

$k$ -Médias

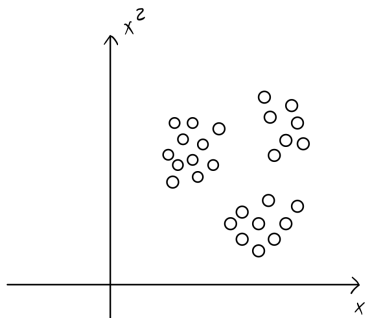
---

Advanced Institute for Artificial Intelligence – AI2

<https://advancedinstitute.ai>

# Introdução

Existem problemas para os quais não temos acesso aos rótulos das classes, ou seja, temos um problema de **aprendizado não supervisionada** (agrupamento). Nestas situações, não observamos os rótulos das amostras, tendo apenas um conjunto de dados definido da seguinte forma:  $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ , em que  $\mathbf{x}_i \in \mathbb{R}^n$ .



Uma das abordagens mais conhecidas é a chamada de  $k$ -Médias, do inglês  $k$ -Means, a qual objetiva agrupar dados com base nas distâncias entre as amostras. Geralmente, a distância Euclidiana é uma das mais utilizadas. O objetivo da técnica é particionar as amostras em  $k < m$  grupos.

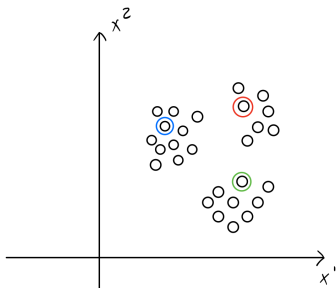
Seja  $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_k\}$  o conjunto de grupos em que  $\boldsymbol{\mu}_i \in \mathbb{R}^n$  corresponde ao centroide (ponto médio) do grupo  $\mathcal{S}_i$ . O algoritmo do  $k$ -Médias tenta resolver a seguinte equação:

$$\mathcal{S}^* = \arg \min_{\mathcal{S}} \sum_{i=1}^k \sum_{\mathbf{x} \in \mathcal{S}_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2. \quad (1)$$

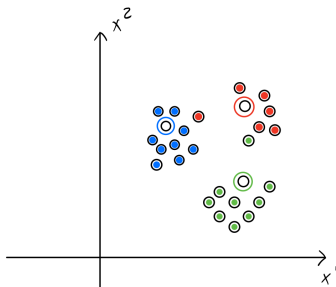
Note que desejamos os centros dos grupos que estão mais próximos dos dados, ou seja, queremos criar agrupamentos de forma a **minimizar** o espalhamento dos dados.

No entanto, a solução para a Equação 1 é um problema NP-Difícil para um número arbitrário de  $k$ , ou seja, não conhecemos um algoritmo polinomial que consegue resolvê-lo. O algoritmo é simples e envolve os seguintes passos:

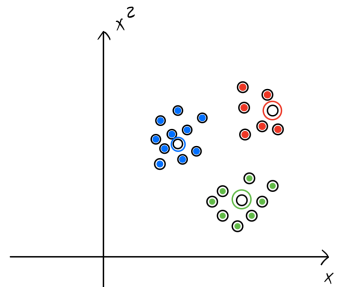
- ➊ Dado um número  $k$ , inicialize os centroides  $\mu_i$  de maneira aleatória,  $i = 1, 2, \dots, k$ . Inicialize, também, os grupos  $\mathcal{S}_i \leftarrow \{\}$ ,  $i = 1, 2, \dots, k$ .
- ➋ Associe cada amostra  $x \in \mathcal{X}$  ao seu centroide  $\mu_j$  mais próximo e faça  $\mathcal{S}_j \leftarrow \mathcal{S}_j \cup \{x\}$ ,  $j = 1, 2, \dots, k$ .
- ➌ Calcule o novo centroide  $\mu_j$  de cada grupo  $\mathcal{S}_j$ .
- ➍ Faça  $\mathcal{S}_j \leftarrow \{\}$ ,  $j = 1, 2, \dots, k$ .
- ➎ Execute os passos 2 – 4 até a convergência, ou seja, até os centroides não se moverem mais.



Inicialização  
dos centros



Atribuição de centros  
mais próximos



Configuração  
final

Alguns pontos importantes:

- Qual função de distância utilizar?
- Qual o valor de  $k$ ?
- $k$ -Médias assume que os agrupamentos são **circulamente simétricos** quando utilizamos a distância Euclidiana.