Aprendizado de Máquina

Sistemas de Informação Inteligente

Prof. Leandro C. Fernandes

Adaptado a partir do material de: Ricardo J. G. B. Campello, Eduardo R. Hruschka e André C. P. L. F. de Carvalho

Tópicos a serem abordados:

- Aprendizado de Máquina Supervisionado
  - Classificação
  - Algoritmo k–NN
- Aprendizado de Máquina Não Supervisionado
  - Agrupamento de Dados
  - Algoritmo das k-médias (k-Means)

Classificação

- Técnica supervisionada que classifica novas instâncias em uma ou mais classes conhecidas
  - Número definido de classes
  - Frequentemente apenas duas (classificação binária)

- Exemplos
  - Diagnóstico, Análise de crédito, ...

Classificação

- Existem várias técnicas, para diferentes contextos de aplicação
  - Sucesso de cada método depende do domínio de aplicação e do problema particular em mãos
  - Técnicas simples muitas vezes funcionam bem!

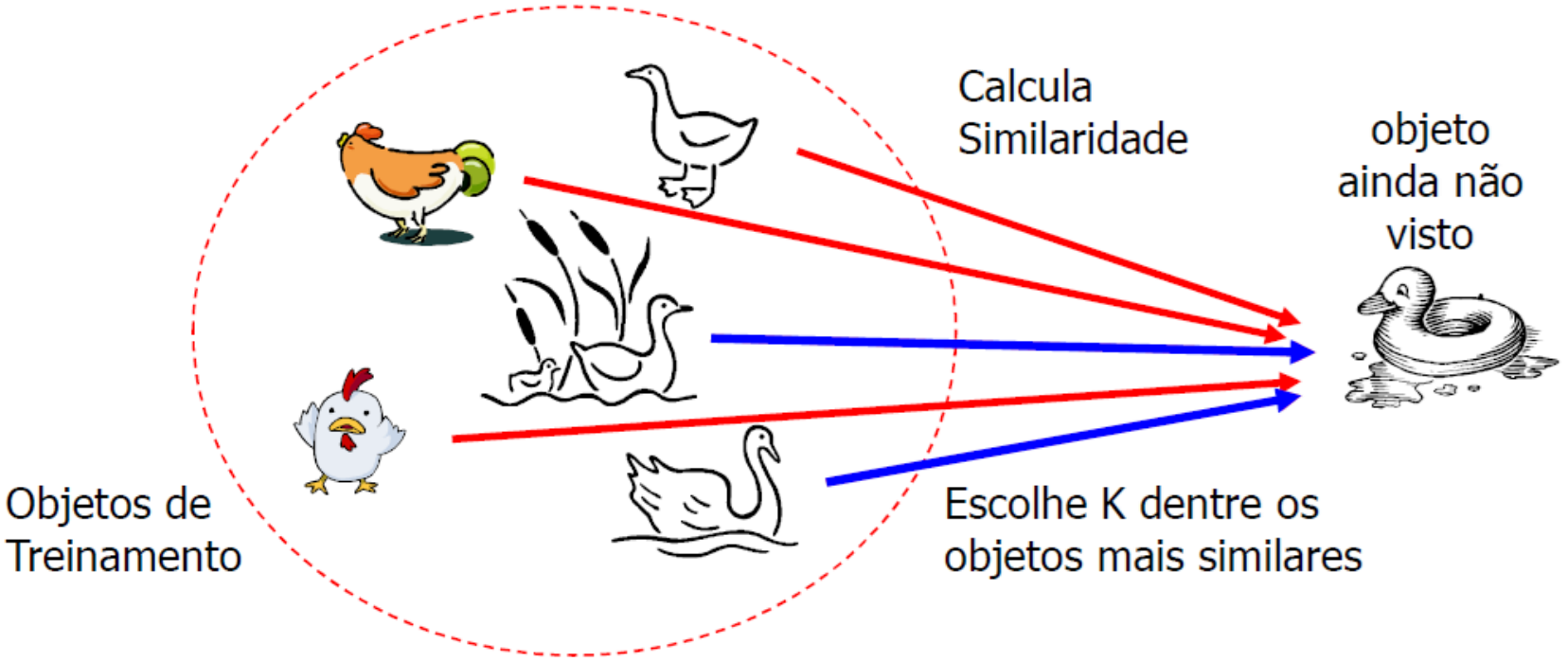
- Análise Exploratória de Dados!

