Clustering

Introdução ao algoritmo K-Means

Pontifícia Universidade Católica de Campinas

Prof. Dr. Denis M. L. Martins

Clustering ou Agrupamento

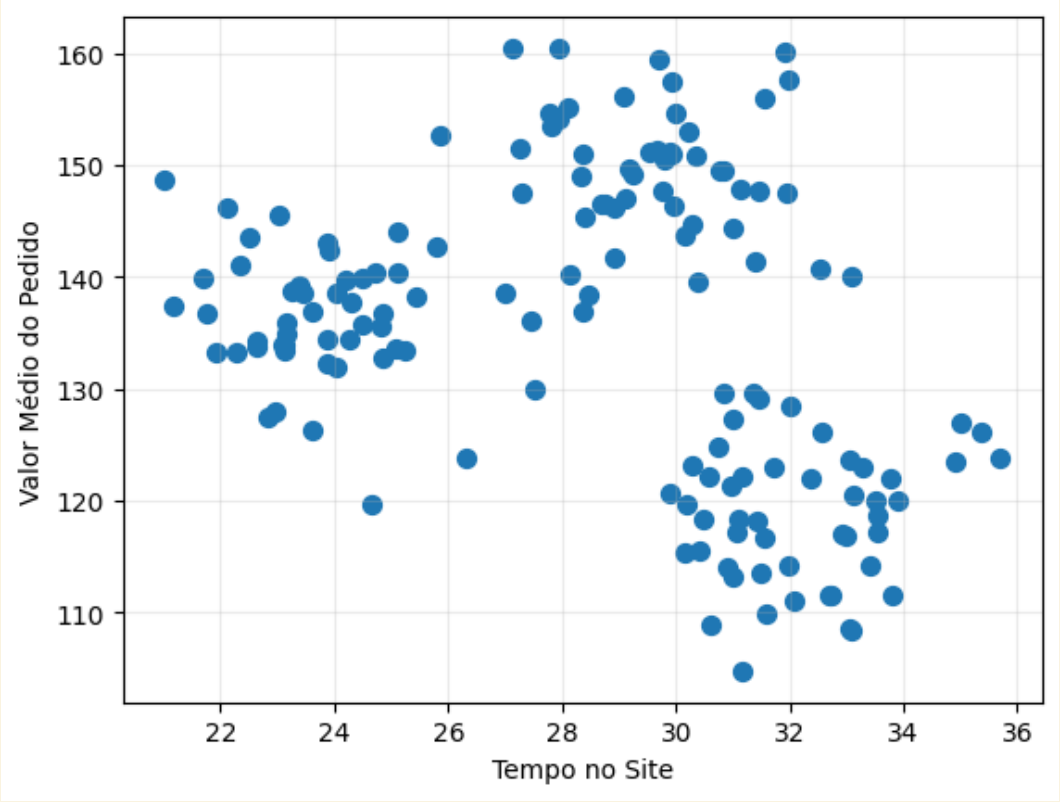
- Categoria de técnicas de aprendizado não supervisionado
- Permite descobrir estruturas ocultas em dados, mesmo sem conhecer a resposta correta previamente
  - Sem rótulos: *unlabeled data*
  - Rótulos podem ser caros de coletar. Por exemplo: exames clínicos, opiniões de especialistas, etc.
- O objetivo do **agrupamento**: éencontrar um grupos nos dados baseados em padrões:
  - Itens no mesmo grupo sejam mais semelhantes entre si do que aos itens de grupos diferentes

