



Insper

Machine Learning

Aula 16 – Clustering

2021 – Engenharia
Fábio Ayres <fabioja@insper.edu.br>

Aprendizado supervisionado

2

- Temos variáveis independentes X e a variável dependente y
- Objetivo: Construir modelo preditivo

$$\hat{y} = h(x, \theta)$$

↑
estimativa

↑
parâmetros do modelo

- Para construir o modelo (aprender os parâmetros)
 - Define uma função de perda
 - Aplica um algoritmo de otimização

Aprendizado não-supervisionado

3

- Não temos variável dependente ← não tem classe verdadeira!
- Objetivo: análise exploratória para obter insight

Exempb: clustering

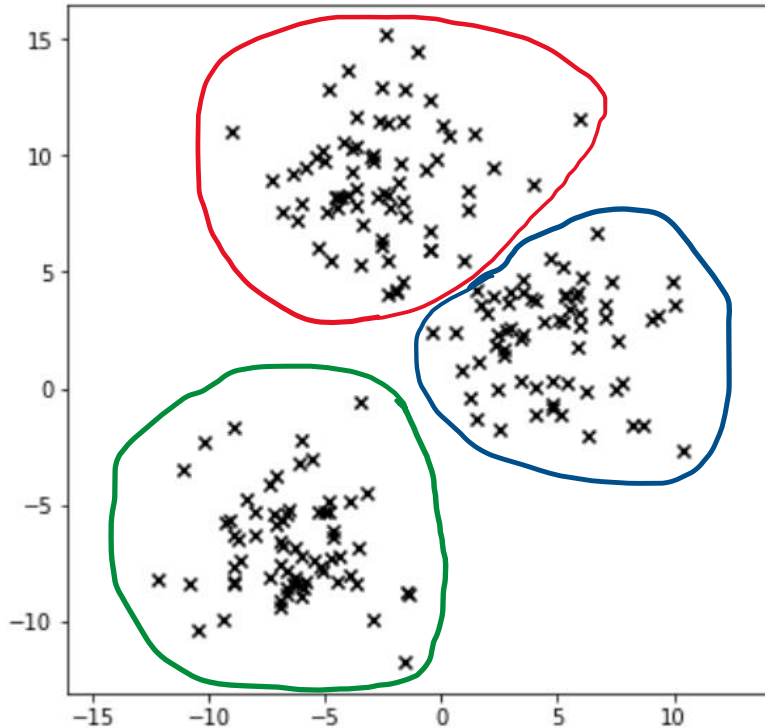
$$\hat{y} = h(x, \theta)$$

cluster

parametros

- Ao invés de otimizar uma "função de perda" (pois não tem y), otimizar algum critério de "qualidade de clustering".

Quais os agrupamentos naturais?

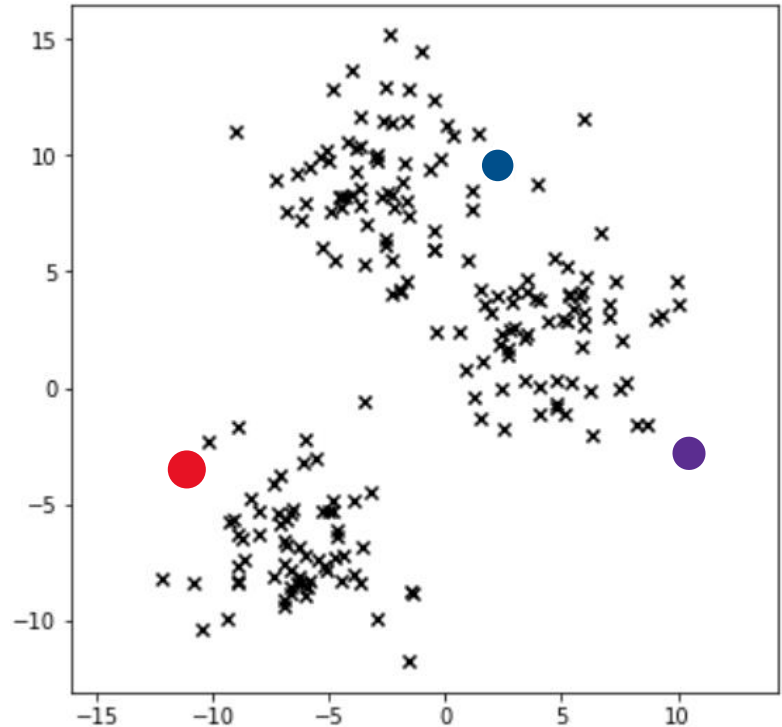


Clustering

- k -means
- Mean Shift
- Clustering Hierarquico

K-means

1. Definir quantos clusters queremos
 $k = n^{\circ}$ clusters
2. Inicialização:
Sorteia k pontos como "centroides" de cluster



