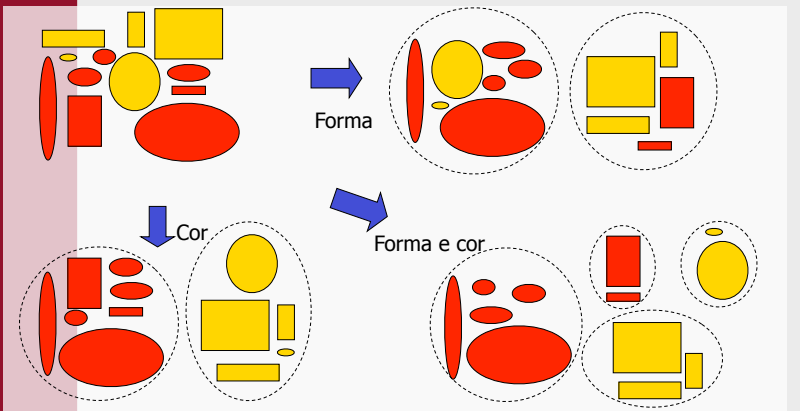
# Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina

## Análise de agrupamentos

Prof. Tiago A. Almeida



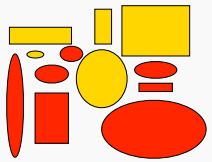
## Análise de Agrupamentos






## Análise de Agrupamentos

- Objetivo de técnica de agrupamento: encontrar uma estrutura de clusters (grupos) nos dados
  - Objetos em cada cluster compartilham alguma característica ou propriedade
    - São de alguma maneira similares



Como organizar?



## Aplicações

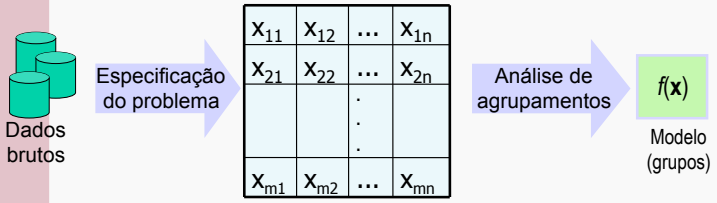
- Mercado
  - Segmentação de clientes para direcionar propagandas
- Redes sociais
- Distribuição do processamento de dados em *Data Centers*
- Etc...

# Análise de Agrupamentos

- Considere os objetos como pontos em um espaço de dimensão $n$
  - $n$ = número de atributos
- Cluster: coleção de objetos próximos ou que satisfazem alguma relação espacial



# Análise de Agrupamentos






# Clusters

- Algumas definições comuns:

**Cluster bem separado**
Conjunto de pontos tal que qualquer ponto está mais próximo (é mais similar) a cada outro nesse cluster do que a qualquer outro ponto não pertencente a ele

**Cluster baseado em centro**
Conjunto de pontos tal que qualquer ponto está mais próximo (é mais similar) ao centro desse cluster do que ao centro de qualquer outro cluster

*Centroide: média aritmética dos pontos do cluster*
*Medoide: ponto mais representativo do cluster*



# Clusters

- Algumas definições comuns:

**Cluster contínuo ou encadeado**
Conjunto de pontos tal que qualquer ponto está mais próximo (é mais similar) a um ou mais pontos nesse cluster do que a qualquer outro ponto não pertencente a ele

**Cluster baseado em densidade**
Região densa de pontos, separada por outras regiões de alta densidade por regiões de baixa densidade

**Cluster baseado em similaridade**
Conjunto de pontos que são similares, enquanto pontos em clusters diferentes não são similares

# Clusters

- Cada definição resulta em um critério de agrupamento
  - Forma de selecionar uma estrutura de clusters (modelo) que melhor se ajuste aos dados
- Cada algoritmo de agrupamento:
  - Baseado em um critério de agrupamento
  - Usa uma medida de proximidade
  - Usa um método de busca para encontrar uma estrutura
    - De acordo com o critério de agrupamento adotado



# Critérios de agrupamento

- Categorias:

