K-Means / Clustering

Lucas Cambuim

Introdução

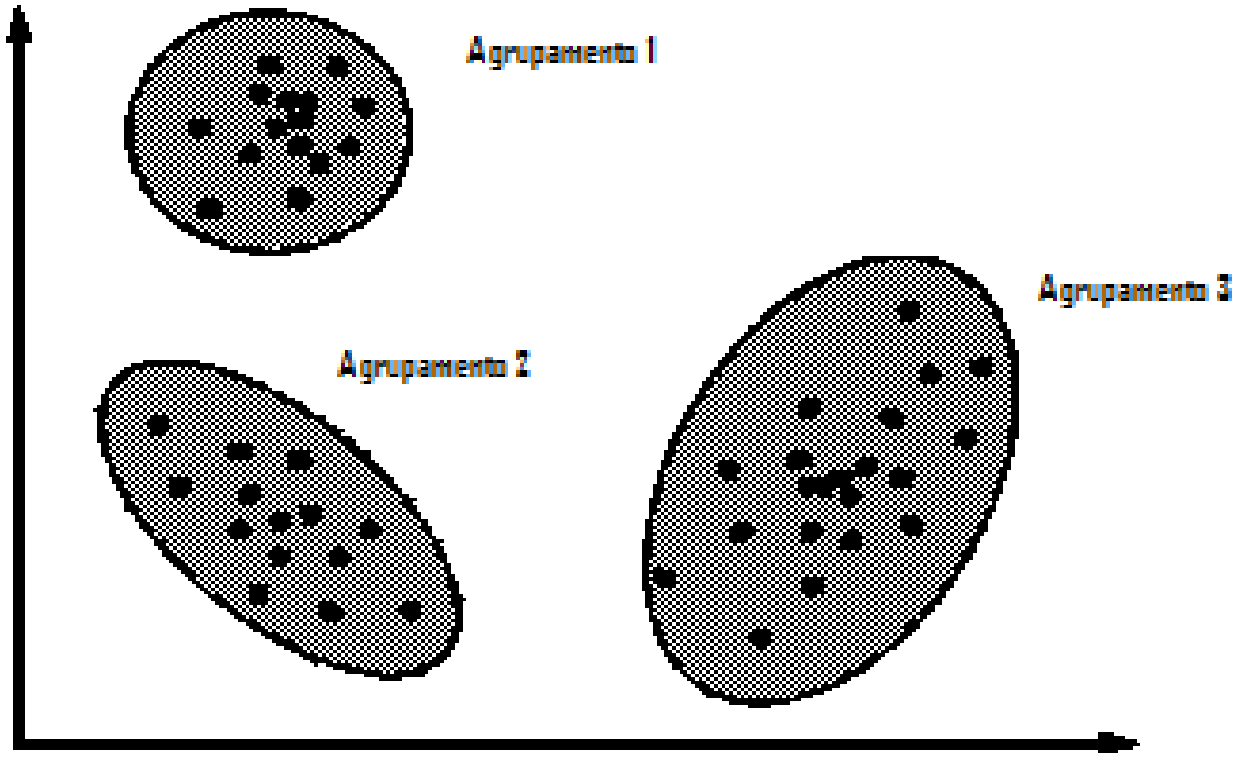
- A Análise de Agrupamentos (*Clustering Analysis*) tem por objetivo a separação de um conjunto de dados em grupos, de forma que objetos de um mesmo conjunto sejam mais similares entre si que objetos pertencentes a conjuntos diferentes

- Aplicação em diversas áreas: biometria, mineração de dados, engenharia, ciências sociais, medicina, etc.

- A qualidade das partições finais irá depender da técnica utilizada para a realização da tarefa de agrupamento.

Introdução

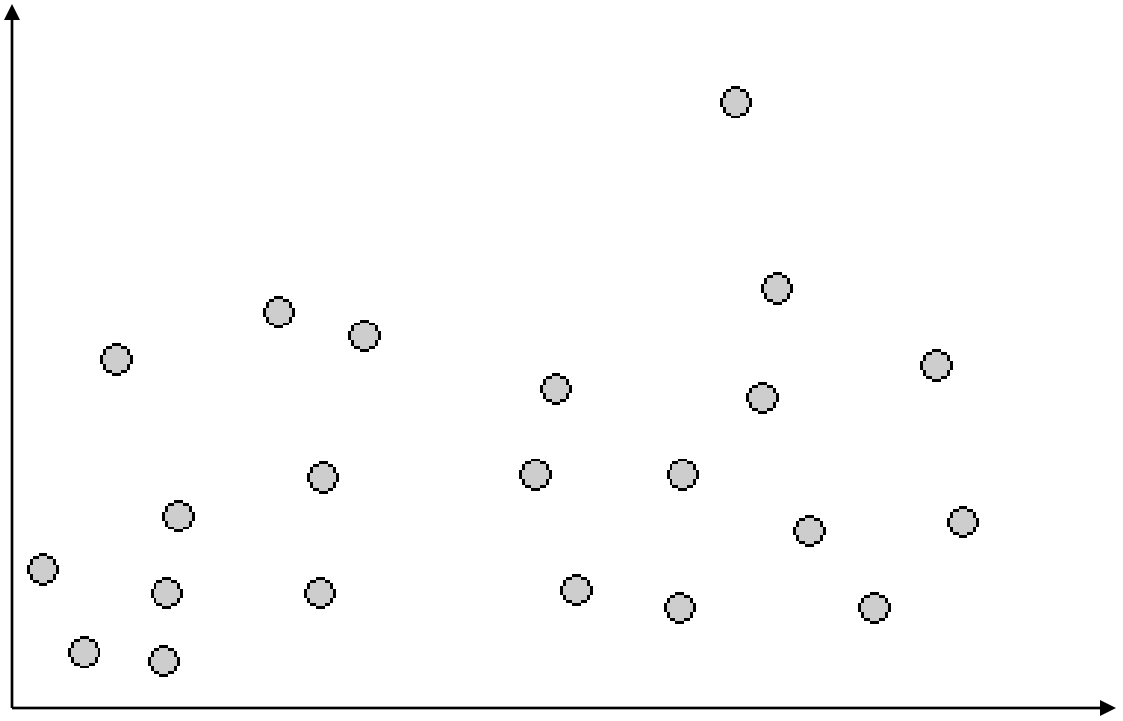
- Exemplo de agrupamentos:



Aprendizagem Não-Supervisionada

- O que pode ser feito quando se tem um conjunto de exemplos mas não se conhece as categorias envolvidas?

Como classificar esses pontos?



Por que estudar esse tipo de problema?

Aprendizagem Não-Supervisionada

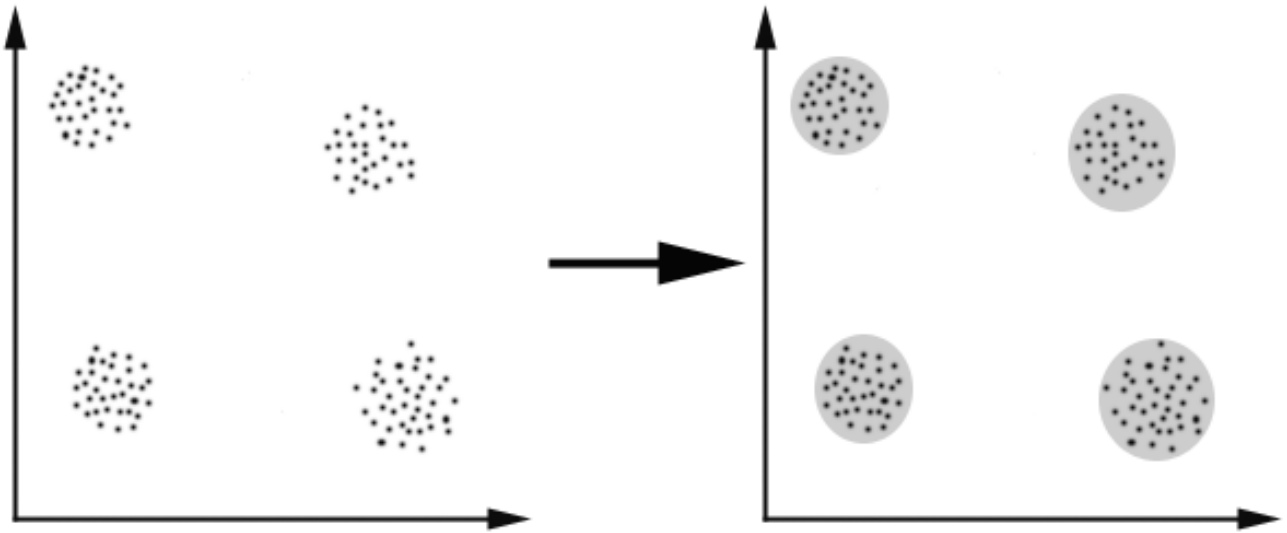
- Primeiramente, coletar e rotular bases de dados pode ser extremamente caro.
  - Ex: Gravar voz é barato, mas rotular todo o material gravado é caro.
- Segundo, muitas vezes não se tem conhecimento das classes envolvidas.
  - Trabalho exploratório nos dados (ex. *Data Mining*.)