Exemplo

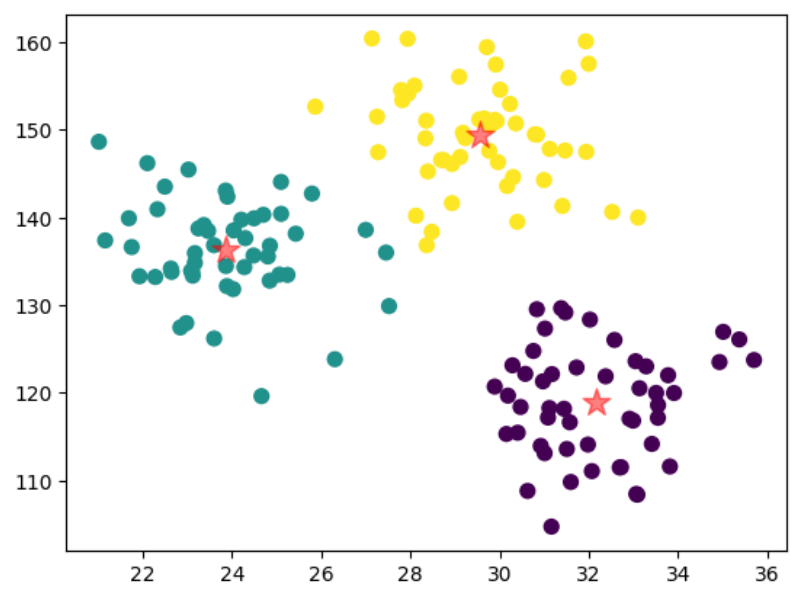
- **Problema**: Uma loja online que vende produtos artesanais deseja entender melhor seus clientes
- **Objetivo**: Personalizar ofertas e melhorar a experiência de compra.
- **Dados**: "Tempo no Site" (em minutos) e "Valor Médio do Pedido" (em reais).
- **Tarefa**: Segmentar os clientes em grupos distintos para direcionar campanhas de marketing mais eficazes.

Exemplo

Com um pouco de atenção, é possível ver três grupos nos dados coletados.