K-means

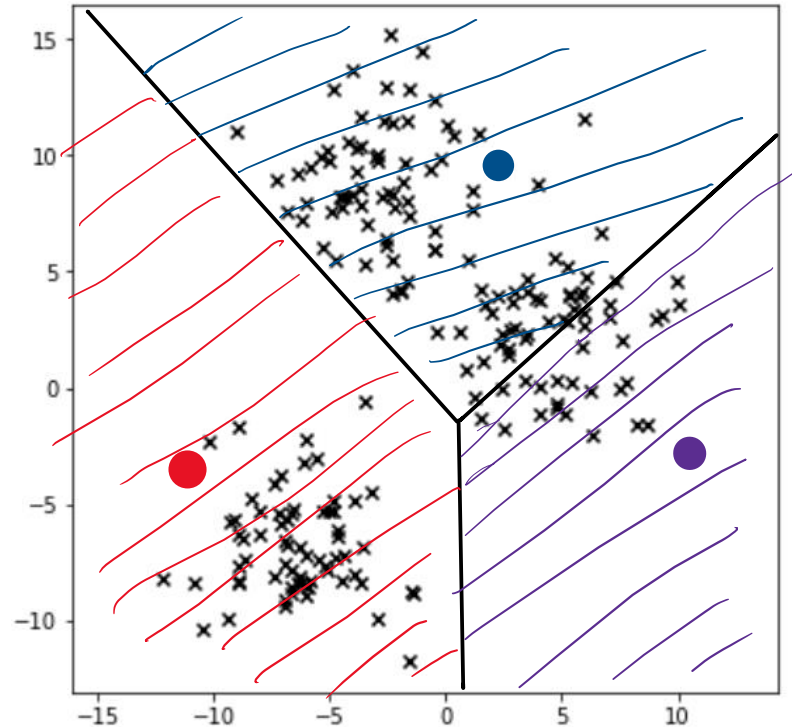
- Associa cada ponto ao centroide mais próximo

mapa de Voronoi

dual

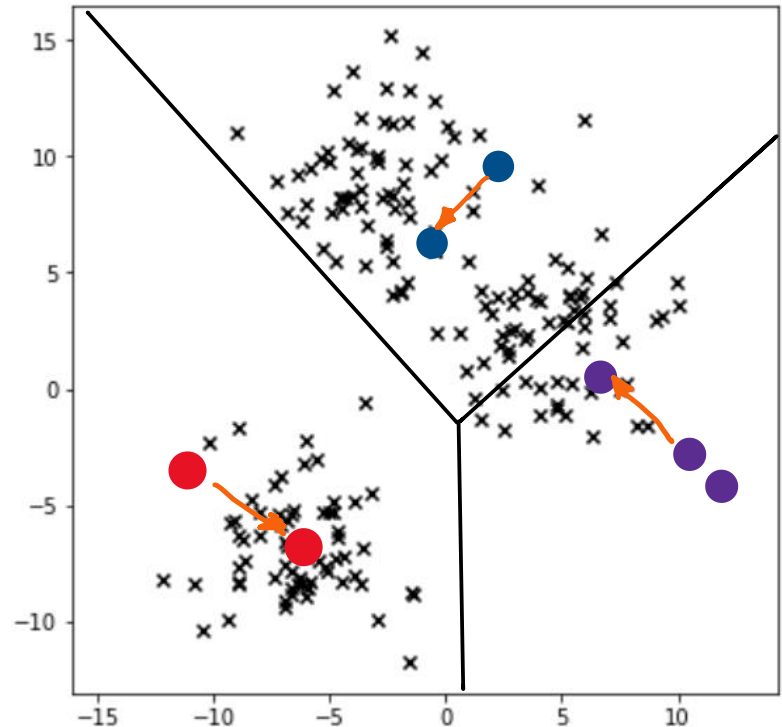
triangulação de
Delaunay

Computational
Geometry



K-means

- Recalcula os centroides
- Repetir até convergência:
 - Associar pontos
 - Recalcular centroide



K-means

Vantagens

- Simples
- Só depende da definição de distância entre pontos
- Escalável

Desvantagens

- Tem que definir a priori o número de clusters.
- Sensível ao chute inicial dos centroides.
- O que mais?