Aprendizagem Não-Supervisionada

- Pré-classificação:
  - Suponha que as categorias envolvidas são conhecidas, mas a base não está rotulada.
  - Pode-se utilizar a aprendizagem não-supervisionada para fazer uma pré-classificação, e então treinar um classificador de maneira supervisionada.

Clustering

- É a organização dos objetos similares (em algum aspecto) em grupos.



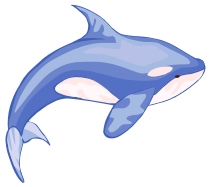
Quatro grupos (clusters)

O que é formação de agrupamentos (clustering)?

❖ Dado um conjunto de objetos, colocar os objetos em grupos baseados na similaridade entre eles

❖ Como agrupar os animais seguintes?

❖ Inerentemente é um problema não definido claramente



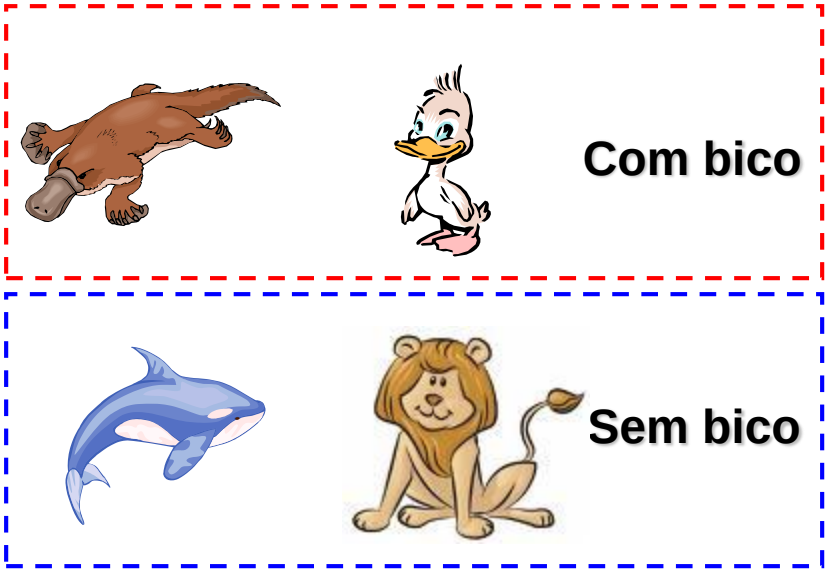
Adaptado de material de Marcílio C. P. de Souto - DIMAp/UFRN

O que é formação de agrupamentos (clustering)?

❖ Dado um conjunto de objetos, colocar os objetos em grupos baseados na similaridade entre eles

❖ Como agrupar os animais seguintes?

❖ Inerentemente é um problema não definido claramente

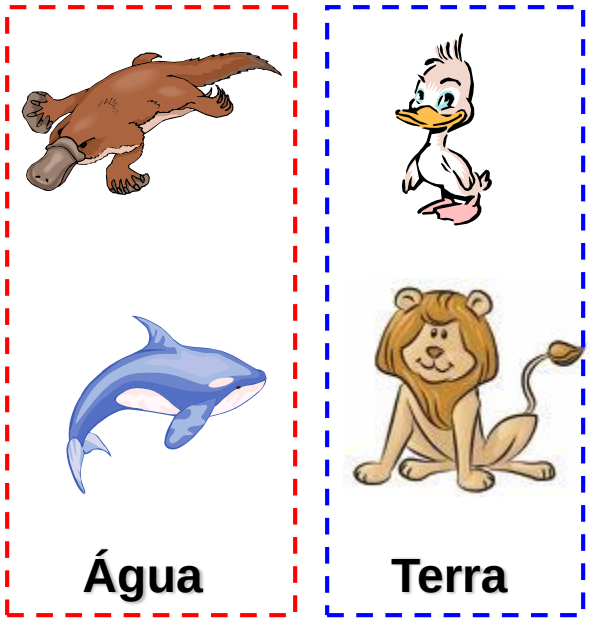


O que é formação de agrupamentos (clustering)?

❖ Dado um conjunto de objetos, colocar os objetos em grupos baseados na similaridade entre eles

❖ Inerentemente é um problema não definido claramente

❖ Como agrupar os animais seguintes?

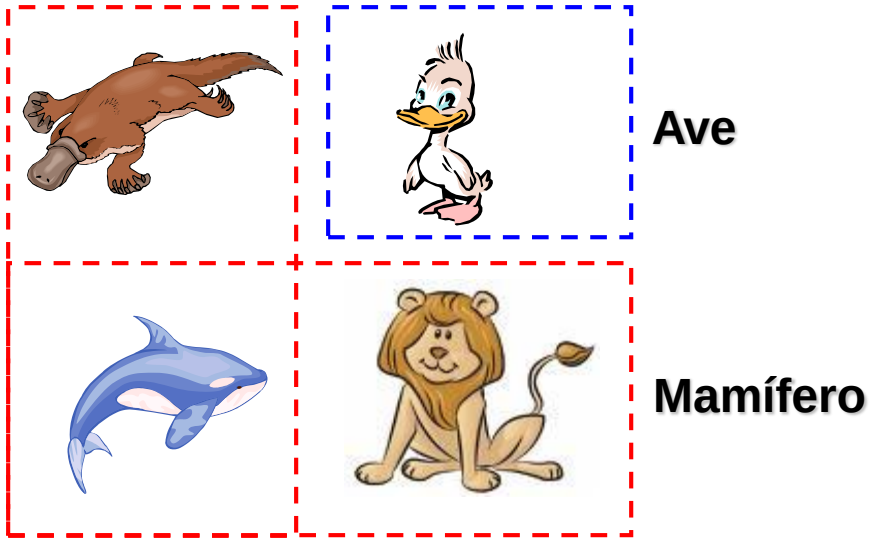


O que é formação de agrupamentos (clustering)?

❖ Dado um conjunto de objetos, colocar os objetos em grupos baseados na similaridade entre eles

❖ Como agrupar os animais seguintes?