Clustering com K-Means

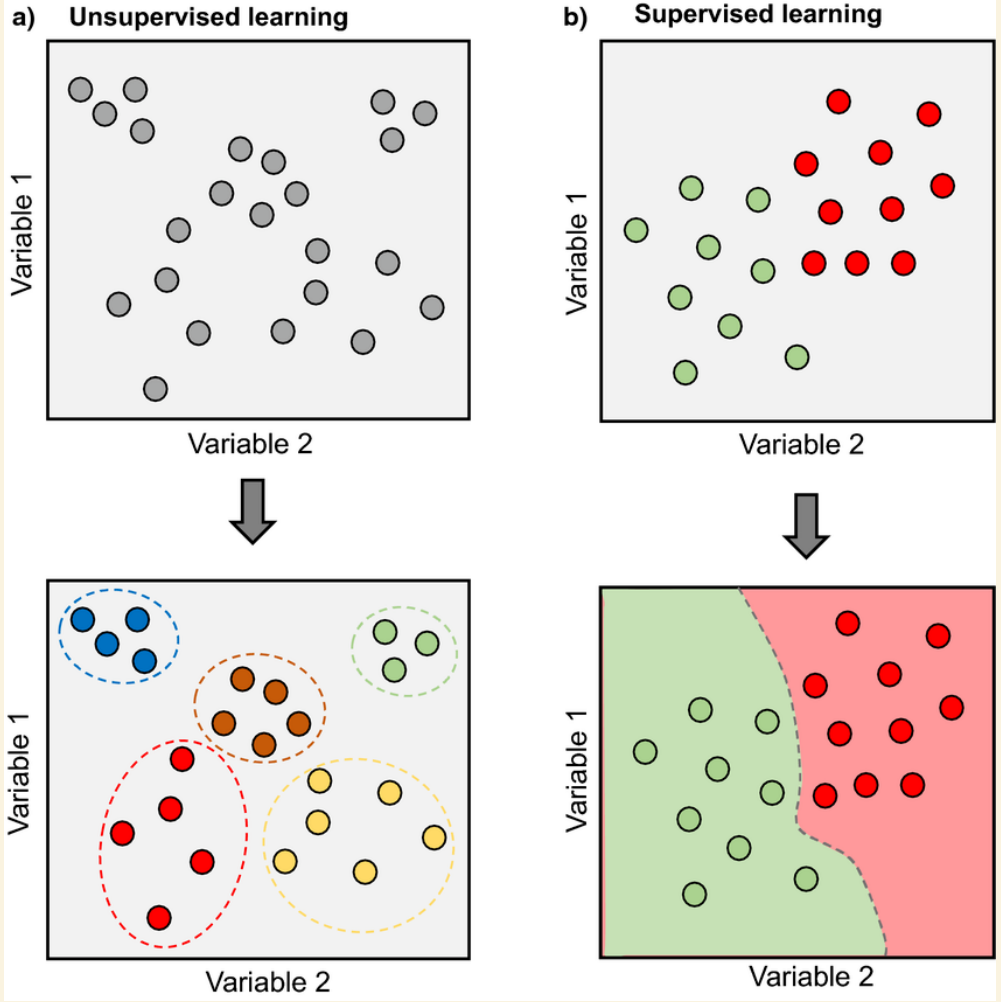


- **Compradores Eficientes:** Baixo tempo no site e valor médio do pedido moderado. Esses clientes são valiosos para a loja e devem receber ofertas personalizadas e programas de fidelidade.
- **Compradores Interessados:** Alto tempo no site e alto valor médio do pedido. Esses clientes podem ser atraídos a retornar ao site com conteúdo relevante e promoções direcionadas.
- **Navegadores:** Tempo médio no site e baixo valor médio do pedido. Esses clientes podem precisar de incentivos adicionais para aumentar seus gastos, como descontos ou frete grátis.

Exemplos de Cénarios de Clustering

- **Segmentação de Clientes:** No marketing, o agrupamento pode ser usado para segmentar clientes com base em suas características demográficas, hábitos de compra e outras características.
- **Segmentação de Imagens:** Na visão computacional, o agrupamento pode ser usado para segmentar uma imagem em diferentes regiões com base em cor, textura ou outras características.
- **Agrupamento de Documentos:** Na recuperação de informação, o agrupamento pode ser usado para agrupar documentos semelhantes, facilitando a busca e recuperação de informações relevantes.
- **Detecção de Anomalias:** Na segurança de redes, o agrupamento pode ser usado para identificar padrões incomuns no tráfego de rede, que podem indicar um ataque ou uma falha do sistema.

Tipos de problema de aprendizado de máquina



K-Means

O algoritmo *k-means* busca encontrar um número predeterminado de grupos (clusters) dentro de um conjunto de dados multidimensionais não rotulados.

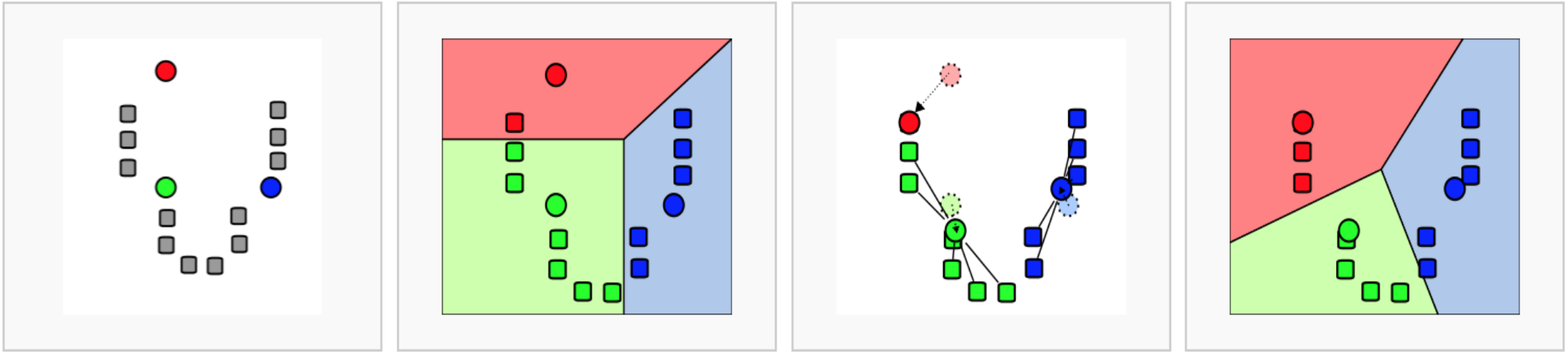
Ele faz isso usando uma concepção simples do que a agrupamento ideal deve ser:

- O "centroide" do grupo é a média aritmética de todos os pontos pertencentes ao grupo.
- Cada ponto está mais próximo do seu próprio centroide do que dos outros centroides.

Essas duas premissas são a base do modelo *k-means*.

