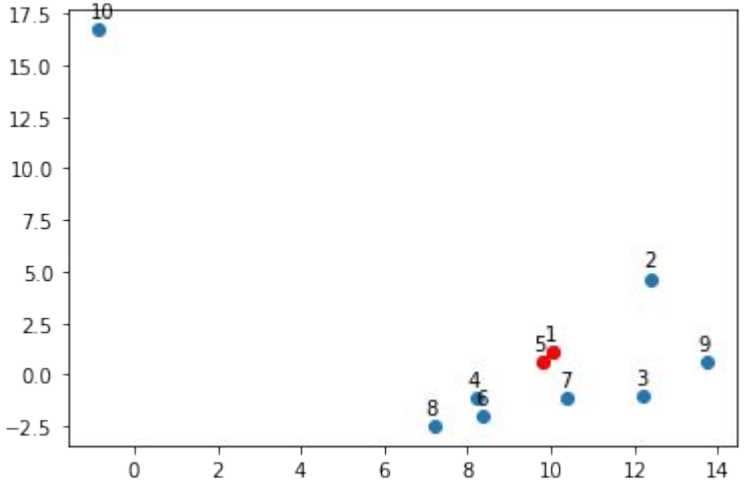




Agglomerative Clustering

Dendrograma

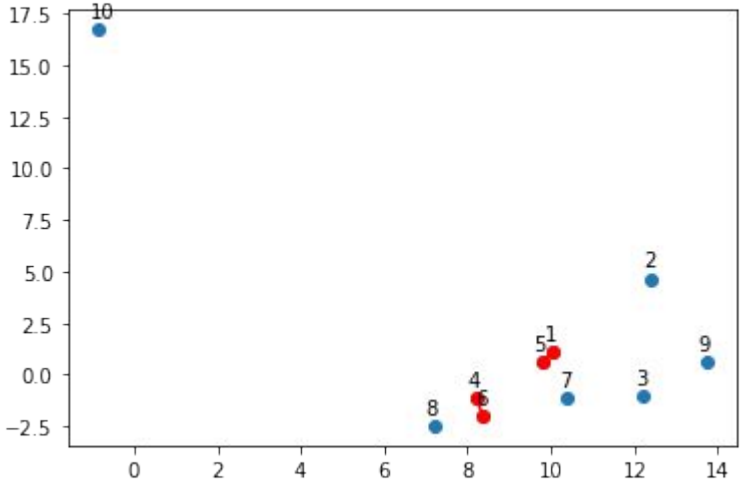
Conecte os dois grupos (instâncias) da distância encontrada



Agglomerative Clustering

Dendrograma

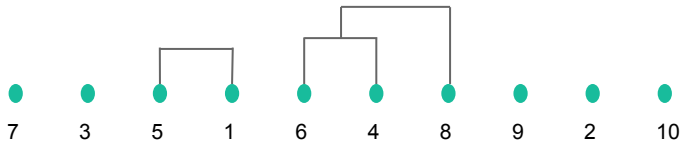
Repita o processo até que todas as instâncias estejam conectadas