k-Nearest-Neighbors (kNN)

- O Algoritmo k-Vizinhos-Mais-Próximos, ou kNN (*k-Nearest-Neighbors*) é um dos mais simples e bem difundidos algoritmos do paradigma baseado em instâncias.

kNN

- Ideia Básica:
  - Se anda como um pato, “quacks” como um pato, então provavelmente é um pato



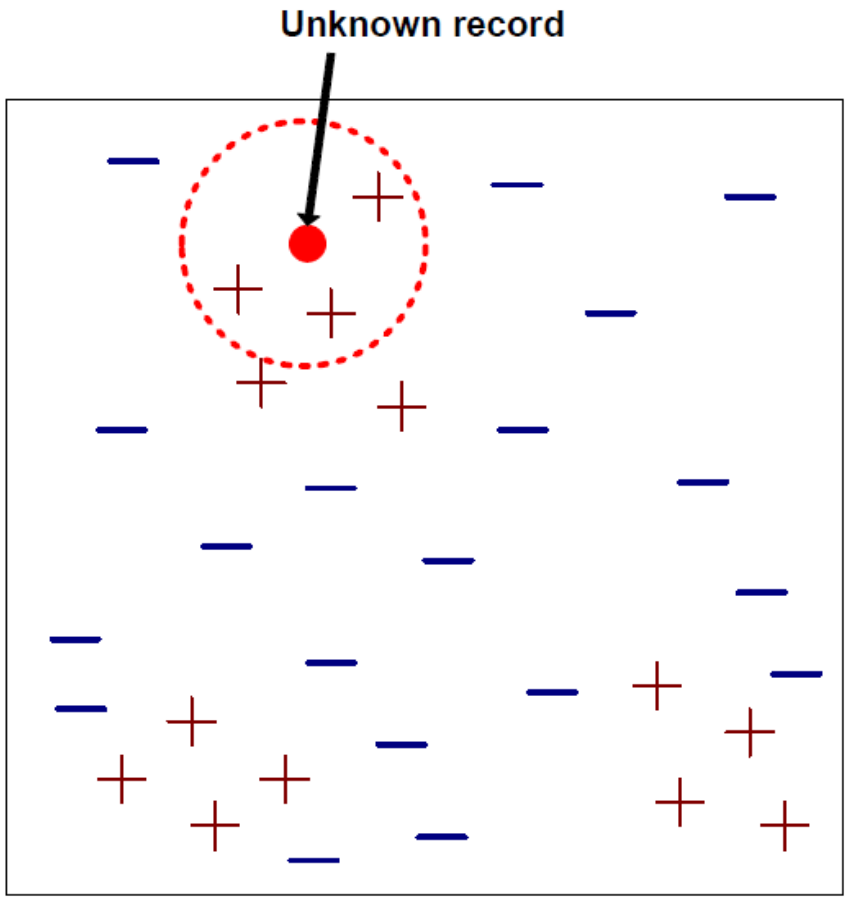
kNN

Em outras palavras, o kNN tem por objetivo:

- Consiste em classificar a instância t atribuindo a ela o rótulo mais frequentemente dentre as k amostras mais próximas.