Compactação ou homogeneidade
Associada a variação intracluster pequena
Efetivos na descoberta de clusters esféricos e/ou bem separados
Podem falhar para estruturas mais complexas

Encadeamento ou ligação
Conceito mais local (objetos vizinhos devem compartilhar o mesmo cluster)
Apropriado para detectar clusters de formas arbitrárias
Não robusto para quando há pouca separação espacial entre os clusters



# Critérios de agrupamento

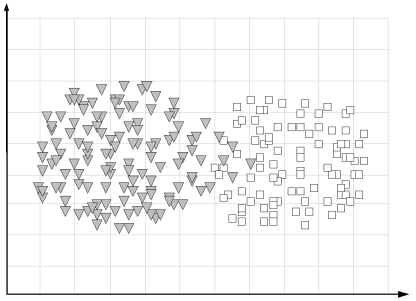
- Categorias:

Separação espacial
Considera distâncias entre os clusters
Fornece pouca orientação durante o agrupamento, podendo levar a soluções triviais
É comumente empregado em conjunto com outros



# Critérios de agrupamento

- Exemplo:



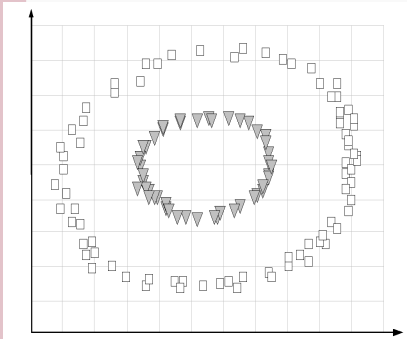
Conjunto de dados `globular`:

Dois clusters esféricos **bem separados**

Algoritmos baseados em **compactação** conseguem captar essa estrutura

# Critérios de agrupamento

- Exemplo:



Conjunto de dados `anel`:

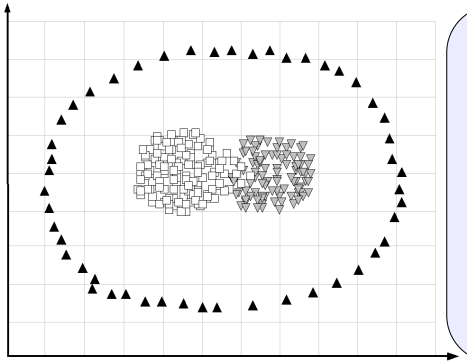
Dois clusters **bem distintos** na forma de anel

Algoritmos baseados em **encadeamento** conseguem captar essa estrutura



# Critérios de agrupamento

- Exemplo: e quando conjunto é heterogêneo?



Conjunto de dados com clusters em conformidade com critérios de agrupamento diferentes

Um cluster em anel
Dois clusters globulares

*Não existe um único algoritmo de agrupamento capaz de encontrar todos os tipos de agrupamentos*



# Algoritmos de agrupamento

- Existe um grande número
  - Cada um buscando clusters de acordo com um critério diferente
    - ⇒ critério de agrupamento representa principal aspecto de um algoritmo de agrupamento
  - Exemplos:
    - Algoritmo k-médias: procura clusters compactos
    - Algoritmo hierárquico ligação média: otimizam critério baseado em encadeamento

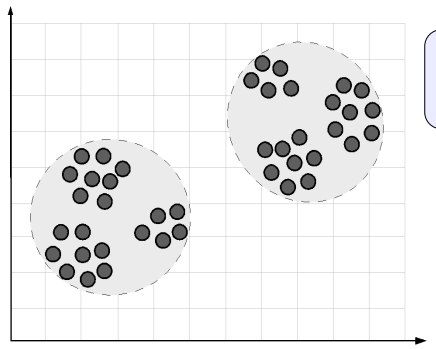


# Algoritmos de agrupamento

- Outro aspecto: nível de refinamento
  - Algoritmos podem encontrar estruturas em diferentes níveis de refinamento
    - Números de clusters diferentes ou de densidades diferentes
    - Dependendo de suas configurações de parâmetros
      - ⇒ Importância de ajuste de parâmetros