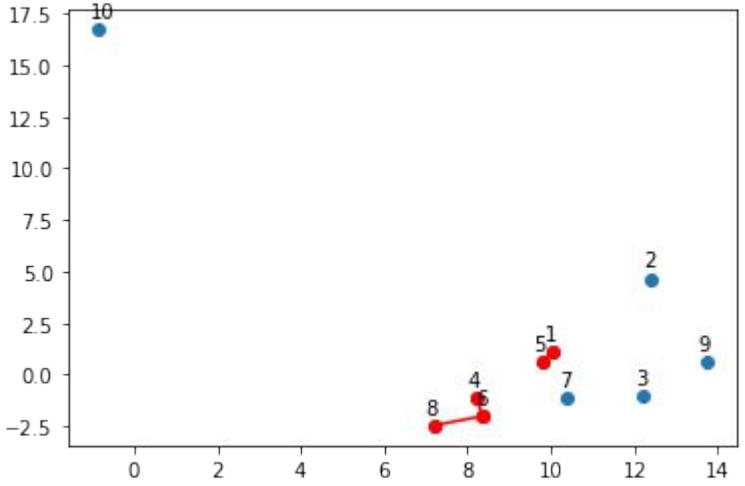


Agglomerative Clustering

Dendrograma

Repita o processo até que todas as instâncias estejam conectadas

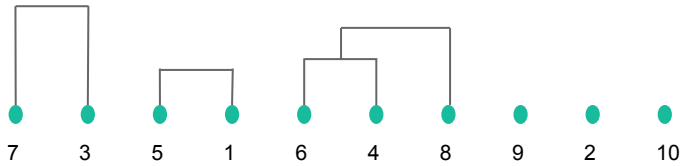


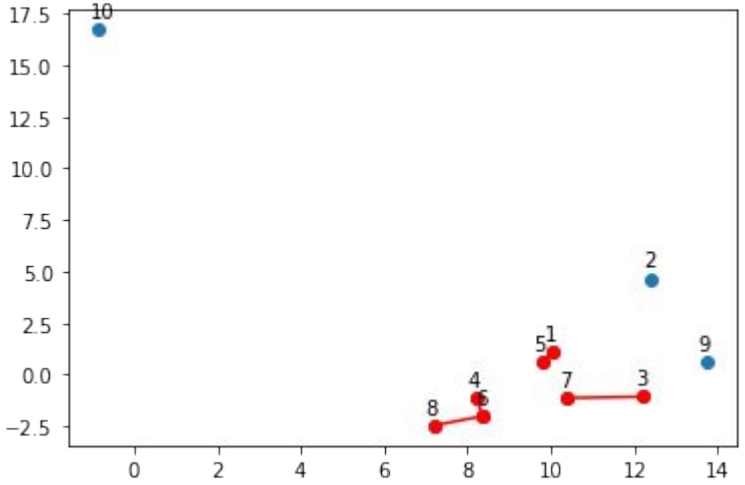


Agglomerative Clustering

Dendrograma

Repita o processo até que todas as instâncias estejam conectadas

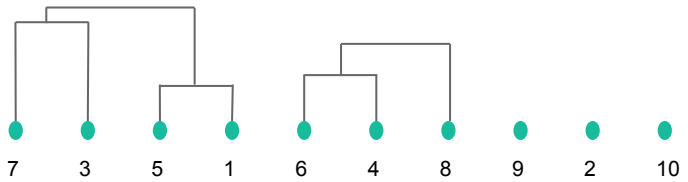


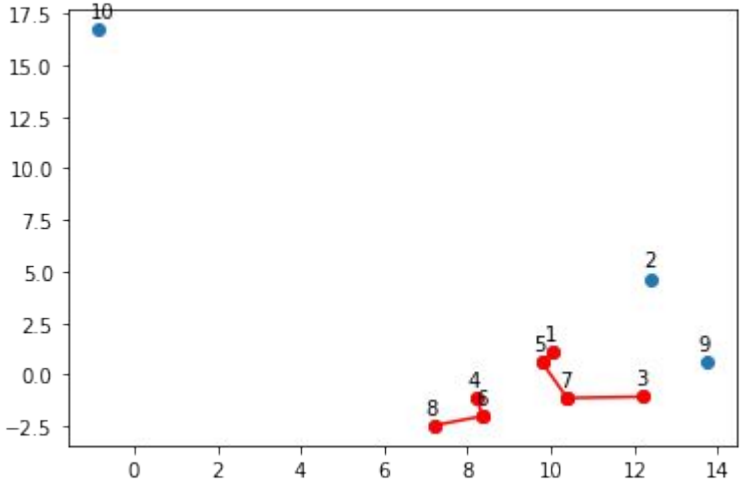