  Uma medida de proximidade bastante utilizada é a distância Euclidiana: $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

kNN



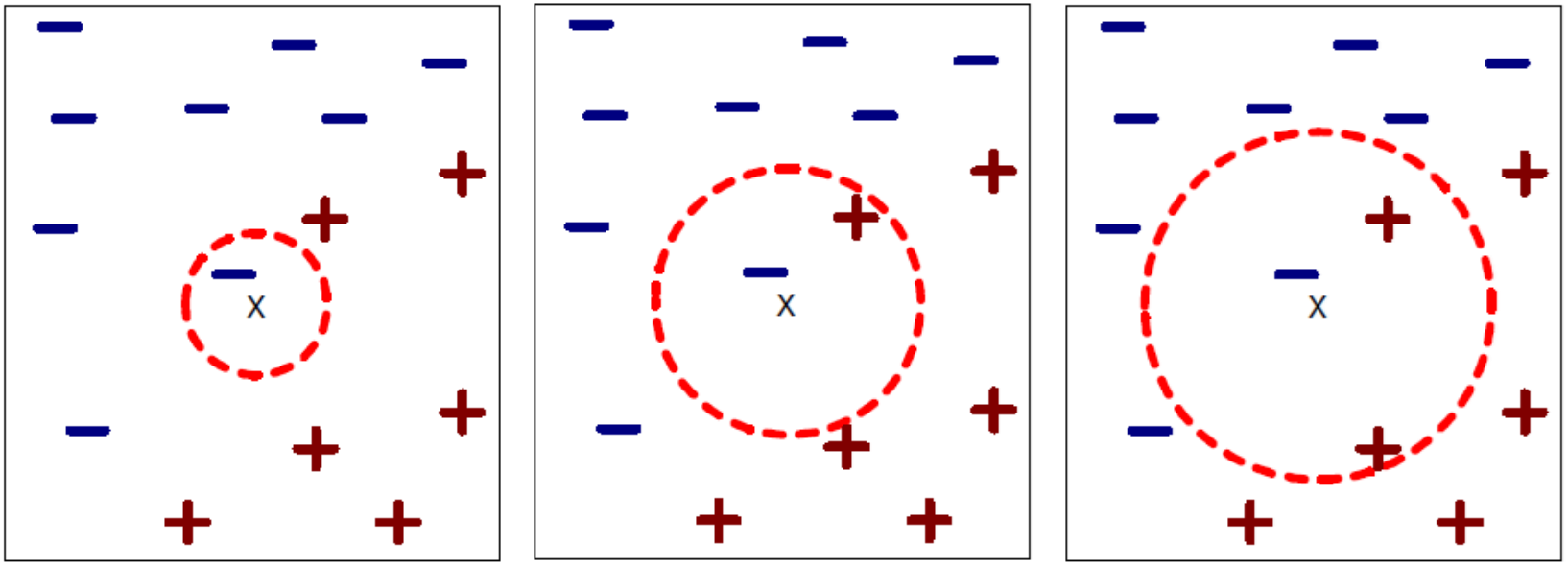
Requer 3 coisas:

- A base de dados de treinamento
- Uma medida de (dis)similaridade entre os objetos da base
- O valor de K: no. de vizinhos mais próximos a recuperar

Para classificar um objeto não visto:

- Calcule a (dis)similaridade para todos os objetos de treinamento
- Obtenha os K objetos da base mais similares (mais próximos)
- Classifique o objeto não visto na classe da maioria dos K vizinhos