❖ Inerentemente é um problema não definido claramente

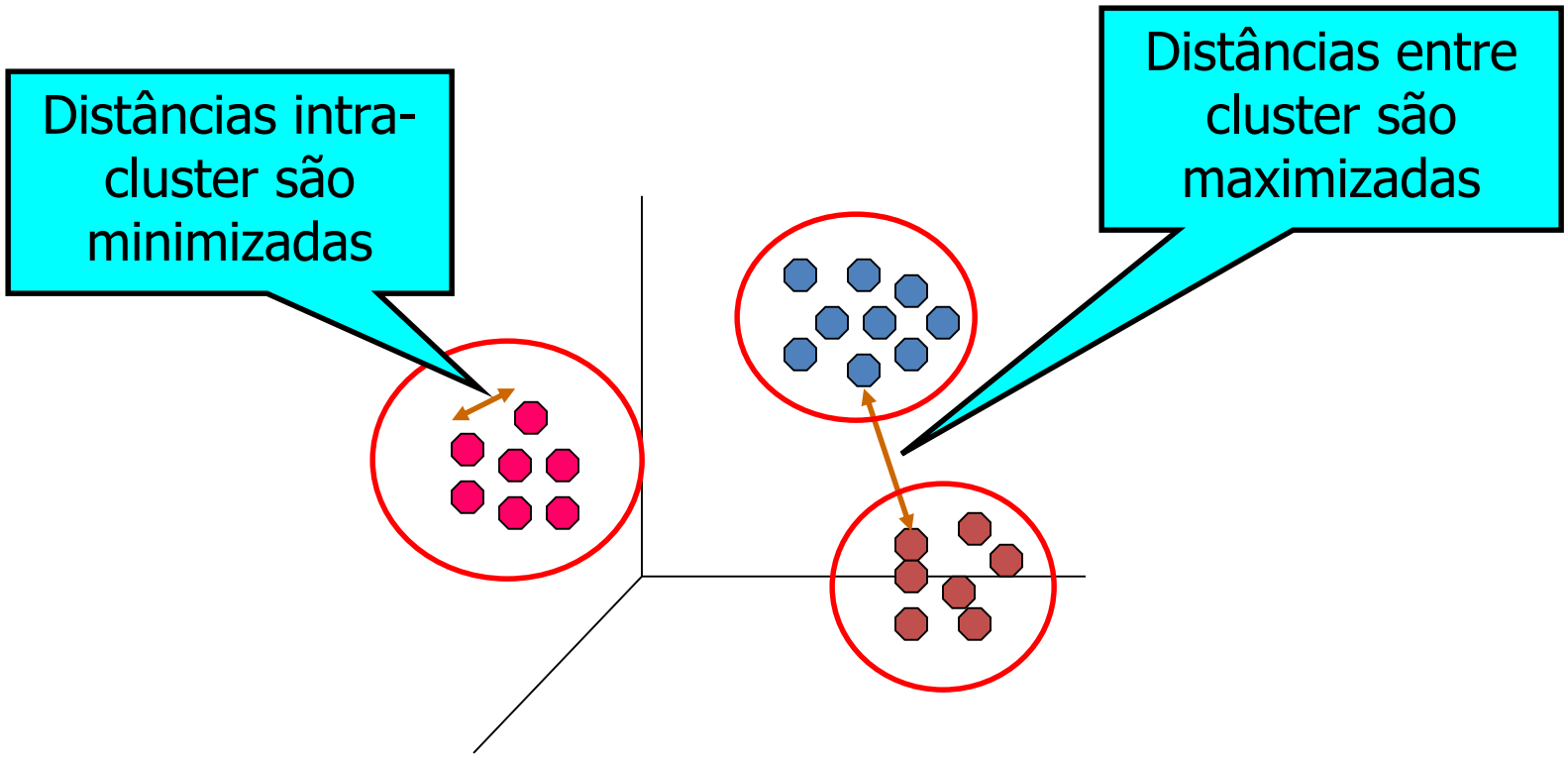


Cluster

- Uma coleção de objetos que são similares entre si, e diferentes dos objetos pertencentes a outros *clusters*.
- Isso requer uma medida de similaridade.
- No exemplo anterior, a similaridade utilizada foi a *distância.*
  - *Distance-based Clustering*

O que é formação de agrupamentos (clustering)?

- Dado um conjunto de objetos, colocar os objetos em grupos baseados na similaridade entre eles

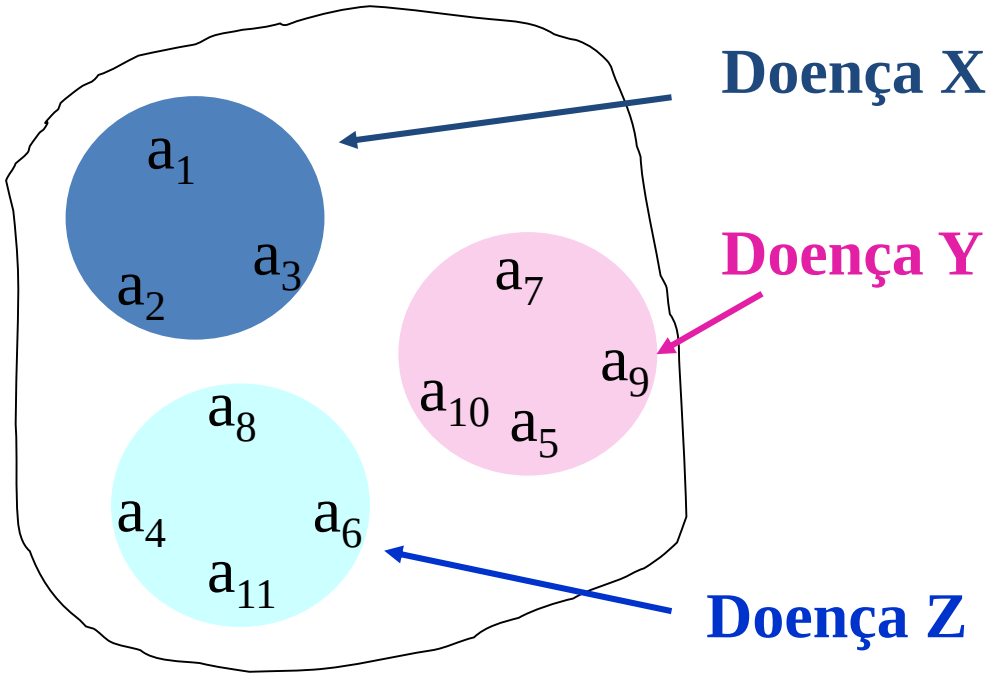


Agrupamento -Análise de Clusters

a_1	a	F	1	0	1	1
a_2	b	M	0	0	1	1
	c	F	1	1	1	0
•	d	F	1	0	0	0
•	e	M	1	1	0	1
•						

Nome Sexo Sintomas

Número de Clusters = 3



Conceito = Doença

Análise de Clusters: Objetivos

- **Compreensão dos dados**
  - Existe algum conceito inerente a cada grupo.
  - Que conceito é este ?
- **Utilidade em outras tarefas**

  Cada cluster pode ser representado por um *objeto protótipo* que caracteriza o cluster

  – ***Sumarização*** : Algoritmos aplicados em grandes volumes de dados podem ser aplicados apenas aos protótipos, reduzindo assim o tempo de execução

  – ***Compressão*** : o objeto protótipo representa cada objeto dentro do seu cluster

  – ***Otimização do cálculo dos vizinhos mais próximos***:

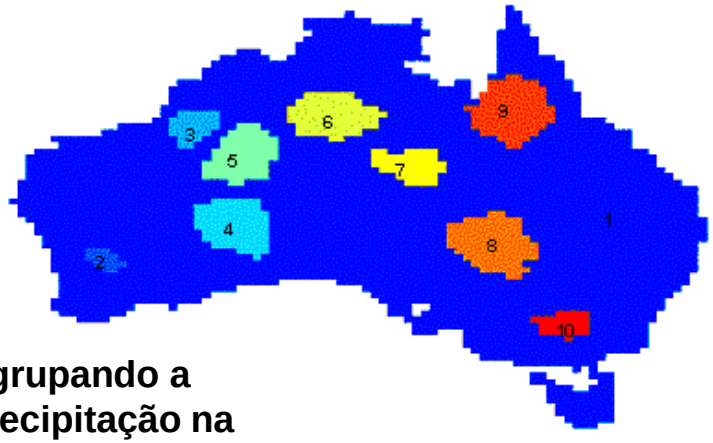
  Se dois protótipos estão distantes então os objetos nos respectivos clusters também estão distantes.

  Os objetos mais próximos do objeto X devem ser procurados no cluster correspondente ao protótipo mais próximo de X.

Aplicações de *clustering*

- **Entendimento**
  - Agrupar documentos relacionados, agrupar proteínas com funcionalidades similares, agrupar ações com as mesmas flutuações de preço
- **Sumarização**
  - Reduzir o tamanho de grandes conjuntos de dados

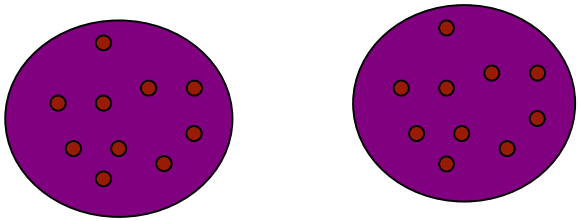
	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP



Agrupando a precipitação na Austrália

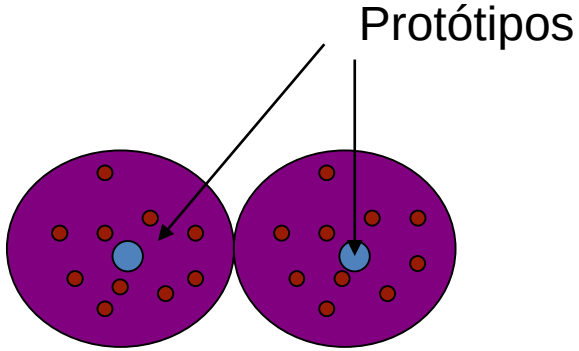
O que é um cluster ?