Agglomerative Clustering

Dendrograma

Repita o processo até que todas as instâncias estejam conectadas

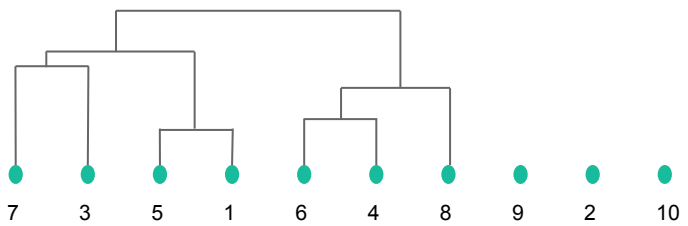


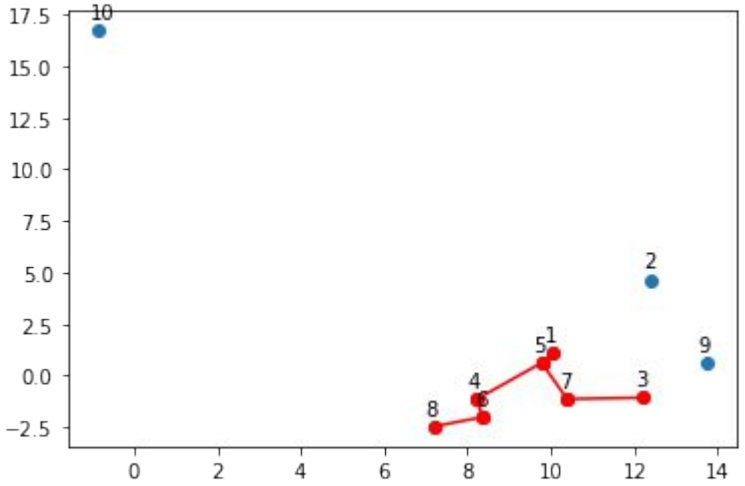


Agglomerative Clustering

Dendrograma

Repita o processo até que todas as instâncias estejam conectadas

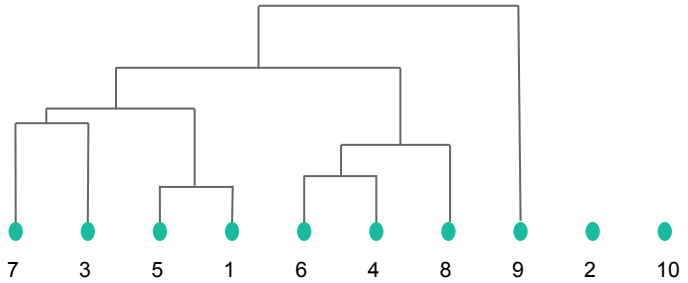


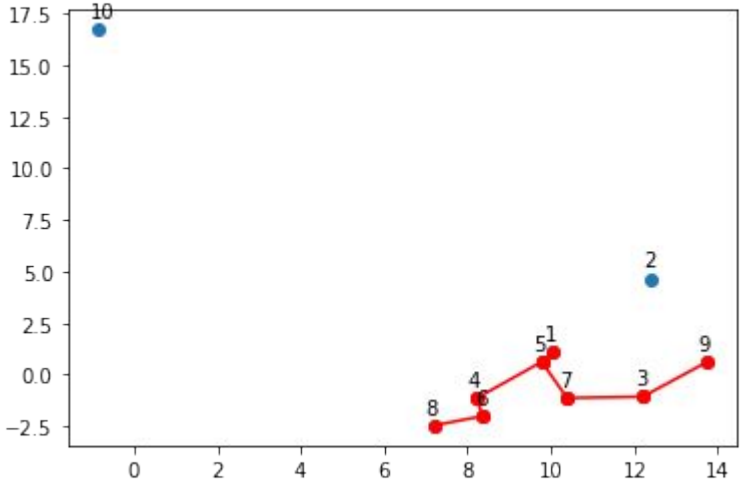


Agglomerative Clustering