# Nível de refinamento

- Exemplo:

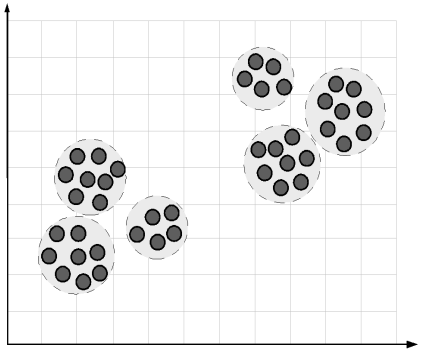


Estrutura **compacta** com dois clusters



# Nível de refinamento

- Exemplo:



Estrutura **compacta** com seis clusters



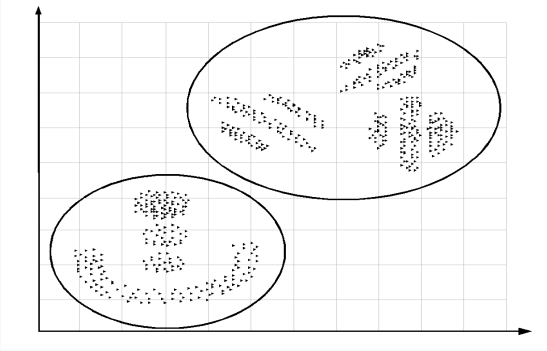
# Estruturas

- Um mesmo conjunto de dados pode ter mais de uma estrutura relevante
  - Cada uma representando uma diferente interpretação dos dados
  - Cada estrutura pode ser compatível com um critério de agrupamento diferente, estar em um nível diferente e/ou ser heterogênea



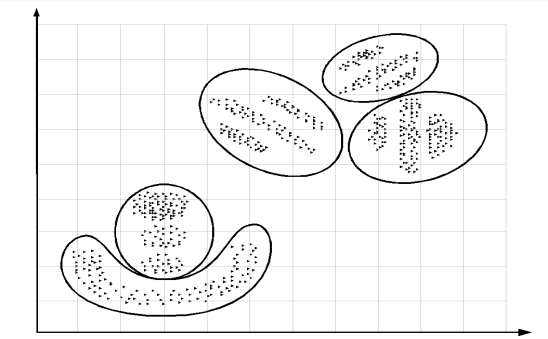
# Estruturas

- Exemplo:



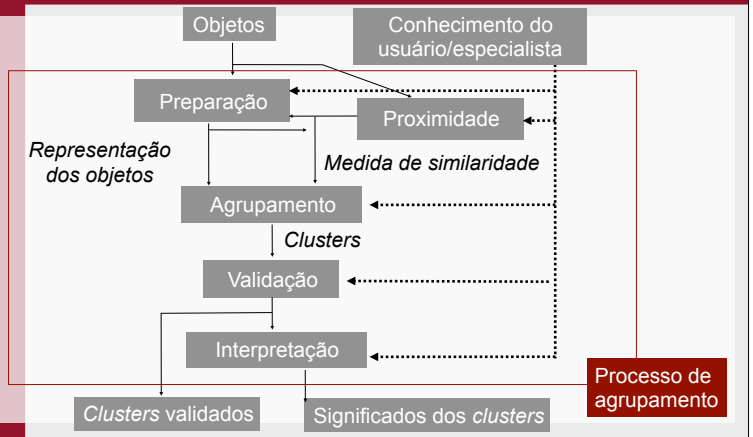
# Estruturas

- Exemplo:





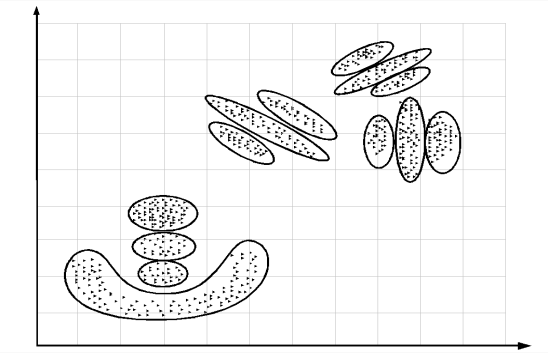
# Etapas Análise de Agrupamentos






# Estruturas

- Exemplo:





# Preparação dos Dados

- Representação e pré-processamento
  - Representação:
    - Geralmente atributo valor
    - Ou relação de proximidade entre objetos (matrizes e grafos de similaridade/dissimilaridade)
  - Pré-processamentos podem incluir:
    - Normalizações
    - Conversões de tipos
    - Redução de atributos
    - Extração de características

## Agrupamento

- Etapa central
  - Um ou mais algoritmos de agrupamento são aplicados aos dados



## Interpretação

- Processo de examinar os clusters e rotulá-los
  - Descrevendo a natureza de cada um
  - Também é forma de validação dos clusters



## Validação

- Avalia o resultado do agrupamento
  - Determinar se clusters são significativos
  - Também pode ajudar na definição de parâmetros do algoritmo

# Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina

## k-médias

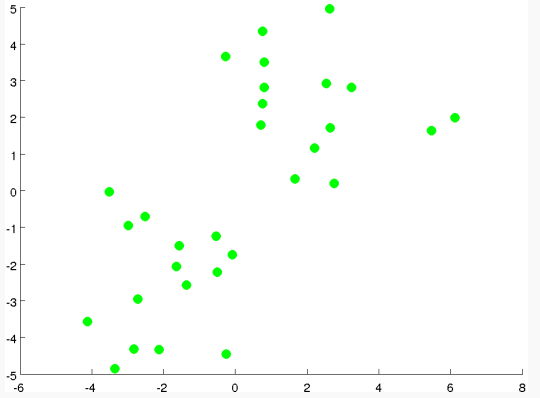
Prof. Tiago A. Almeida

# k-Médias

- Realiza agrupamento particional
  - O número de grupos ($k$) tem que ser definido a *priori*



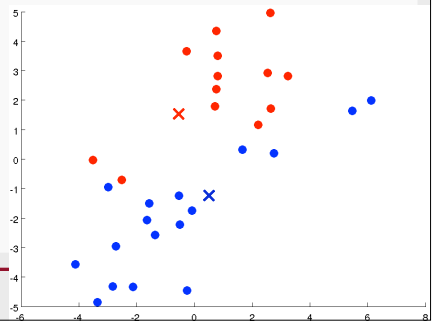
# k-Médias: Objetos





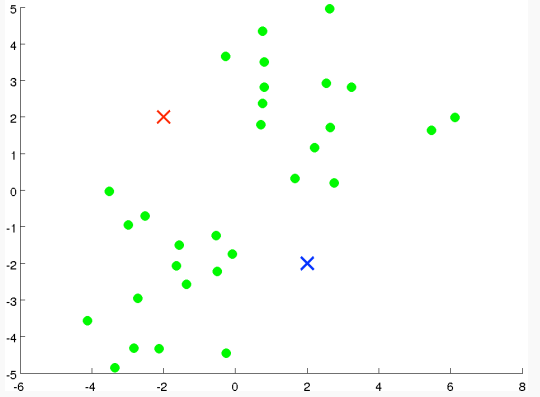
# k-Médias: Algoritmo

- **Passo 1**: Inicializar ***k* centróides** aleatoriamente
- **Passo 2**: Atribuir cada objeto ao grupo associado ao centro mais próximo (atribuição de centróide)
- **Passo 3**: Computar novo centro para cada grupo (mover centróide)
- **Passo 4**: Repetir **Passo 2** (com os novos centros) e **Passo 3** até que não haja mudança nos centros