Como definir a noção de Cluster ?



Bem separados

Um *cluster* é um conjunto de objetos no qual cada objeto está mais próximo (ou é mais similar) a objetos dentro do cluster do que qualquer objeto fora do cluster.



Baseados em Protótipos

Um *cluster* é um conjunto de objetos no qual cada objeto está mais próximo ao *protótipo que define o cluster* do que dos protótipos de quaisquer outros clusters.
Em geral: Protótipo = centróide

Os clusters encontrados tendem a ser *globulares*.

Tipos de Clusterização

- Algoritmos hierárquicos: estes encontram clusters sucessivos usando clusters previamente estabelecidos.
  - Agglomerativo ("de baixo para cima"): os algoritmos agglomerativos começam com cada elemento como um cluster separado e mescla-os em clusters sucessivamente maiores
  - Divisivo ("de cima para baixo"): os algoritmos divisivos começam com um conjunto inteiro e dividi-o sucessivamente.
- Clusterização particional: os algoritmos particionais determinam todos os clusters ao mesmo tempo.
  - Eles incluem:
    - K-means e derivados
    - Fuzzy c-means clustering
    - QT clustering algorithm

Medidas de distâncias comuns

- A medida de distância determinará como a similaridade de dois elementos é calculada e influenciará a forma dos clusters
- A distância euclidiana é dada por:

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

- A distância de Manhattan:

$$d(x,y) = \sum_{i=1}^{p} |x_i - y_i|$$

- A distância máxima:

$$d(x,y) = \max_{1 \le i \le p} |x_i - y_i|$$

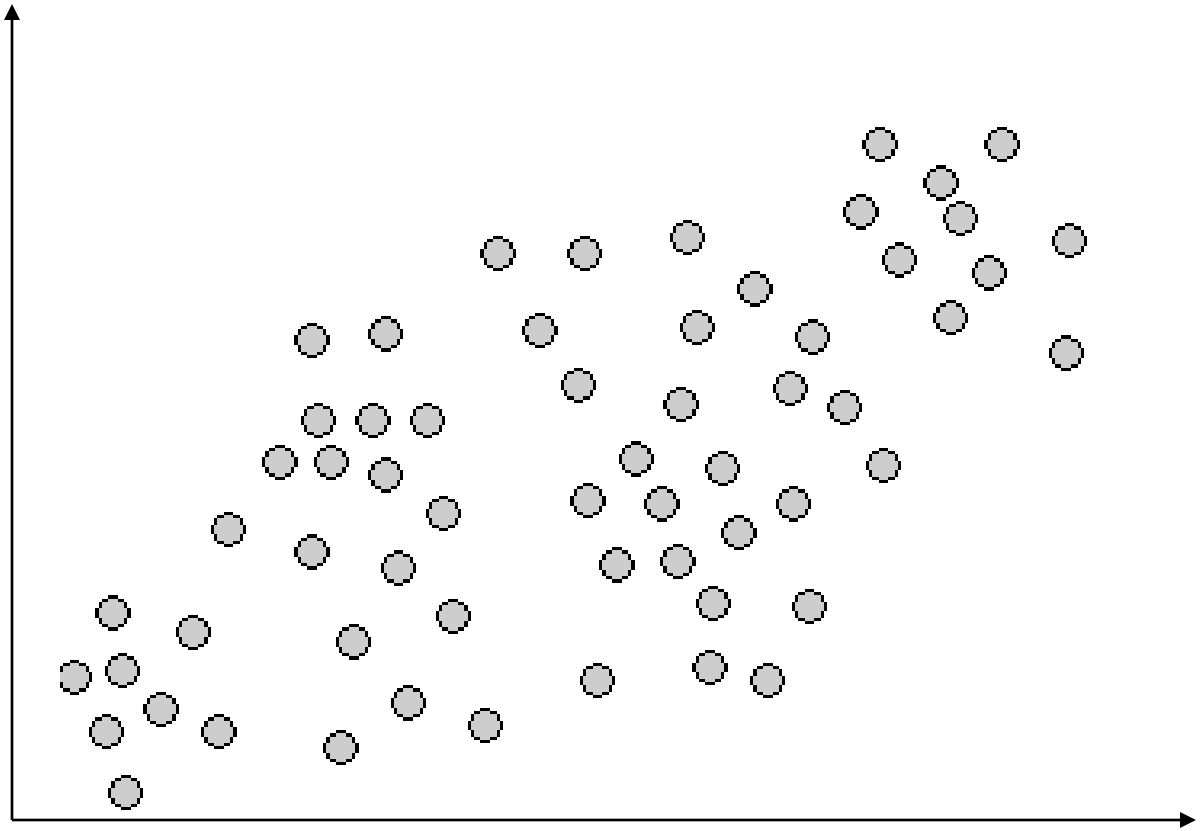
k-Means Clustering

- É a técnica mais simples de aprendizagem não supervisionada.
- Consiste em fixar *k* centróides (de maneira aleatória), um para cada grupo (clusters).
- Associar cada indivíduo ao seu centróide mais próximo.
- Recalcular os centróides com base nos indivíduos classificados.

Algoritmo *k-Means*

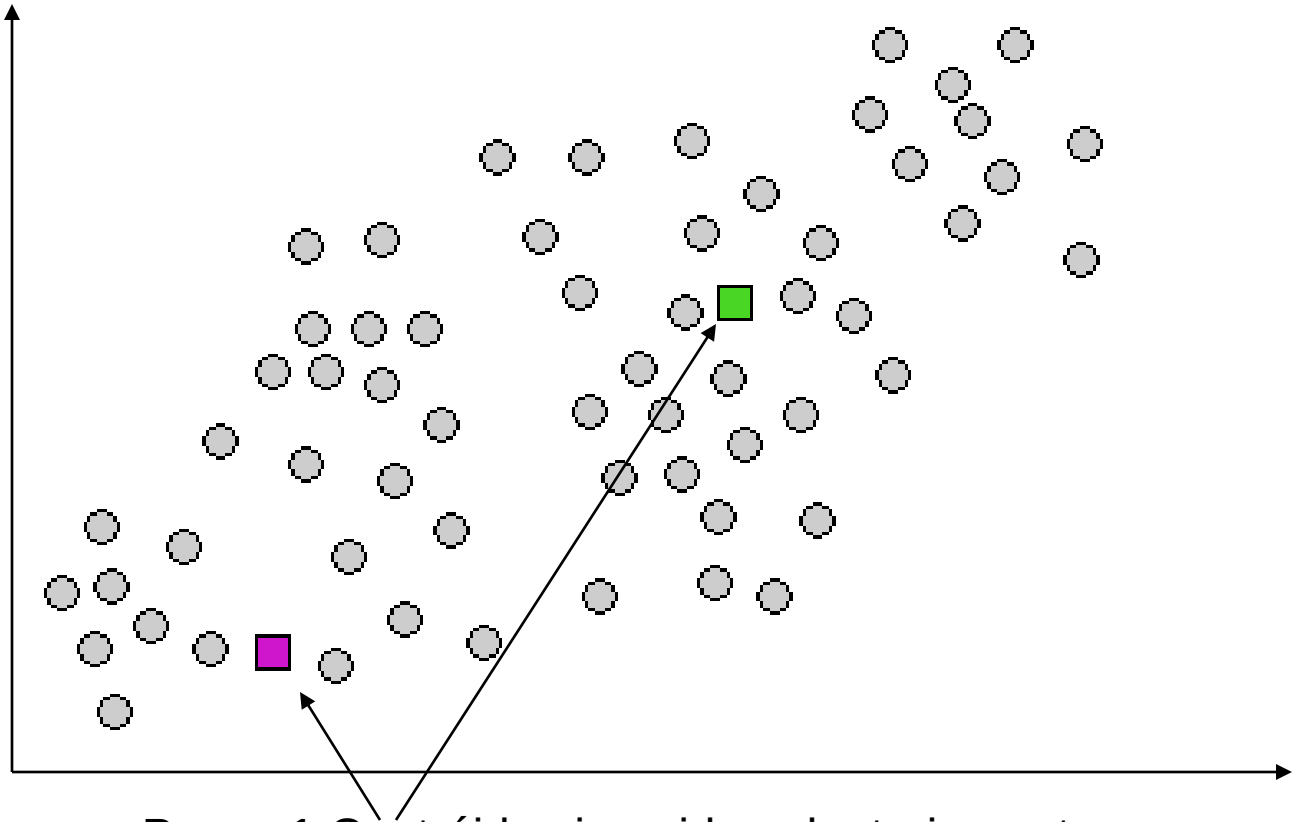
1. Determinar os centróides
2. Atribuir a cada objeto do grupo o centróide mais próximo.
3. Após atribuir um centróide a cada objeto, recalcular os centróides.
4. Repetir os passos 2 e 3 até que os centróides não sejam modificados.

