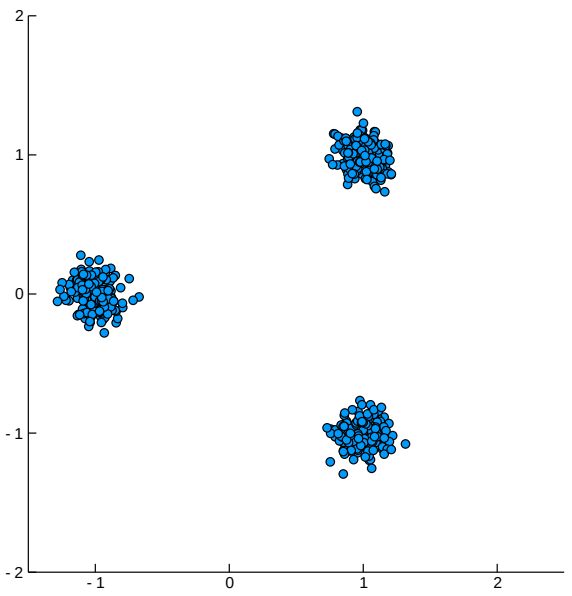
Aglutinação de dados via *k-means*

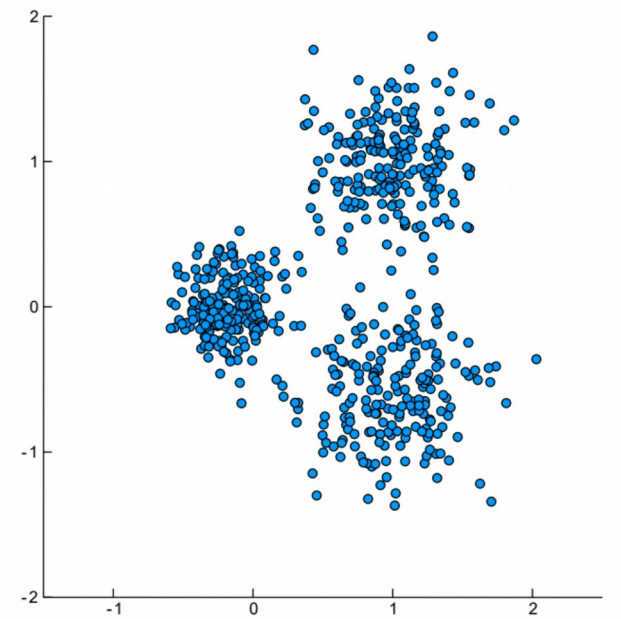
Bruno Santiago

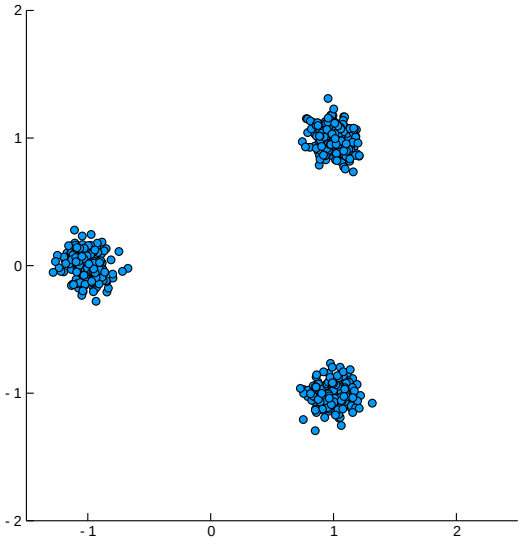
Professor Adjunto A
Departamento de Análise
Instituto de Matemática e Estatística
Universidade Federal Fluminense