Dendrograma

Repita o processo até que todas as instâncias estejam conectadas

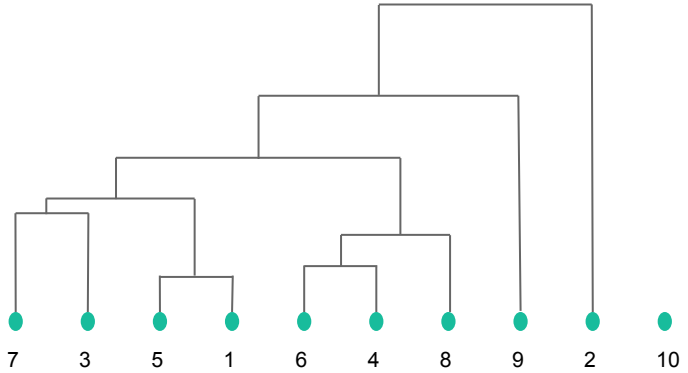


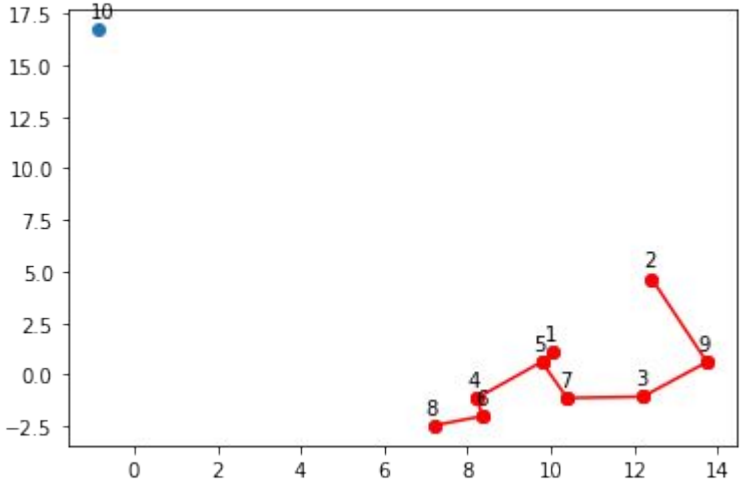


Agglomerative Clustering

Dendrograma

Repita o processo até que todas as instâncias estejam conectadas

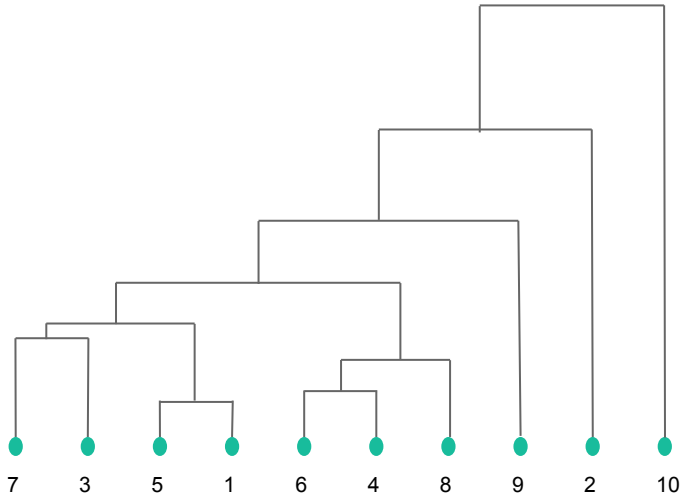


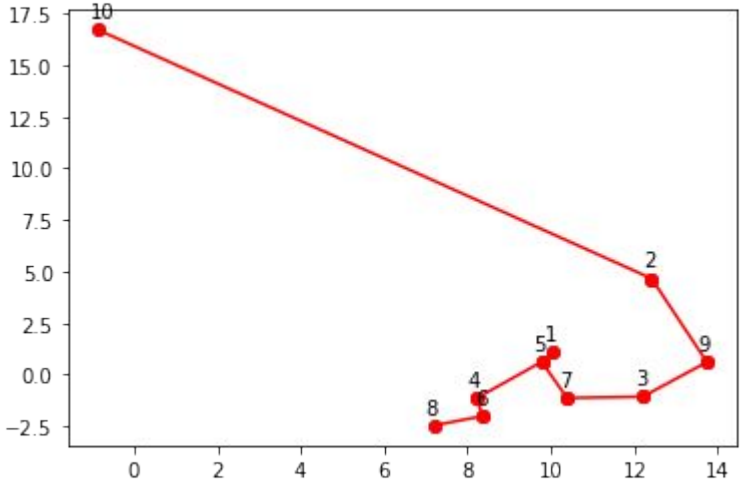


Agglomerative Clustering

Dendrograma