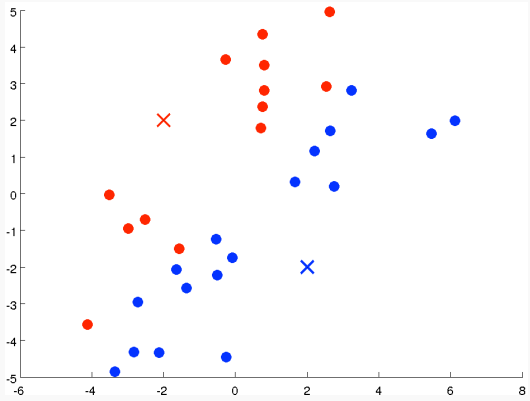




# Passo 1: Inicialização (k = 2)

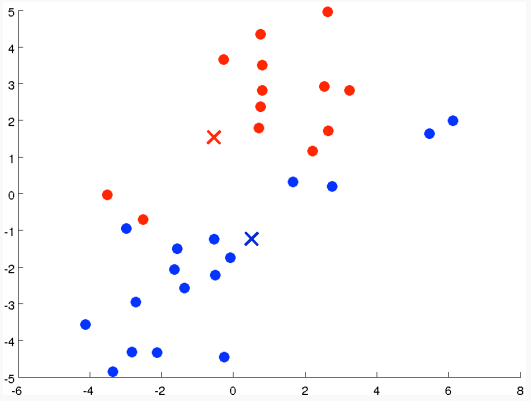


## Passo 2: Atribuição aos centros



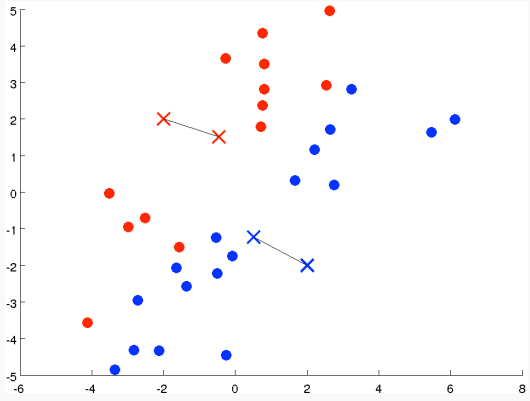



## Passo 2: Atribuição



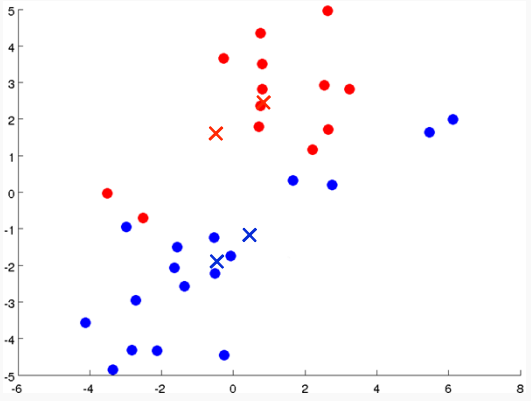



## Passo 3: Computar novos centros

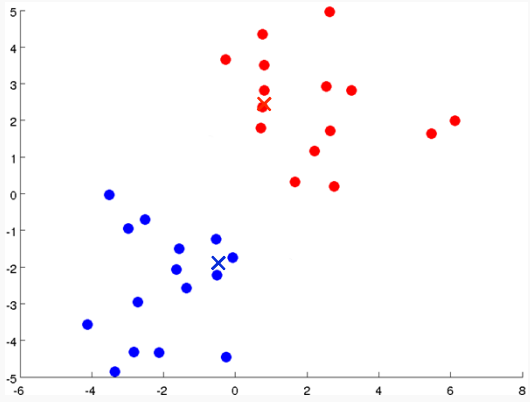





## Passo 3: Computar novos centros

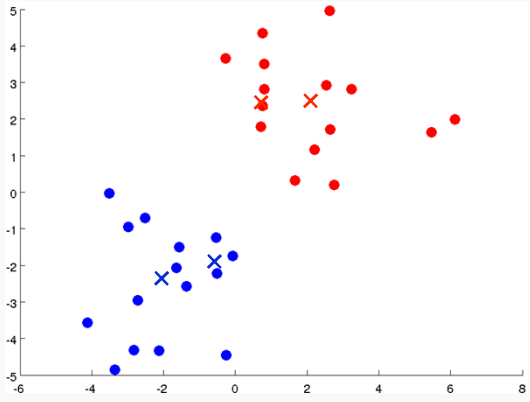


## Passo 2: Atribuição



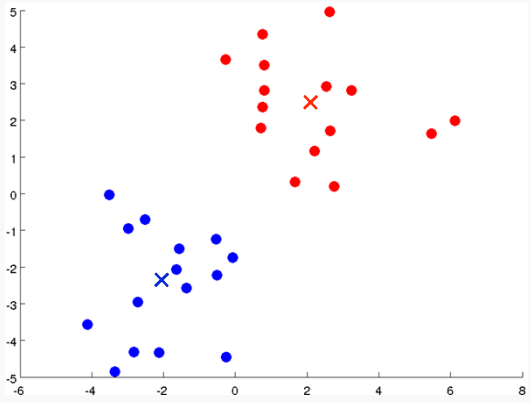


## Passo 3: Computar novos centros





## Passo 2: Atribuição





## k-Médias: algoritmo

- Entradas
  - *Quantidade de clusters: K*
  - *Dados de treinamento:* $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

# k-Médias: algoritmo

**k-Médias**

Entrada: $K$, $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$
Saída: Conjunto de $K$ clusters

Inicializar aleatoriamente $K$ centróides $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$
**Repita**
  **Para** $i$ = 1 **até** $m$
    $c^{(i)}$ := índice (1 a $K$) da centróide mais próxima de $x^{(i)}$

  **Para** $k$ = 1 **até** $K$
    $\mu^{(i)}$ := média dos pontos associados ao cluster $k$
**Fim-repita**



# k-Médias: Função Custo

- $c^{(i)}$: índice do cluster (1, ..., $K$) ao qual a amostra $x^{(i)}$ está associada
- $\mu^{(k)}$: centróide $k$ ($\mu_k \in \mathbb{R}^n$)
- $\mu_{c^{(i)}}$: centróide do cluster no qual a amostra $x^{(i)}$ está associada
- **Função Custo:**