K-Means

Demonstration of the standard algorithm



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

Funcionamento do K-Means

1. Escolha aleatoriamente *k* centroides dos exemplos como centros iniciais dos grupos.
2. Atribua cada exemplo ao centroide mais próximo.
3. Mova os centroides para o centro de todos os exemplos atribuídos a eles.
4. Repita os passos 2 e 3 até que as atribuições se tornem estáveis (ou seja, nenhum item mude de grupo) ou um número máximo de iterações seja atingido.

Funcionamento do K-Means

Distância euclidiana ao quadrado entre dois pontos *x* e *y* em um espaço *m*-dimensional:

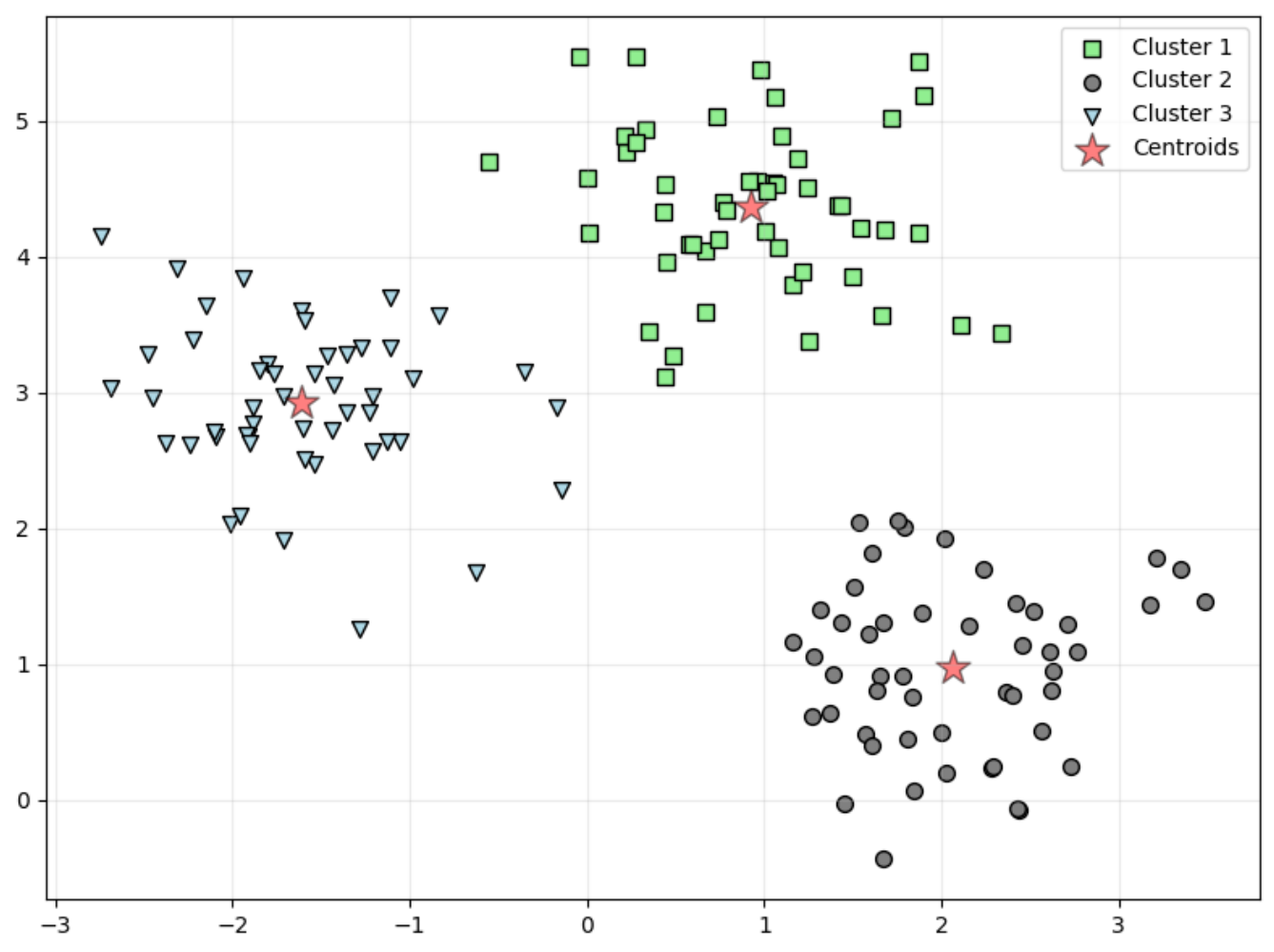
$$d(x, y)^2 = \sum_{j=1}^{m} (x_j - y_j)^2$$