Repita o processo até que todas as instâncias estejam conectadas





Agglomerative Clustering

Para dividir os dados em grupos:

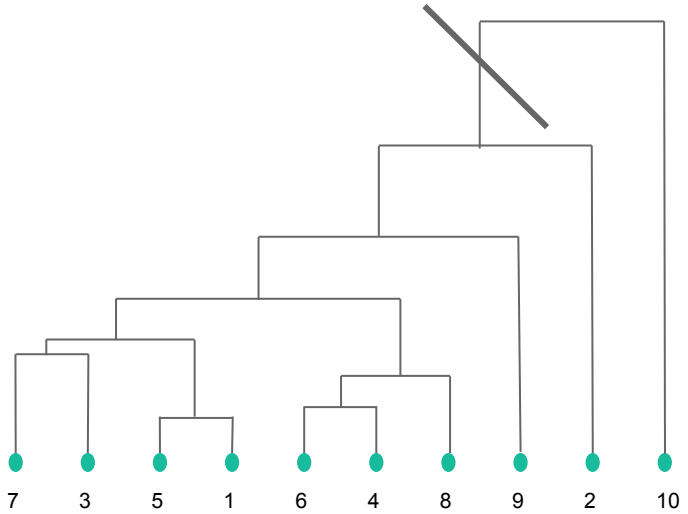
Realize um corte no dendrograma onde a distância entre os dois grupos for a maior

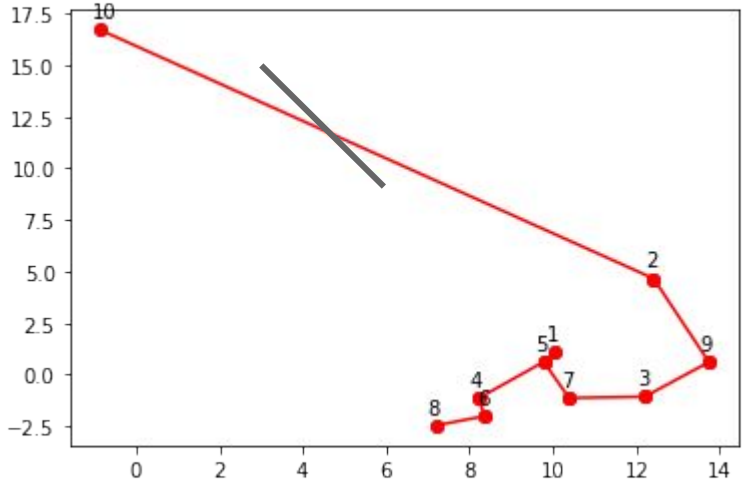
Uma boa regra para realizar agrupamentos é que os dados devem possuir bastante similaridade intra-grupos e pouca similaridade inter-grupos