$$J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K) = \frac{1}{m} \sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2$$

$$\min_{\substack{c^{(1)},\ldots,c^{(m)},\\ \mu_1,\ldots,\mu_K}} J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$$



# k-Médias: Função Custo


- $c^{(i)}$: índice do cluster (1, ..., $K$) ao qual a amostra $x^{(i)}$ está associada
- $\mu^{(k)}$: centróide $k$ ($\mu_k \in \mathbb{R}^n$)
- $\mu_{c^{(i)}}$: centróide do cluster no qual a amostra $x^{(i)}$ está associada




# k-Médias: inicialização das centróides

**k-Médias**

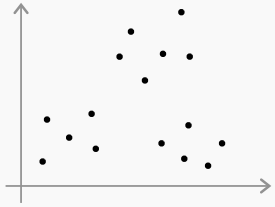
Entrada: $K$, $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$
Saída: Conjunto de $K$ clusters

Inicializar aleatoriamente $K$ centróides $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$
**Repita**
  **Para** $i$ = 1 **até** $m$
    $c^{(i)}$ := índice (1 a $K$) da centróide mais próxima de $x^{(i)}$

  **Para** $k$ = 1 **até** $K$
    $\mu^{(i)}$ := média dos pontos associados ao cluster $k$
**Fim-repita**

## k-Médias: inicialização das centróides

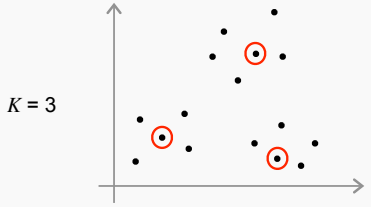
1. Definir $K < m$
2. Escolher aleatoriamente $K$ amostras da base
3. Definir $\mu_1, \ldots, \mu_K$ iguais às $K$ amostras selecionadas





## k-Médias: inicialização das centróides

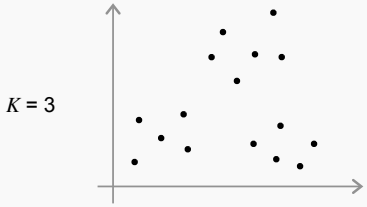
1. Definir $K < m$
2. **Escolher aleatoriamente $K$ amostras da base**
3. Definir $\mu_1, \ldots, \mu_K$ iguais às $K$ amostras selecionadas