k-Means – Um Exemplo



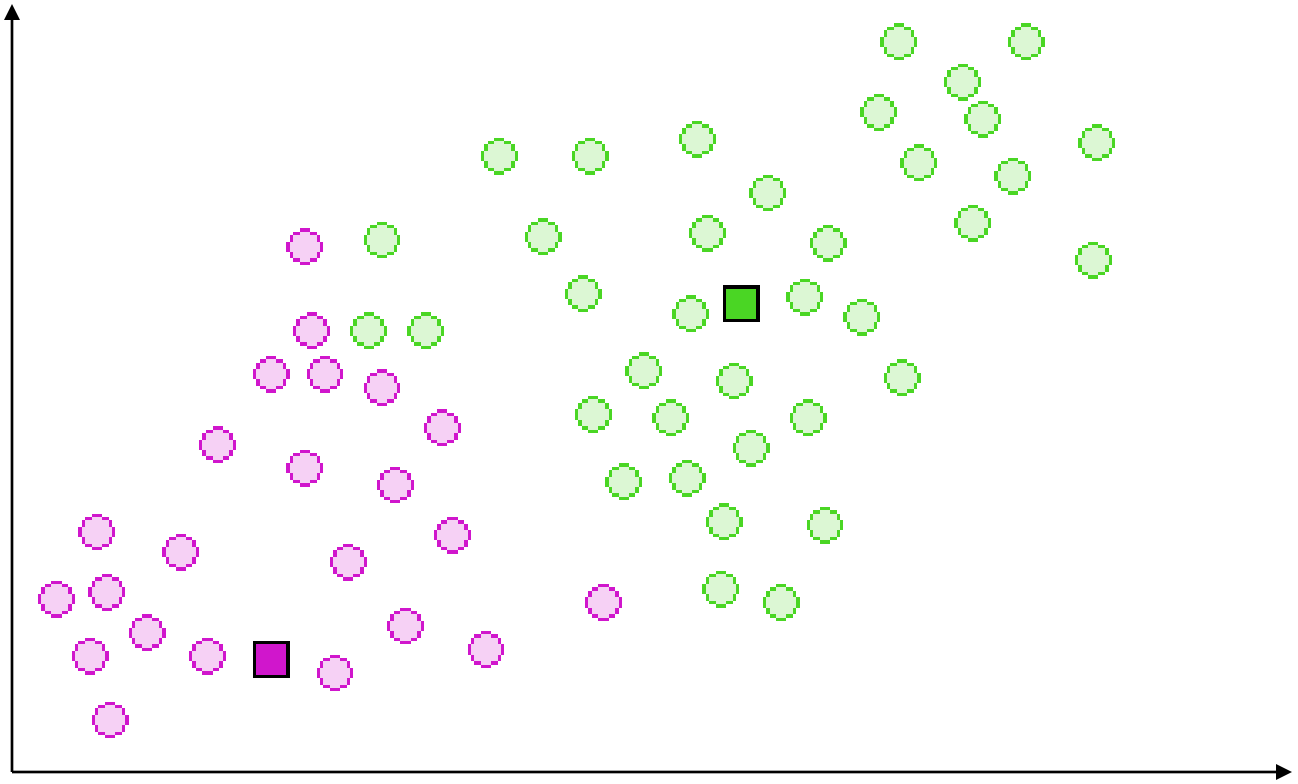
Objetos em um plano 2D

k-Means – Um Exemplo



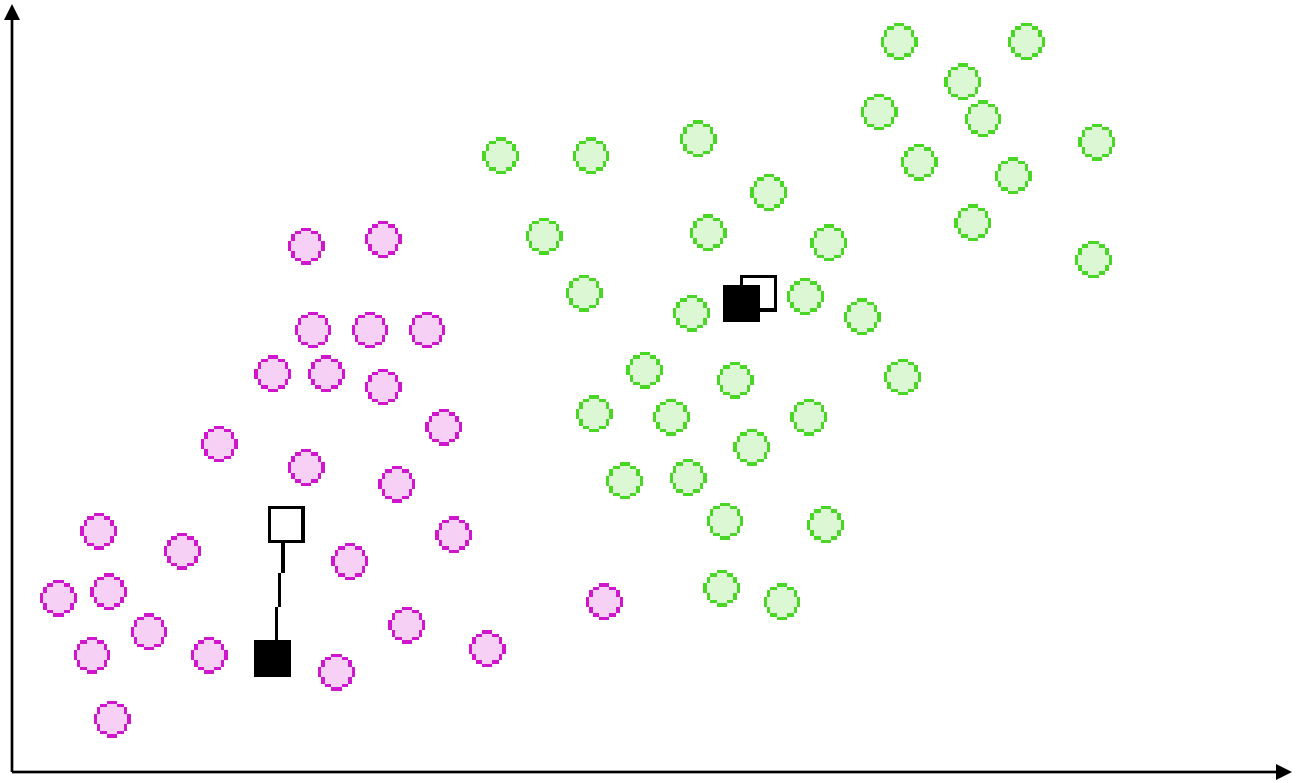
Passo 1:Centróides inseridos aleatoriamente