kNN

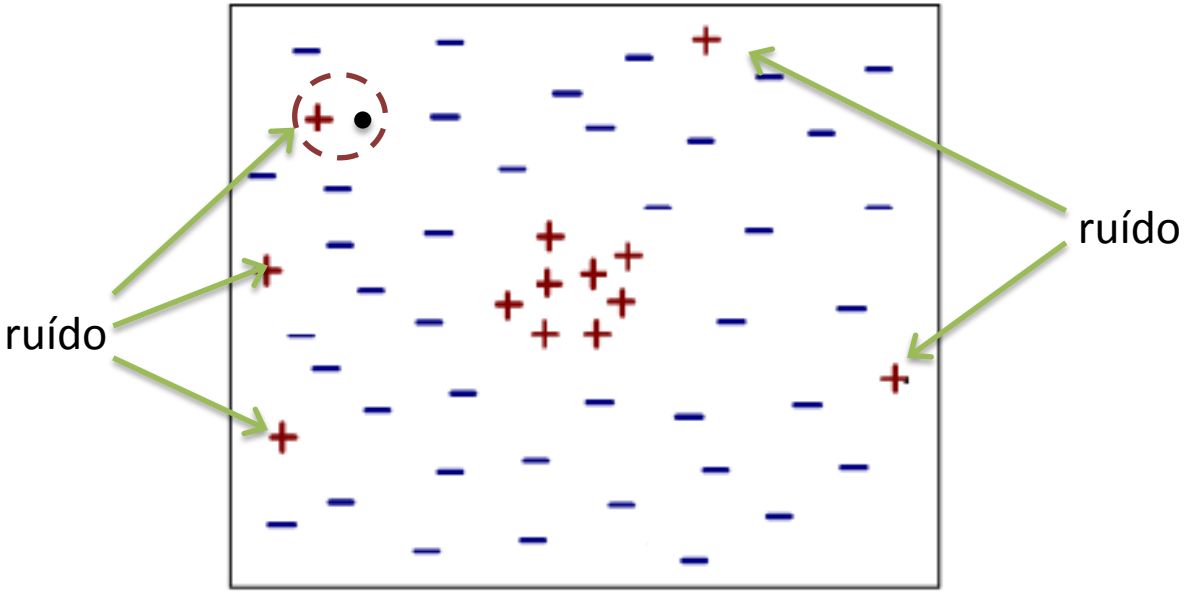


(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

k-NN: Visão geométrica para 2 atributos contínuos e dissimilaridade por distância Euclidiana. K = 1, 2 e 3

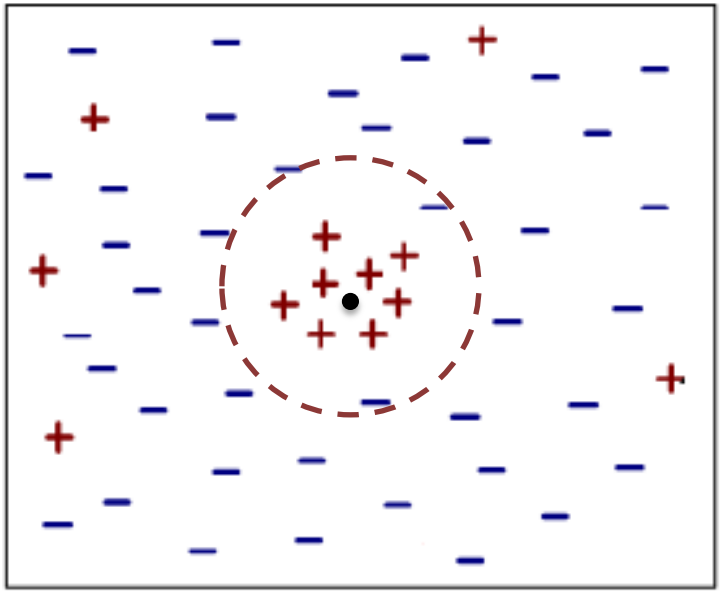
kNN: Escolha do Valor de K

- Muito pequeno:
  - discriminação entre classes muito flexível
  - porém, sensível a ruído
    - classificação pode ser instável (p. ex. K = 1 abaixo)