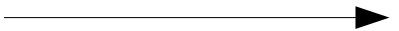
Aglutinação de Dados

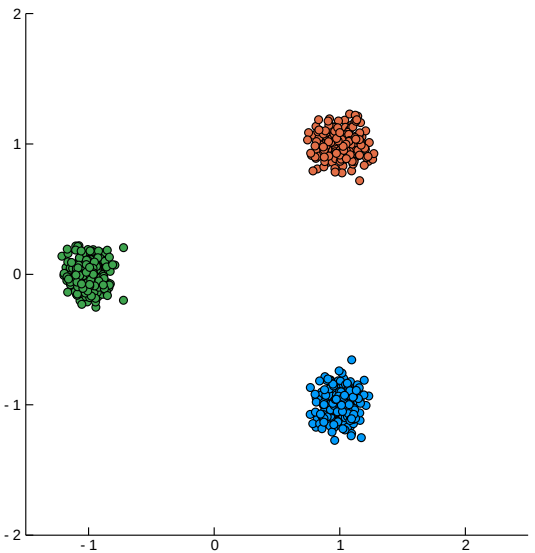
- Dado um conjunto $\{x_1, ..., x_n\} \subset \mathbb{R}^d$ grande de pontos num espaço euclideano (*data set),* como classificá-lo em "aglutinações/ agrupamentos"?

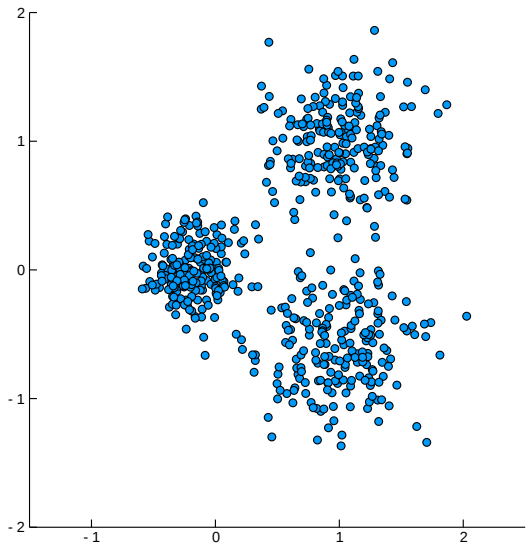


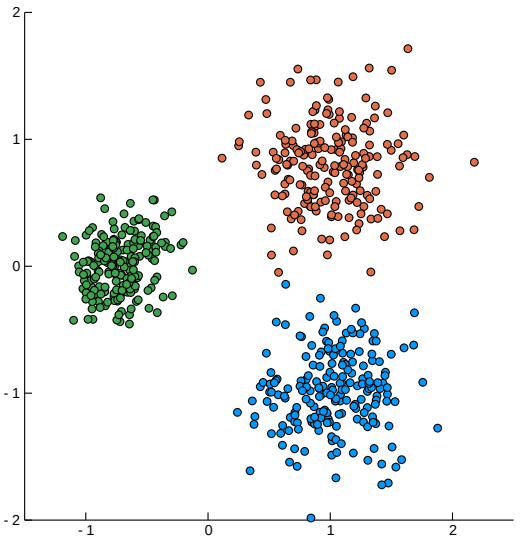



2
1
0
-1
-2
-1
0
1
2




2
1
0
-1
-2
-1
0
1
2


2
1
0
-1
-2
-1
0
1
2


2
1
0
-1
-2
-1
0
1
2

Motivações

- Algorítimos automáticos de descoberta de tópico
- Estudos sobre o câncer
- Estudos sobre o ENEM
- Padrões de uso de energia elétrica
- Padrão de consumo
- Etc...

Formalização do Problema

- Considere um conjunto de dados $X = \{x_1, ..., x_n\} \subset \mathbb{R}^d$.
- Um agrupamento de X em k pedaços é uma família de subconjuntos $G_1 \subset X, \ldots, G_k \subset X$ satisfazendo