k-Means – Um Exemplo



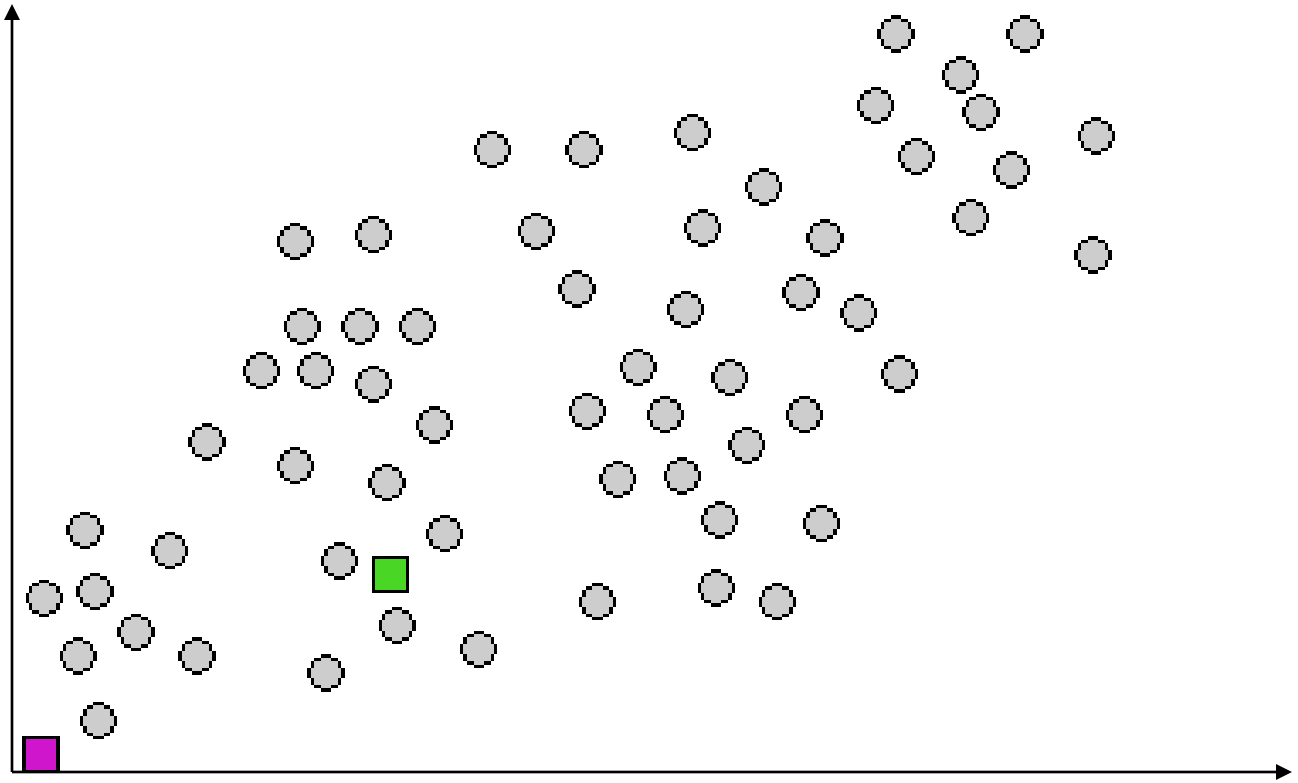
Passo 2: Atribuir a cada objeto o centróide mais próximo

k-Means – Um Exemplo



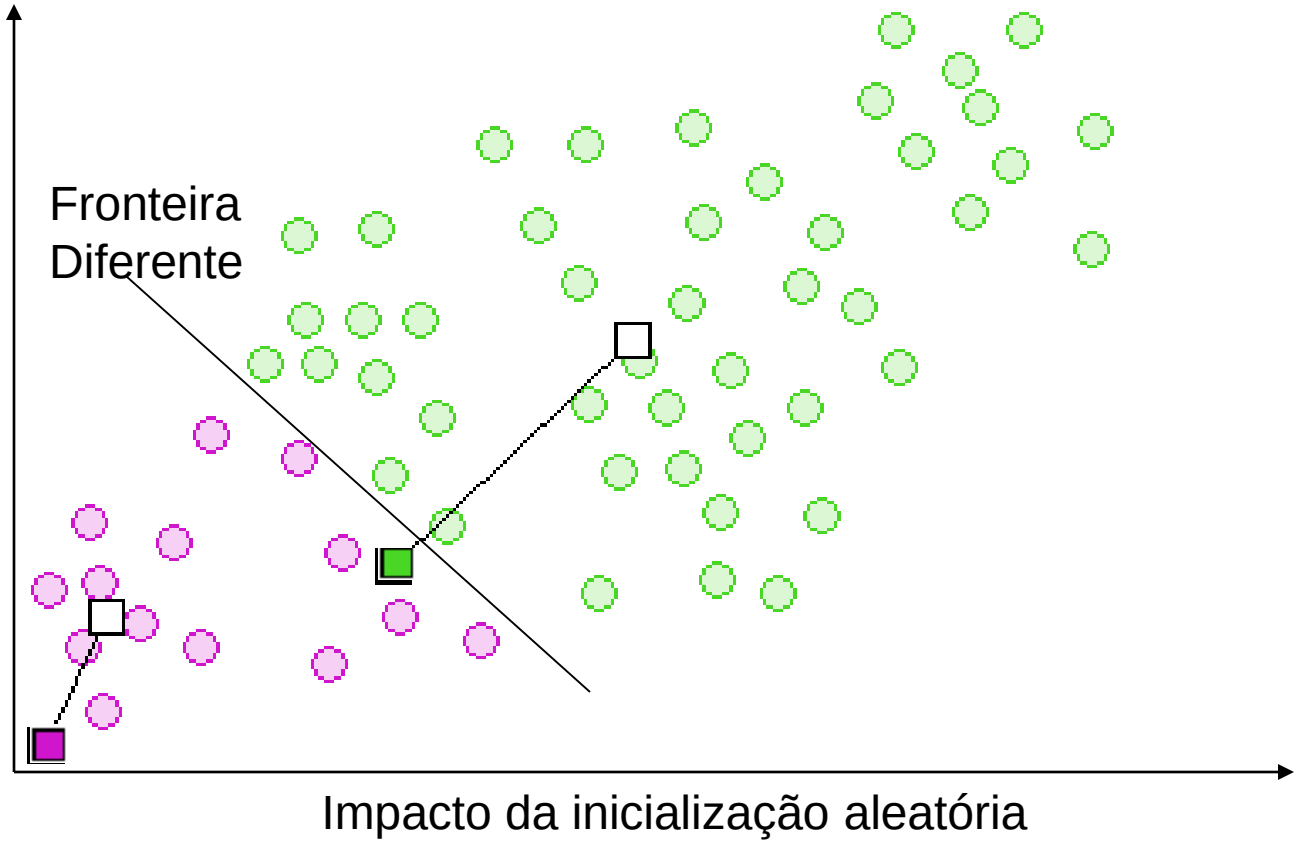
Passo 3: Recalcular os centróides

k-Means – Um Exemplo



Impacto da inicialização aleatória.