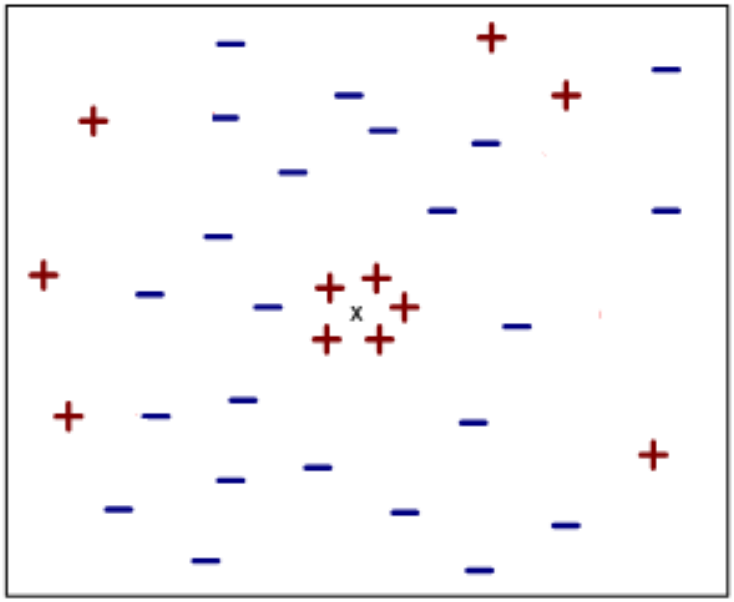
kNN: Escolha do Valor de K

- Muito grande:
  - mais robusto a ruído
  - menor flexibilidade de discriminação entre classes
    - privilegia classe majoritária...



kNN: Configuração

- Valor ideal?
  - Depende da aplicação
  - Análise Exploratória dos Dados

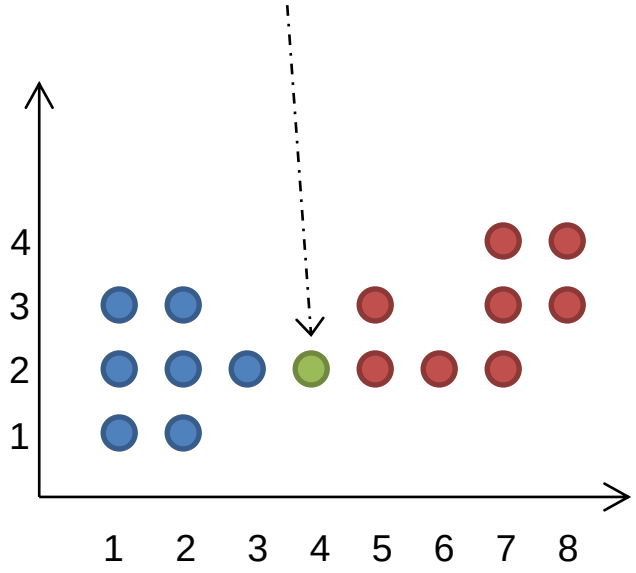


kNN

- Como calcular as (dis)similaridades... ?
  - Existem dezenas de medidas, sendo que aquela mais apropriada depende:
    - do(s) tipo(s) do(s) atributos!
    - do domínio de aplicação!
- Por exemplo:
  - Euclidiana, City Block, Mahalanobis, Casamento Simples (Simple Matching), Jaccard, Cosseno, Pearson, ...

kNN: Um Exemplo

A qual classe pertence este ponto?
Azul ou vermelho?



Calcule para os seguintes valores de k: 1, 3, 5 e 7

- k = 1: não se pode afirmar
- k = 3: vermelho
- k = 5: vermelho
- k = 7: azul

- A classificação pode mudar de acordo com a escolha de *k.*

kNN

- Além da escolha de uma medida apropriada, é preciso condicionar os dados de forma apropriada
  - Por exemplo, atributos podem precisar ser normalizados para evitar que alguns dominem completamente a medida de (dis)similaridade
- Exemplo:
  - Altura de uma pessoa adulta normal: 1.4m a 2.2m
  - Peso de uma pessoa adulta sadia: 50Kg a 150Kg
  - Salário de uma pessoa adulta: $400 a $30.000

Exercício

- Normalize os dados abaixo para em [0, 1] e utilizando o kNN com Distância Euclidiana, classifique a última instância para K = 1, 3 e 5. Discuta os resultados.