1) $i \neq j \implies G_i \cap G_j = \emptyset$

2) $\cup_{\ell=1}^{k} G_\ell = X$

Função de Coloração

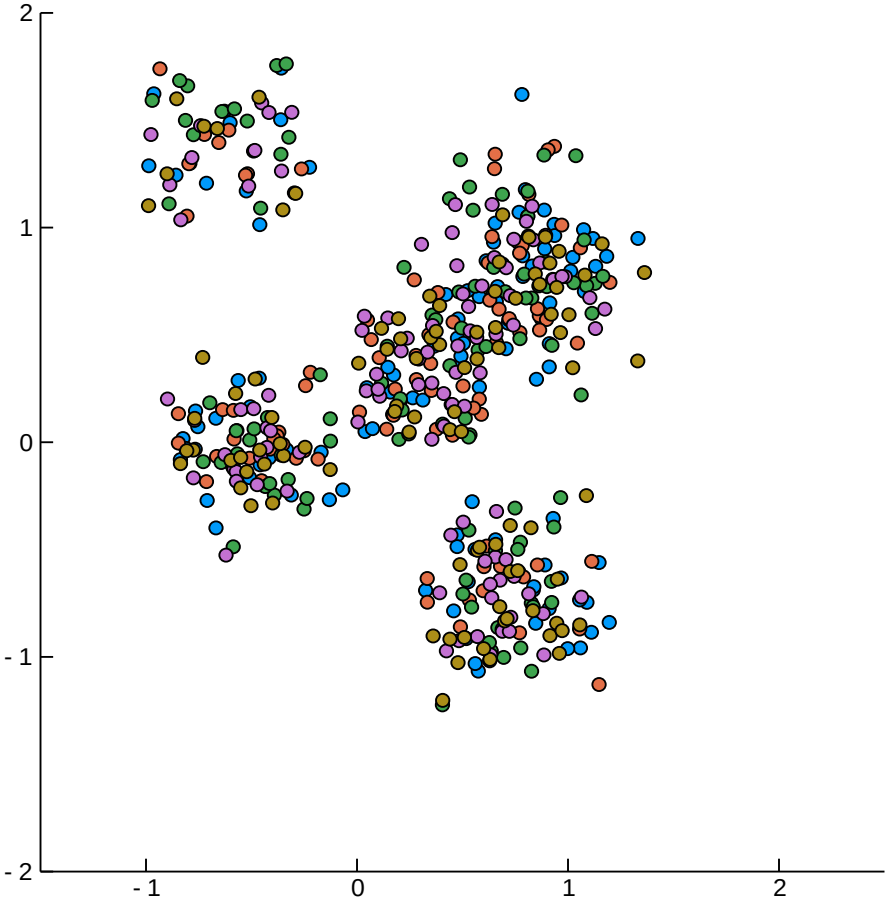
- Um agrupamento é o mesmo que atribuir cores a subconjuntos de forma que todo elemento tenha uma cor. Matematicamente, isso equivale a dar uma função sobrejetiva