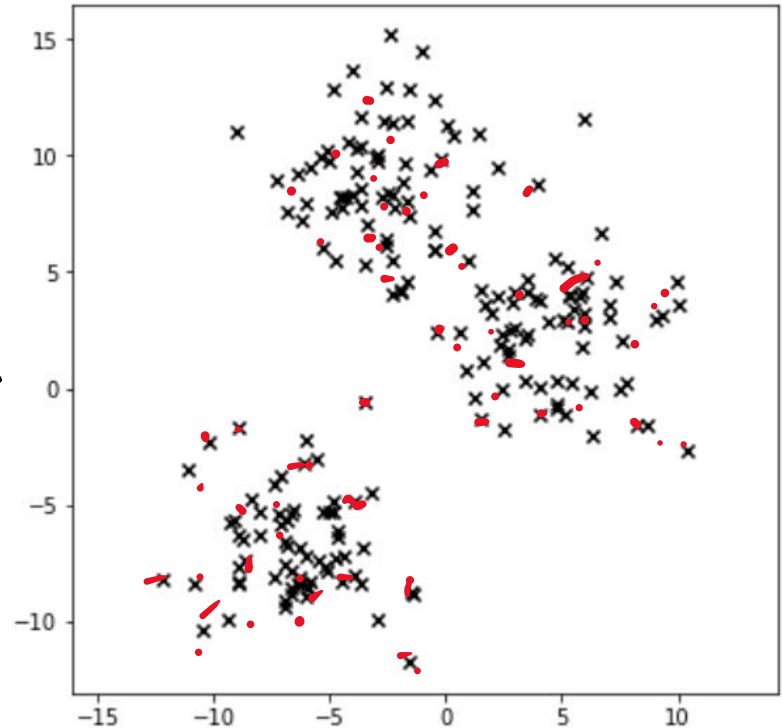
Mean-shift clustering

Inicialização:

Escolhe vários pontos do dataset como "sementes"

Loop:

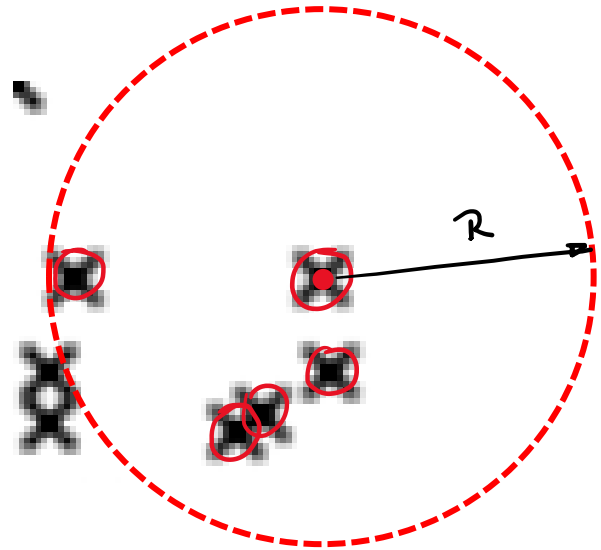
- Cada semente identifica seus novos "amigos" num raio R



Mean-shift clustering

Loop:

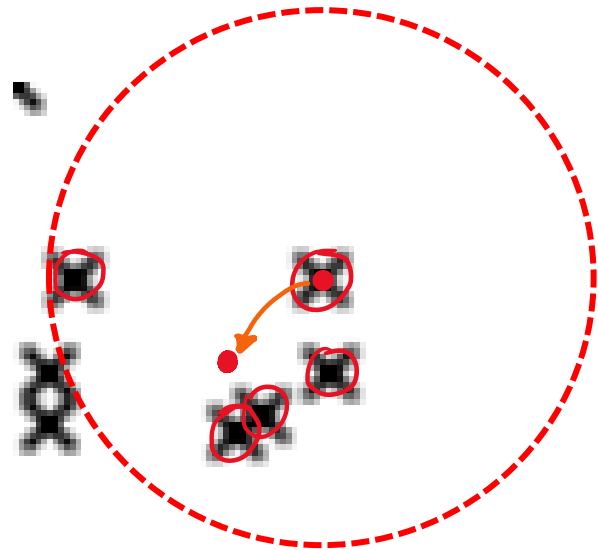
- Cada semente identifica seus novos "amigos" num raio R