## k-Médias: inicialização das centróides

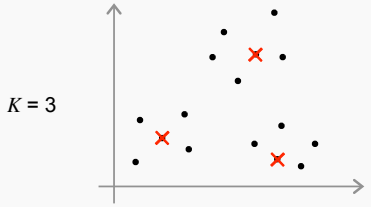
1. **Definir $K < m$**
2. Escolher aleatoriamente $K$ amostras da base
3. Definir $\mu_1, \ldots, \mu_K$ iguais às $K$ amostras selecionadas





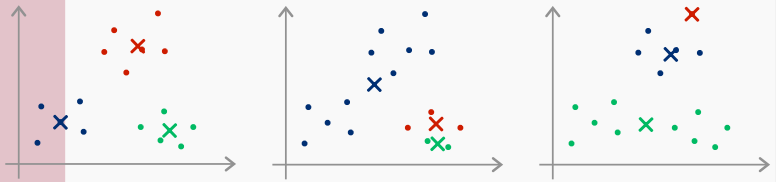

## k-Médias: inicialização das centróides

1. Definir $K < m$
2. Escolher aleatoriamente $K$ amostras da base
3. **Definir $\mu_1, \ldots, \mu_K$ iguais às $K$ amostras selecionadas**



## k-Médias: ótimos locais

- A inicialização aleatória pode impactar diretamente a qualidade final dos clusters
- O $k$-médias apresenta sensibilidade à condição inicial e pode ficar "preso" em ótimos locais





## k-Médias: evitando ótimos locais

**k-Médias de menor custo**

**Para** $i$ = 1 **até** 100

Inicializar aleatoriamente $K$ centróides $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$

Executar $k$-médias. Armazenar $c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K$

Calcular a função custo $J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$

**Fim-Para**

Selecionar resultado que apresentar menor custo $J$



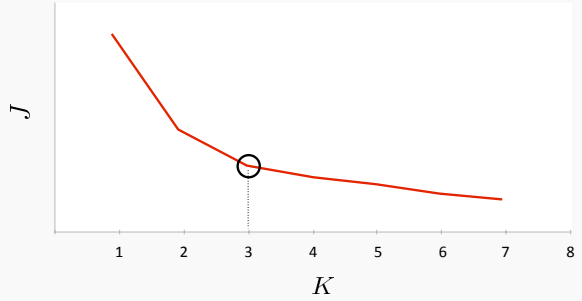
## k-Médias: número de clusters

- Número de cluster depende da aplicação
  - $k$ é geralmente escolhido por
    - Visualização dos dados
    - Conhecimento prévio / Necessidade do problema



## k-Médias: número de clusters

- Método cotovelo para selecionar o valor de $k$



## k-médias

- Aspectos positivos:
  - Eficiente
    - O($m$)
  - Como usa critério de compactação, é indicado para encontrar grupos hiperesféricos



## k-médias

- Aspectos negativos:
  - Pode convergir para ótimos locais
  - Sensível à inicialização
  - Clusters em geral desbalanceados
  - Difícil determinar o valor de $k$



## Avaliação do resultado

- Validação dos agrupamentos
  - Determinar se os clusters são significativos (se a solução é representativa para o conjunto de dados)
  - Um agrupamento é válido se não ocorreu por acaso, já que qualquer algoritmo de agrupamento encontrará clusters independentemente da existência de similaridade entre os dados
  - Existem várias medidas para avaliar agrupamentos
    - Considerando ou não algum agrupamento conhecido nos dados



## Referências


- Ilustrações usadas:
  - http://www.expertstown.com/web-mining/
- Alguns slides foram baseados em apresentações de:
  - Prof Dr André C. P. L. F. Carvalho
  - Prof Ricardo Campello
  - Profa Solange O. Rezende
  - Prof Dr Marcilio C. P. Souto