

# Aprendizagem sem Supervisão

A **Aprendizagem sem Supervisão** é um paradigma de aprendizagem automática (*machine learning*) em que não são conhecidos os resultados sobre os casos, apenas os enunciados dos problemas, tornando necessário a escolha de técnicas de aprendizagem que avaliem o funcionamento interno do sistema.

Problemas de aprendizagem sem supervisão são, normalmente, divididos em duas categorias: - **Segmentação**: quando se pretende organizar os dados em grupos coerentes (por exemplo, agrupar clientes que compram bebidas açucaradas); - **Associação**: quando se pretende encontrar padrões de associação entre os dados (por exemplo, pessoas que compram bebidas açucaradas não compram bebidas alcóolicas).

## Segmentação (Clustering)

A Segmentação (Clustering) de dados é um processo através do qual se **particiona** um conjunto de **dados em segmentos/clusters** de menor dimensão, que agrupam conjuntos de dados semelhantes.

### Exemplo:

Dado um conjunto de dados de clientes de um supermercado, é possível segmentar os clientes em grupos de acordo com as suas preferências de compra.

**Aplicações de Clustering**: - Como fase de pre-processamento, de forma a organizar os dados a submeter a outros algoritmos; - Como uma ferramenta de análise dos dados; - Problemas de conhecimento de padrões; - No processamento de imagens;

## Algoritmos de Clustering

Os algoritmos de Clustering são divididos em duas categorias: - **Hierárquicos**: Criam uma árvore de clusters, onde cada nó é um cluster que contém os clusters filhos; - **Particionais**: Dividem o conjunto de dados em  $k$  clusters, onde  $k$  é um valor pré-definido.

**K-Means** O algoritmo **K-Means** é um dos algoritmos de clustering mais populares. Ele é um algoritmo de particionamento que divide os dados em  $k$  clusters, onde  $k$  é um valor pré-definido.

O algoritmo K-Means é composto por três passos: 1. **Inicialização**: Selecionar  $k$  centroides aleatórios; 2. **Atribuição**: Atribuir cada ponto ao centroide mais próximo; 3. **Atualização**: Recalcular os centroides.

**AGNES (Agglomerative Nesting)** O algoritmo **AGNES** é um algoritmo de clustering hierárquico que cria uma árvore de clusters. Ele é composto por três passos: 1. **Inicialização**: Cada ponto é um cluster; 2. **Agregação**: Agregar os clusters mais próximos; 3. **Repetição**: Repetir o passo 2 até que todos os pontos estejam num único cluster.

As vantagens deste método são que ele é fácil de implementar e que não é necessário definir o número de clusters *a priori*.

**DIANA (Divisive Analysis)** O algoritmo **DIANA** é um algoritmo de clustering hierárquico que cria uma árvore de clusters. Ele é composto por três passos: 1. **Inicialização**: Todos os pontos estão num único cluster; 2. **Divisão**: Dividir o cluster em dois clusters; 3. **Repetição**: Repetir o passo 2 até que cada ponto esteja num cluster.

## Referências

- K-Means Clustering Algorithm