Febre	Enjôo	Mancha	Diagnóstico
0	1	3	doente
1	0	2	saudável
2	1	3	doente
2	0	0	saudável
0	0	4	doente
1	0	1	???

kNN Ponderado

- Na versão básica do algoritmo, a indicação da classe de cada vizinho possui o mesmo peso para o classificador
  - 1 voto (+1 ou -1) por vizinho mais próximo
- Isso torna o algoritmo muito sensível à escolha de K
- Uma forma de reduzir esta sensibilidade é ponderar cada voto em função da distância ao respectivo vizinho
  - **Heurística Usual:** Peso referente ao voto de um vizinho decai de forma inversamente proporcional à distância entre esse vizinho e o objeto em questão

Exercício

- Repita o exercício anterior com a ponderação de votos pelo inverso da Distância Euclidiana e discuta o resultado, comparando com o resultado anterior

Febre	Enjôo	Mancha	Diagnóstico
0	1	3	doente
1	0	2	saudável
2	1	3	doente
2	0	0	saudável
0	0	4	doente
1	0	1	???

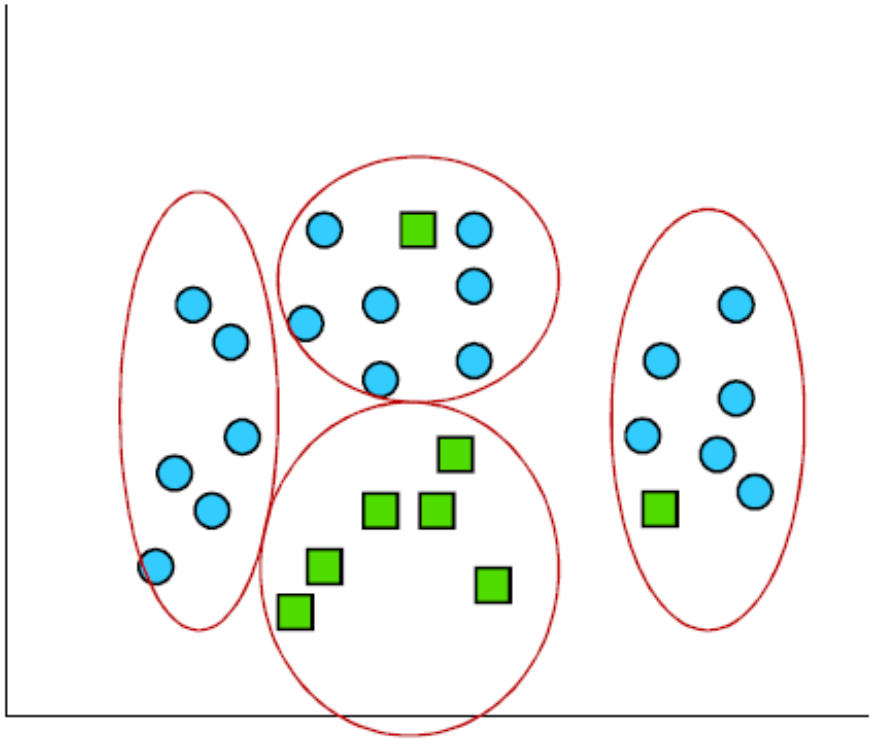
kNN: Características

- kNN não constrói explicitamente um modelo
  - Isso torna a classificação de novos objetos relativamente custosa computacionalmente
  - É necessário calcular as distâncias de cada um dos objetos a serem classificados a todos os objetos da base de instâncias rotuladas armazenada
    - Problema pode ser amenizado com algoritmos e estruturas de dados apropriados (além do escopo deste curso)

kNN: Características

- Sensíveis ao projeto
  - Escolha de K e da medida de (dis)similaridade...
- Podem ser sensíveis a ruído
  - Pouco robustos para K pequeno
- É sensível a atributos irrelevantes
  - distorcem o cálculo das distâncias
- Podem ter poder de classificação elevado
  - Função de discriminação muito flexível para K pequeno