Problema de otimização simples: Minimizar a **soma dos quadrados dos erros dentro do grupo** ("inércia do cluster"):

$$SSE = \sum_{i=1}^{m} \sum_{j=1}^{k} w^{(i,j)} d(x^{(i)}\ \mu^{(j)})^2,$$

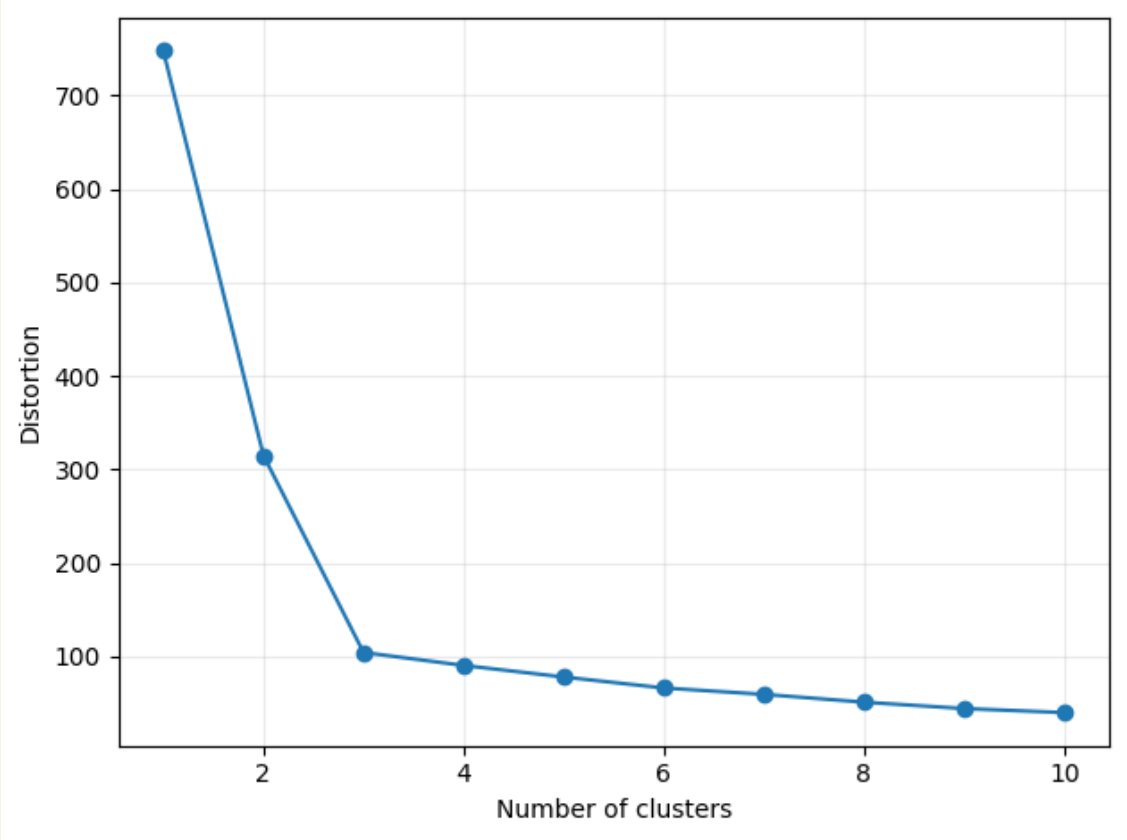
onde $\mu^{(j)}$ é o centroide para o grupo *j*; $w^{(i,j)} = 1$ se $x^{(i)} \in$ grupo *j* e 0 caso contrário.

K-Means



Elbow method para encontrar o número de clusters

Com base na SSE dentro do grupo, podemos usar o **método do cotovelo**, para estimar o número ótimo de *k* de clusters para uma determinada tarefa (*k* = 3 na imagem abaixo).



K-Means para Compressão de Cores

Em muitas imagens, uma grande quantidade dessas cores não são utilizadas e muitos pixels da imagem possuem cores similares ou até idênticas.

- Esta imagem na direita alcança um fator de compressão de cerca de 1 milhão!