Caso este critério não seja atingido, é possível continuar realizando cortes

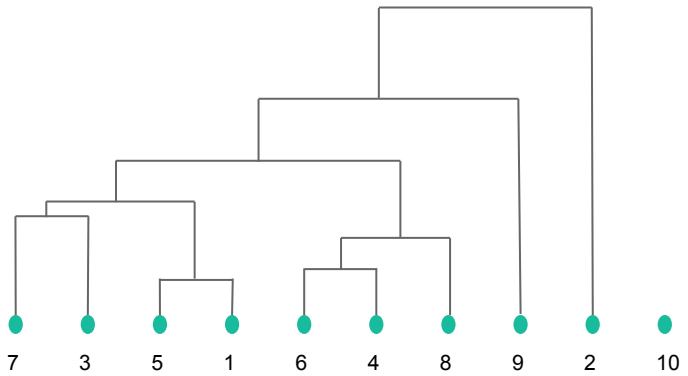
Geralmente utilizam-se medidas de avaliação para verificar este critério

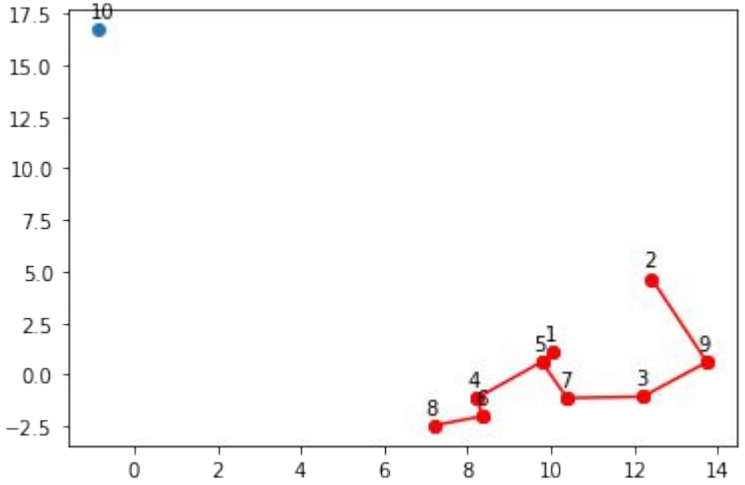
Agglomerative Clustering





Agglomerative Clustering





Agglomerative Clustering

Avalia quantitativamente uma partição/ hierarquia de dados a partir de alguma medida. São divididos em 3 tipos:

Externos: avaliam o grau de correspondência entre a estrutura encontrada e uma solução esperada ou conhecida (golden truth). Ex: Rand Index, Jaccard, Fowlkes-Mallows

Internos: avaliam o grau de compatibilidade entre a estrutura encontrada e os dados (e apenas os dados). Ex: SSE

Relativos: avaliam dentre duas ou mais estruturas, qual a melhor (segundo algum aspecto). Tipicamente são índices internos capazes de quantificar a qualidade dos agrupamentos. Ex: Davis-Bouldin, Silhueta, Dunn, Gap Statistics, Calinski-Harabasz

Clustering - Avaliação

Medidas de avaliação de clustering

Silhueta (Silhouette Width Criterion):

Para cada instância, calcula o índice s(i)