$$c : [1, N] \to [1, k]$$

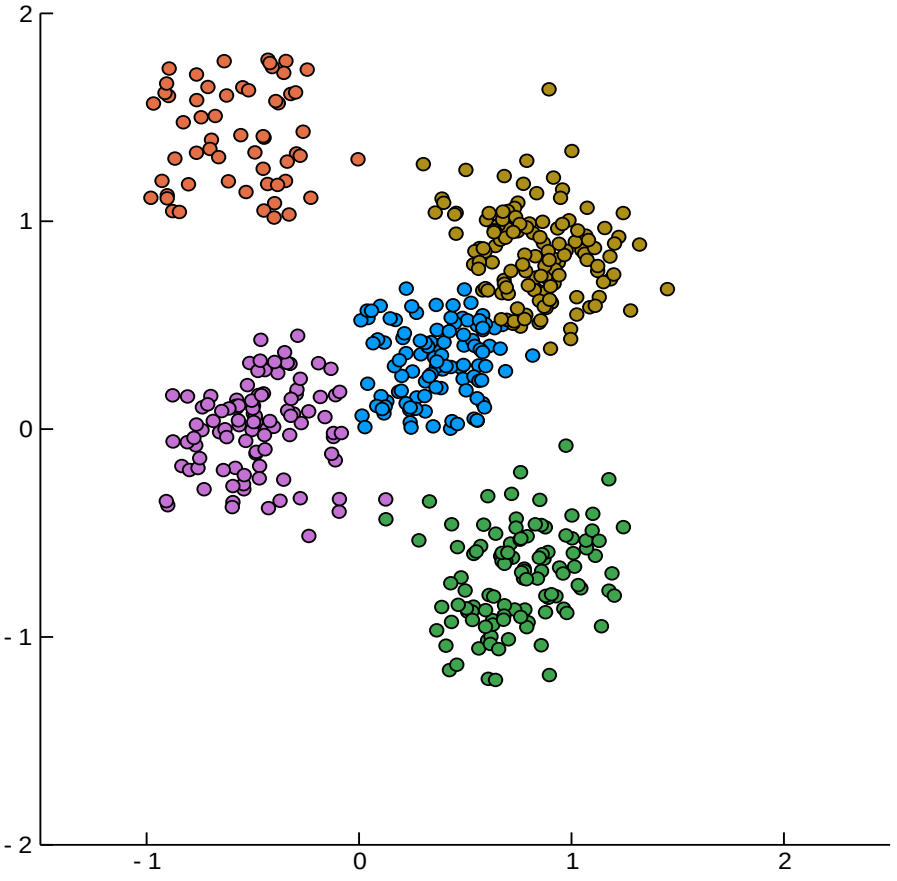
tal que

$$c(\ell) = j \iff x_\ell \in G_j$$

Cores atribuídas de forma aleatória

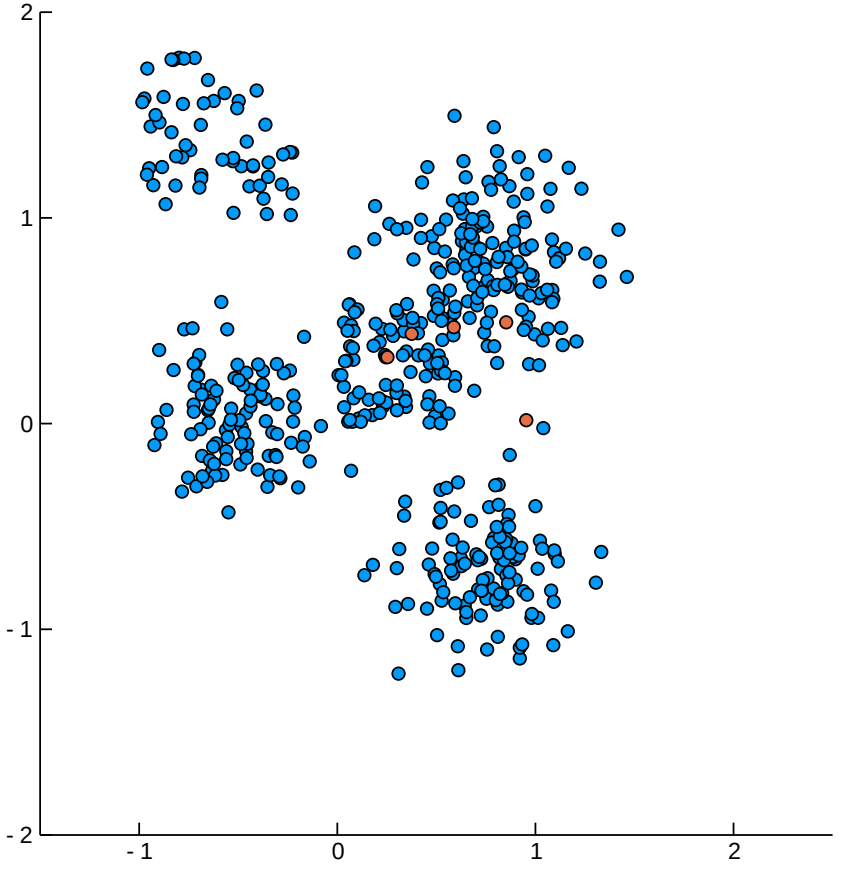


Cores atribuídas matematicamente (usando o algorítimo *k-means)*



Representantes do grupo

- Seja $G_1, ..., G_k$ um agrupamento do conjunto de dados X. Dado um conjunto qualquer $Z = \{z_1, ..., z_k\} \subset \mathbb{R}^d$ com k elementos, dizemos que o vetor z_ℓ é o representante do grupo G_ℓ .