Mean-shift clustering

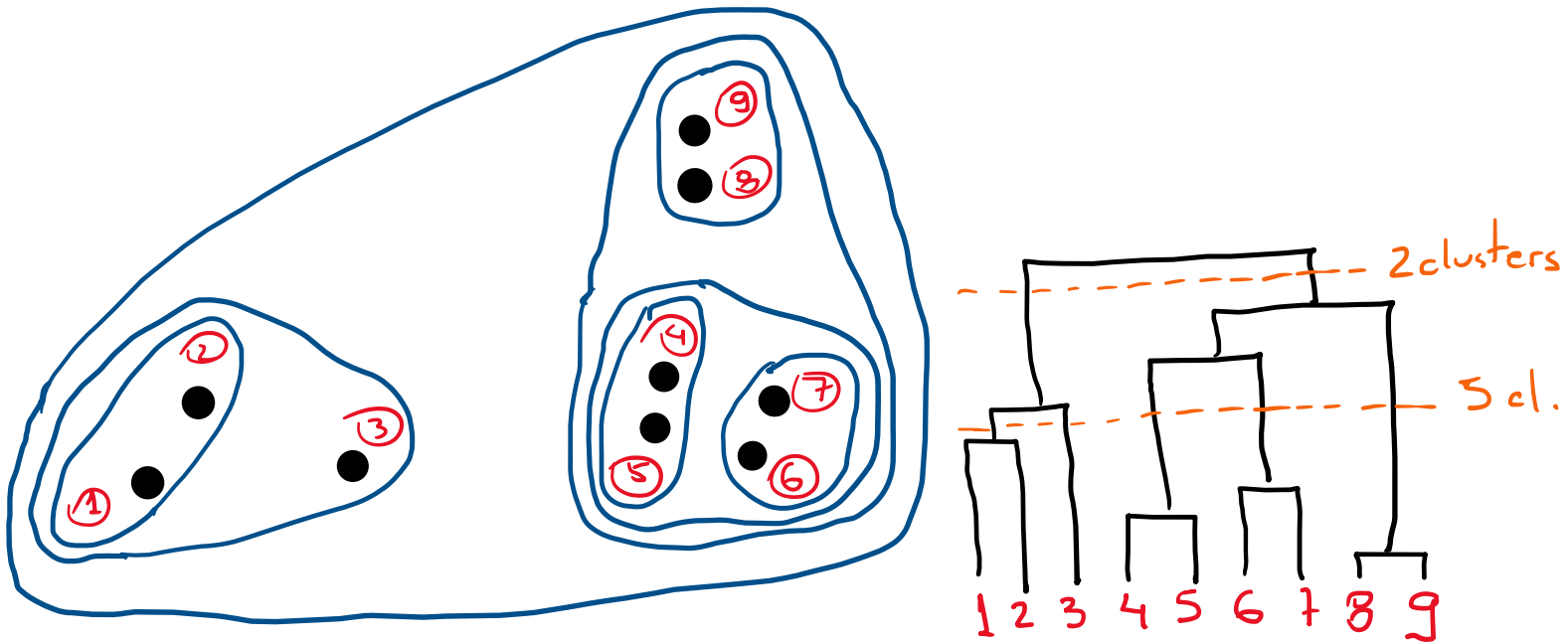
Loop:

- "Corre abraçar os amigos!"
Recalcula posição da semente como a média dos pontos selecionados
- Se duas sementes "se encontram"
→ fusão!



Repetir até convergir

Agglomerative clustering



The background of the slide is composed of several concentric, partial circular arcs in red and grey, creating a dynamic, layered effect. The arcs are of varying thicknesses and are positioned at different radii from the center, which is where the text is located.

Insper