Dissimilaridade média da i-ésima instância ao segundo cluster mais próximo

Dissimilaridade média da i-ésima instância ao cluster mais próximo

$$SWC = \frac{1}{N}\sum_{i=1}^{N} s(i)$$

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

A swc mede quão compactos e bem-separados os dados estão

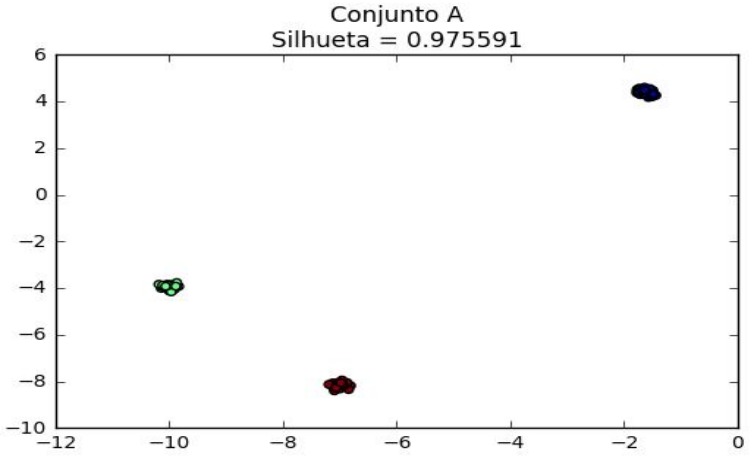
Clustering – Avaliação

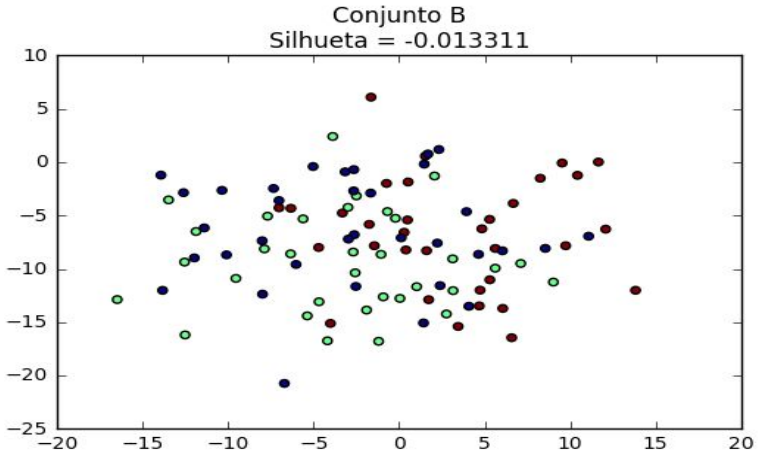
A silhueta varia de -1 até 1:

-1: grupos esparsos e misturados

1: grupos bem separados e compactos

Entre 0 e 0.5 indica que os dados não são bem separados





Clustering – Avaliação

Mais métricas para avaliar agrupamentos:

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

K-Means - sklearn:

```
from sklearn.cluster import KMeans
cl = KMeans(n_clusters=3, init='k-means++', n_init=10, max_iter=300)
cl = cl.fit(df)

labels = cl.labels_
centroids = cl.cluster_centers_
```

Hierarquicos - scipy:

```
from scipy.cluster.hierarchy import dendrogram, linkage
from scipy.cluster.hierarchy import fcluster
Z = linkage(dataset, 'single')
```

Exercício prático

Implemente o algoritmo k-means e single linkage para o dataset iris.

- Remova a classe do conjunto de dados
- Execute com diversos números de grupo (ex, de 2 a 10)
  - Execute apenas uma vez para cada k
- Verifique os agrupamentos utilizando a silhueta
- Qual é o melhor número de grupos?
- Qual é o melhor algoritmo? PQ?

In [77]:

```
from sklearn.datasets import load_iris
import numpy as np
import pandas as pd

data = load_iris()
df = pd.DataFrame(
    data['data'],
    columns=data['feature_names']
)
```

Exercício para entregar

Agrupe os dados do censo americano de 2005 com o K-means e algum algoritmo hierarquico

Não considerar a classe do conjunto de dados

Somente um conjunto de dados

Dados devem ser pré-processados

Usar vários números de grupos (ex: 2 a 20)

Medir a qualidade dos grupos com a silhueta (podem usar outros também)

Qual é o melhor número de grupos? Qual é o melhor algoritmo? Motivos?

Entregar via BB o código python ou jupyter e o passo a passo realizado (ex, motivos do pre-processamento etc.) em um zip.

Conlusão

Leitura recomendada:

Capítulo 8 e 9 de Introduction to Data Mining