Qualidade do agrupamento

- Seja $G_1, ..., G_k$ um agrupamento de X, traduzido pela função de coloração $c : \{1, ..., N\} \to \{1, ..., k\}$ com represantes $Z = \{z_1, ..., z_k\}$. A qualidade do agrupamento é o número

$$J(c, Z) \frac{1}{N} \sum_{\ell=1}^{N} \|x_\ell - z_{c(\ell)}\|^2.$$

Enunciado do Problema

Dado um subconjunto finito $X \subset \mathbb{R}^d$ *com N elementos, encontrar um agrupamento de X com função de coloração* $c : \{1, ..., N\} \to \{1, ..., k\}$ *e representantes* $Z = \{z_1, ..., z_k\}$ *para o qual a qualidade J(c,Z) seja a menor possível.*

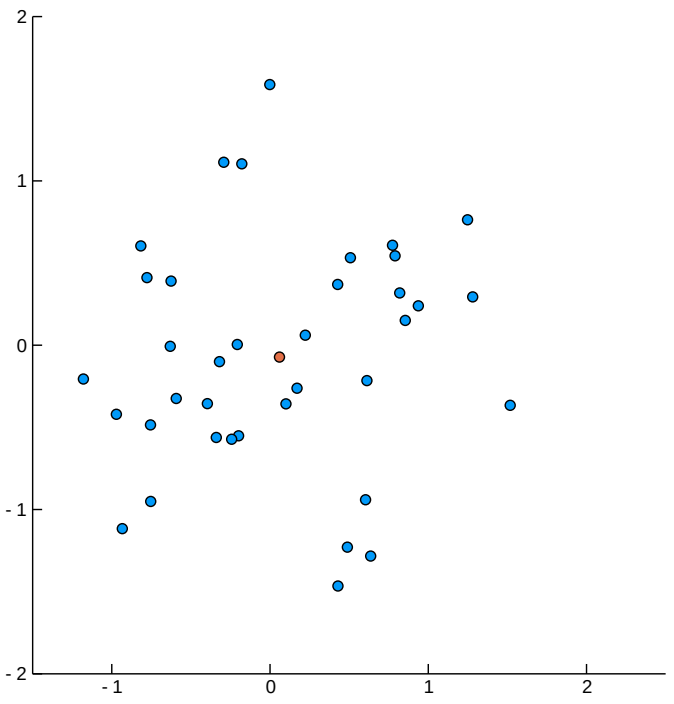
A ideia fundamental

Lema do Centroide

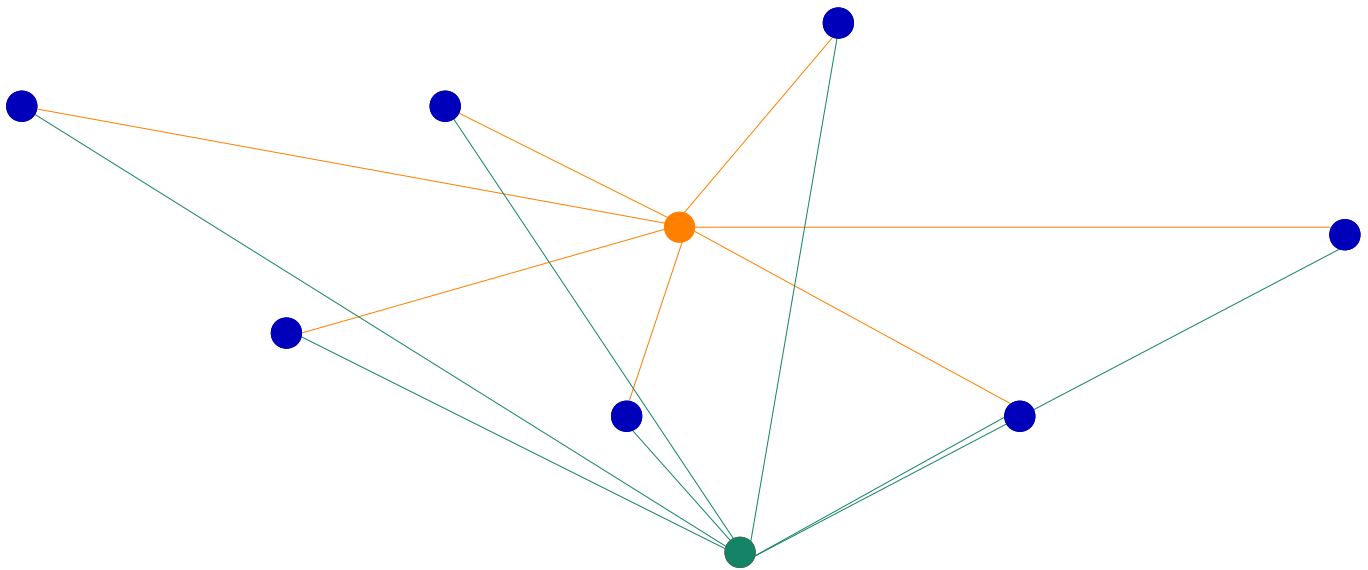
- Sejam $Y = \{y_1, ..., y_N\} \subset \mathbb{R}^d$ e $u = \frac{1}{n}\sum_{j=1}^{n} y_j$ o centroide (ou a media) do conjunto Y. Seja $z \in \mathbb{R}^d$ qualquer. Considere as quantidades