Classificação x *Clustering*



Classificação:

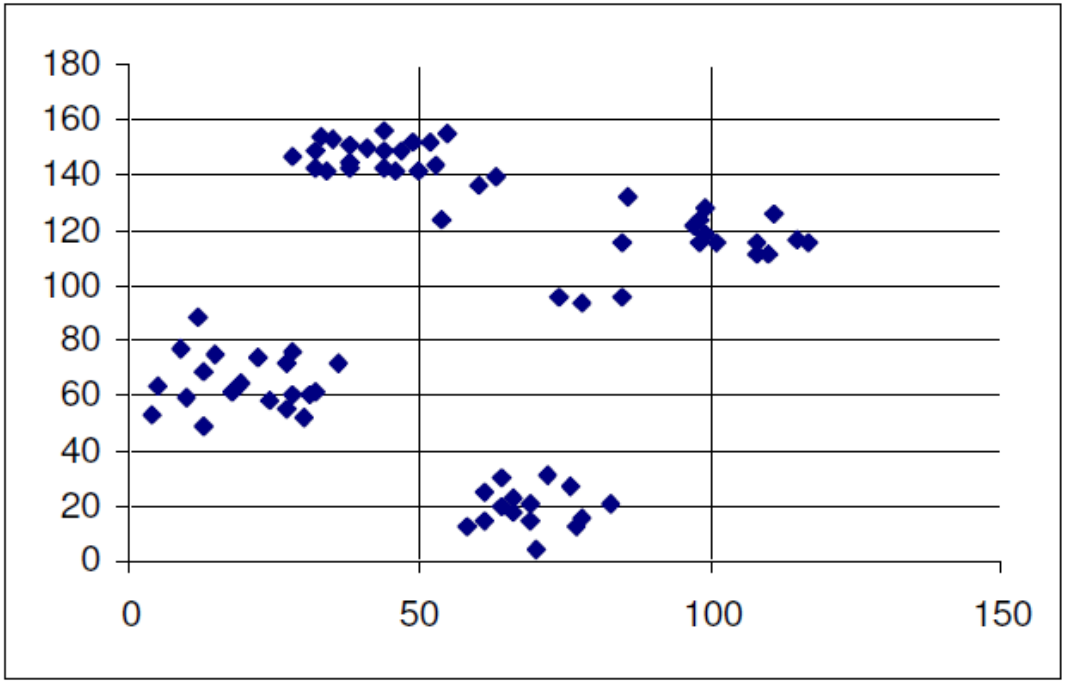
- Aprender um método para predizer as categorias (classes) de instâncias não vistas a partir de exemplos pré-rotulados (classificados)

Agrupamento de Dados (*Clustering*):

- Encontrar os rótulos das categorias (grupos ou **clusters**) diretamente a partir dos dados

Agrupamento de Dados (*Clustering*)

- Aprendizado não supervisionado
  - Encontrar grupos “naturais” de objetos não rotulados...
    - tais que objetos em um mesmo grupo sejam similares ou relacionados entre si e diferentes ou não relacionados aos demais



Definindo o que é um *Cluster*

- Conceitualmente, definições são subjetivas:
  - Homogeneidade (coesão interna)...
  - Heterogeneidade (separação entre grupos)...
- É preciso formalizar matematicamente
  - Existem diversas medidas
  - Em geral, baseadas em algum tipo de (dis)similaridade
    - Por exemplo, distância Euclidiana

Clustering

- Assim como para classificação, existem várias técnicas, para diferentes contextos de aplicação
  - Sucesso de cada método depende do domínio de aplicação e do problema particular em mãos
  - Análise Exploratória de Dados!