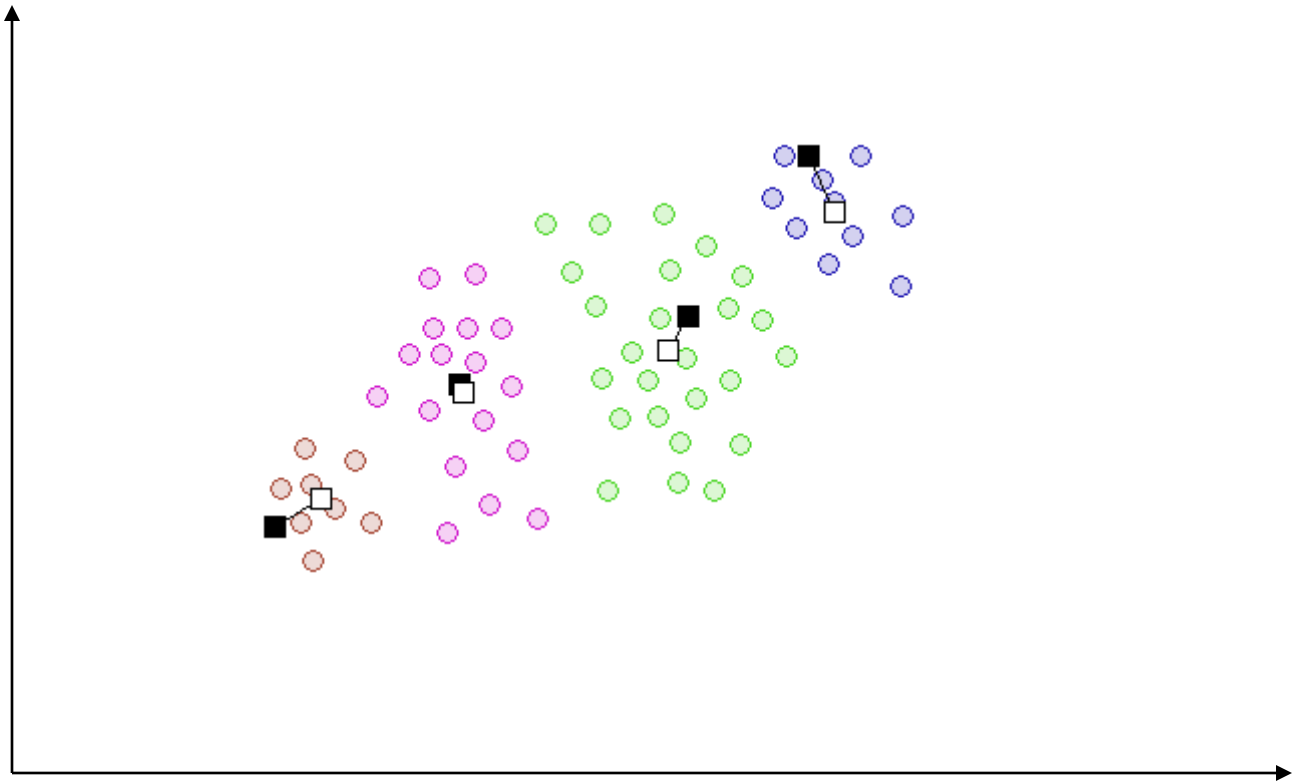
k-Means – Um Exemplo



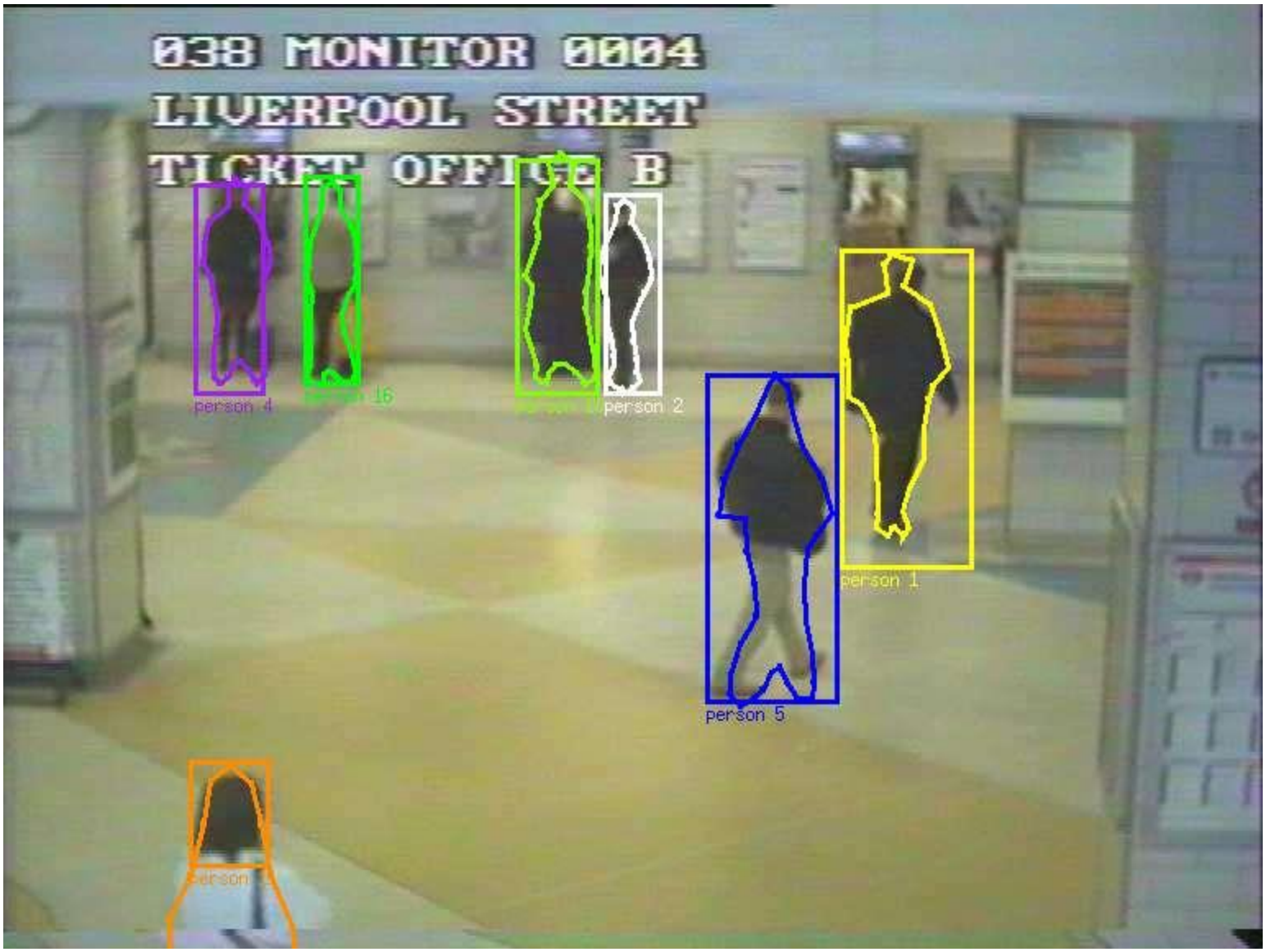
k-Means – Inicialização

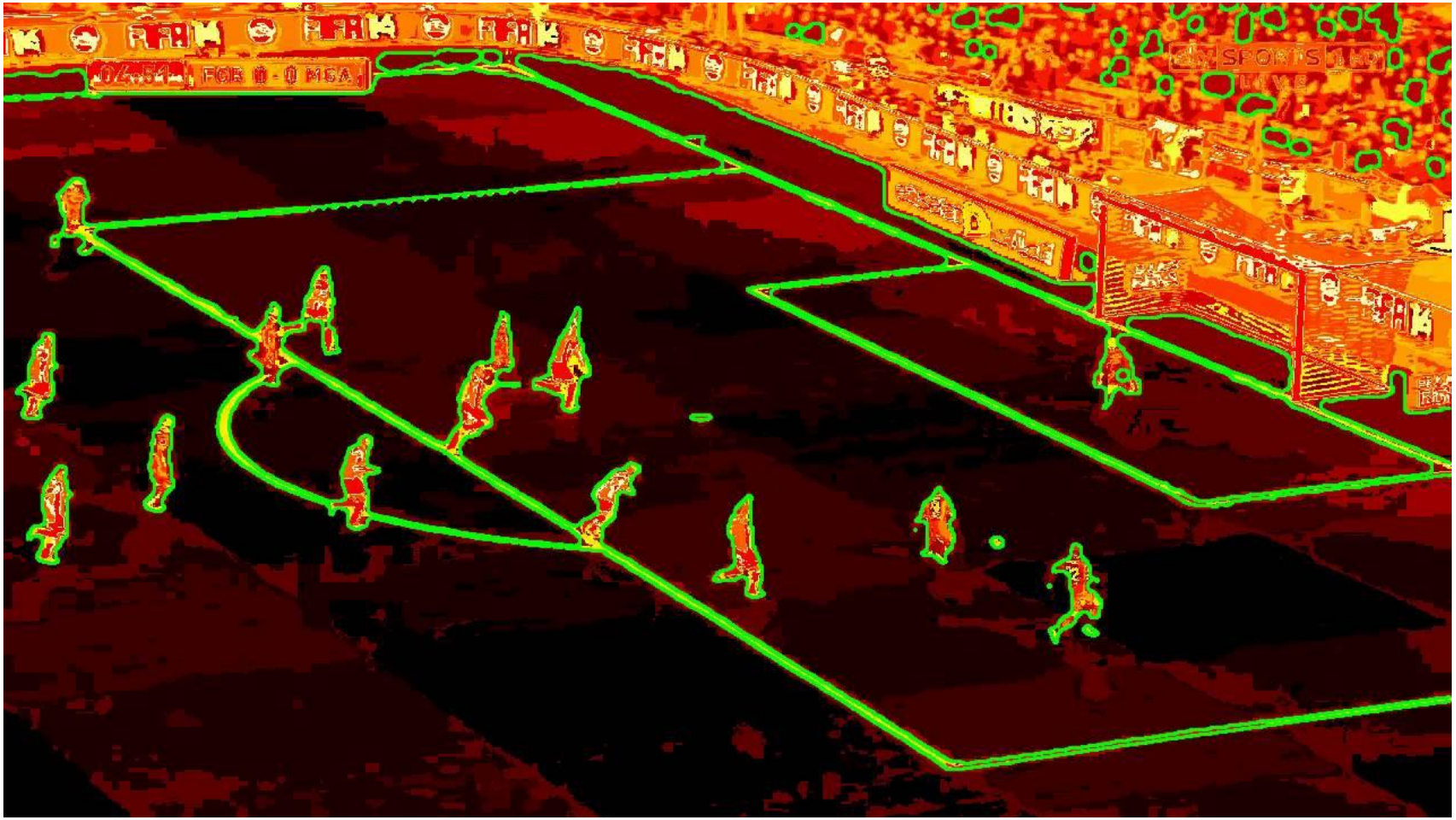
- Importância da inicialização.
- Quando se têm noção dos centróides, pode-se melhorar a convergência do algoritmo.
- Execução do algoritmo várias vezes, permite reduzir impacto da inicialização aleatória.