$$J(u) = \sum_{j=1}^{N} \|y_j - u\|^2 \quad \text{e} \quad J(z) = \sum_{j=1}^{N} \|y_j - z\|^2.$$

Se $z \neq u$ então $J(u) < J(z)$



“Dada uma nuvem de pontos em R^d o centroide da nuvem é o ponto mais próximo dela, dentre todos os outros pontos de R^d”



Demonstração. Vamos usar a *lei dos cossenos em espaços de Hilbert:*[5]

$$\|a+b\|^2 = \|a\|^2 + 2\langle a, b\rangle + \|b\|^2 \tag{6.1}$$

Seja $j \in \{1, ..., N\}$ qualquer. A equação (6.1) com $a = y_j - u$ e $b = -(z-u)$ nos leva a

$$\|y_j - z\|^2 = \|y_j - u - (z-u)\|^2 = \|y_j - u\|^2 - 2\langle y_j - u, z-u\rangle + \|z-u\|^2.$$

Logo,

$$\begin{aligned} J(z) = \sum_{j=1}^{N} \|y_j - z\|^2 &= \sum_{j=1}^{N} \left(\|y_j - u\|^2 - 2\langle y_j - u, z-u\rangle\right) + N\|z-u\| \\ &= \sum_{j=1}^{N} \|y_j - u\| - 2\sum_{j=1}^{N}\langle y_j - u, z-u\rangle + N\|z-u\|. \end{aligned}$$

Pela linearidade do produto interno,

$$\sum_{j=1}^{N}\langle y_j - u, z-u\rangle = \langle\sum_{j=1}^{N}(y_j - u), z-u\rangle.$$

O leitor questionador deve ter se perguntado: de onde raios saiu essa escolha marota de $a = y_j - u$ e $b = -(z - u)$. Pois aqui vem a resposta: a definição de u implica que

$$\sum_{j=1}^{N}(y_j - u) = \sum_{j=1}^{N} y_j - Nu = 0.$$

Concluímos assim que

$$J(z) = \sum_{j=1}^{N} \|y_j - u\| + N\|z - u\| = J(u) + N\|z - u\|.$$

Portanto, se $z \neq u$ então $\|z - u\| > 0$ e segue que $J(z) > J(u)$. □

O algorítimo *k-means*

- Dada uma lista de vetores $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ e uma lista inicial $\{z_1, \ldots, z_k\} \subset \mathbb{R}^d$ de k representantes, repita os passos a seguir até a convergência.

  Passo 1 Escolha uma função de coloração com a melhor qualidade possível.

  Passo 2 Redefina os representantes como centroides dos grupos.

Aplicações

- Algorítimo de recomendação (**Spotify, Netflix...)**
- Completar as entradas não-conhecidas de alguns vetores dentro de um conjunto grande de vetores (***mising data problem***)
- Classificação automática de textos (***automatic topic discovery***)