k-Means – Um Exemplo



4 Centróides


038 MONITOR 0004
LIVERPOOL STREET
TICKET OFFICE B
person 4
person 16
person 2
person 1
person 5

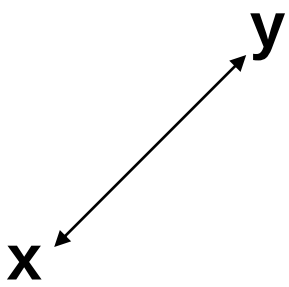



Original Image
K = 2
K = 4
K = 8

Calculando Distâncias

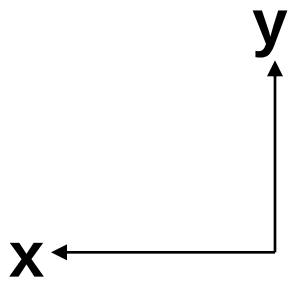
- Distância Euclidiana

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



- Manhattan (City Block)

$$d = \sum_{i=1}^{n} |x_i - y_i|$$



Calculando Distâncias

- Minkowski
  - Parâmetro r
    - r = 2, distância Euclidiana
    - r = 1, City Block

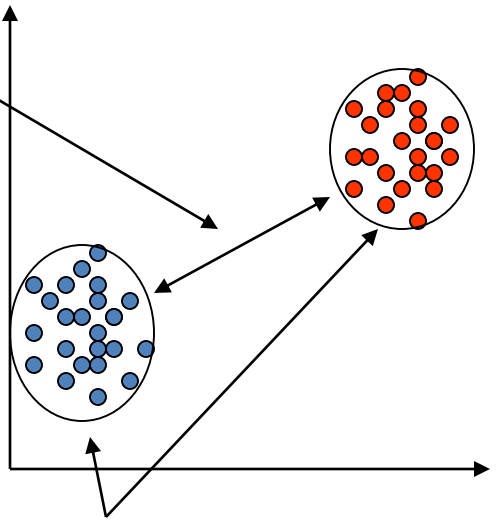
$$d = \left(\sum_{i=1}^{n} \left(x_i - y_i \right)^r \right)^{1/r}$$

Critérios de Dispersão

- Relação Within-Between

Caso ideal

Alto between (S_b)
Clusters distantes um do outro.

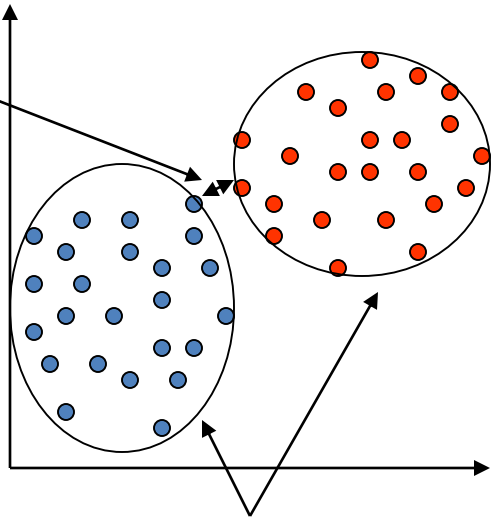


Baixo within (S_w)
(boa compactação)

Critérios de Dispersão

Caso não ideal

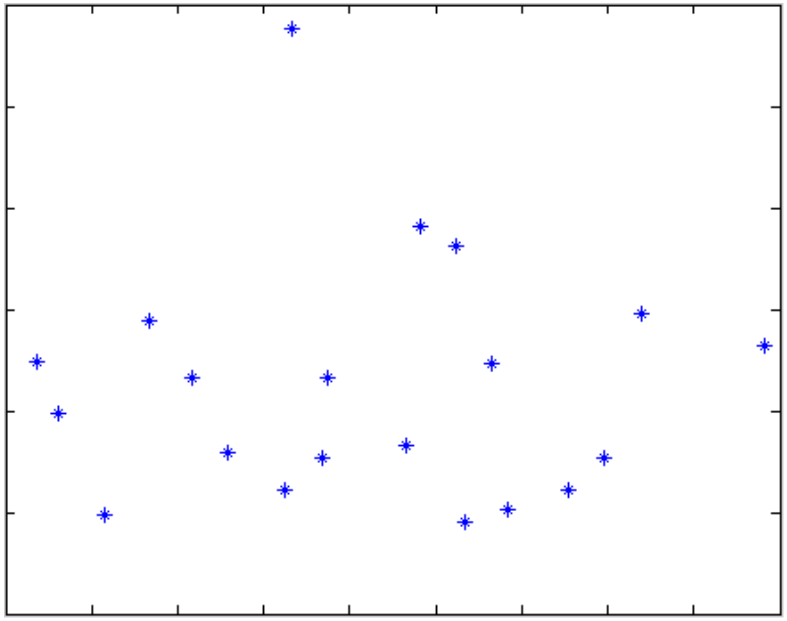
Baixo between (S_b)
Baixa distância entre os clusters.



Clusters dispersos
Alto within

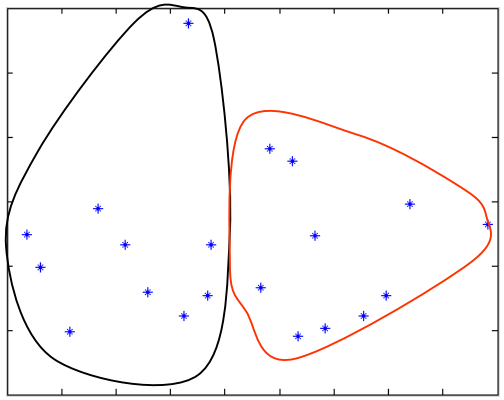
Critérios de Dispersão

- Podemos entender melhor os critérios de dispersão analisando o seguinte exemplo:

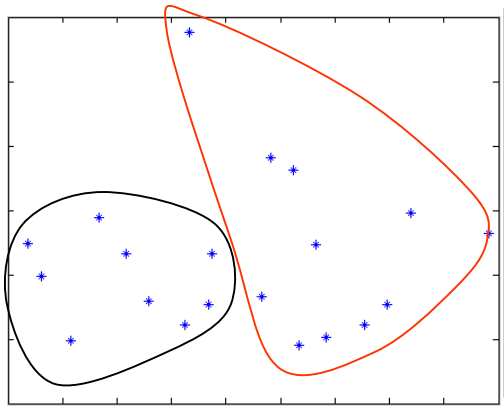


Diferentes clusters para c=2 usando diferentes critérios de otimização

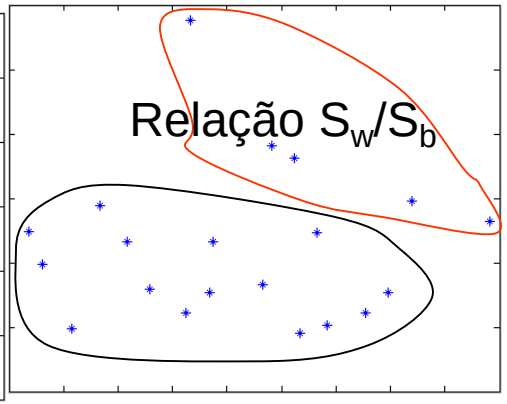
Erro Quadrado



S_w



Relação S_w/S_b



Algumas Aplicações de *Clustering*

- *Marketing*: Encontrar grupos de consumidores com comportamento similares
- Biologia: Classificar grupos de plantas e animais.
- Bibliotecas: Organização de livros.
- Administração: Organização de cidades, classificando casas de acordo com suas características.
- WWW: Classificação de conteúdos.

Problemas

- Vetores de característica muito grandes: tempo de processamento elevado.
- Definição da melhor medida de distância: Depende do problema. As vezes é difícil, especialmente quando se trabalha com grandes dimensões.
- O resultado do *clustering* pode ser interpretado de diferentes maneiras.