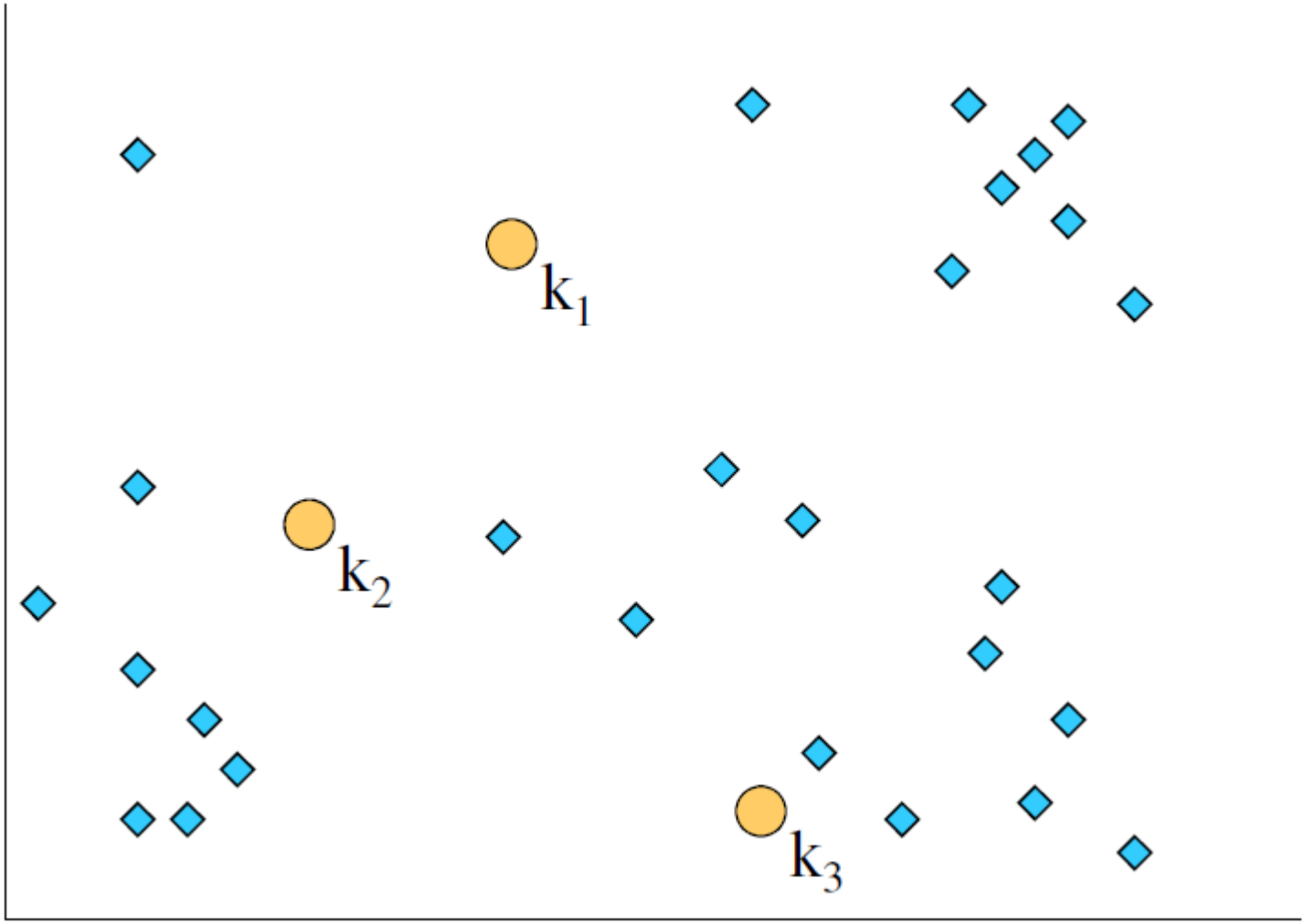
k-Means

- O Algoritmo *k-Means* é um dos mais simples e populares algoritmos de agrupamento de dados.
  - Minimiza as distâncias intra-grupos
    - indiretamente maximiza as distâncias inter-grupos

k-Means (k-Médias)

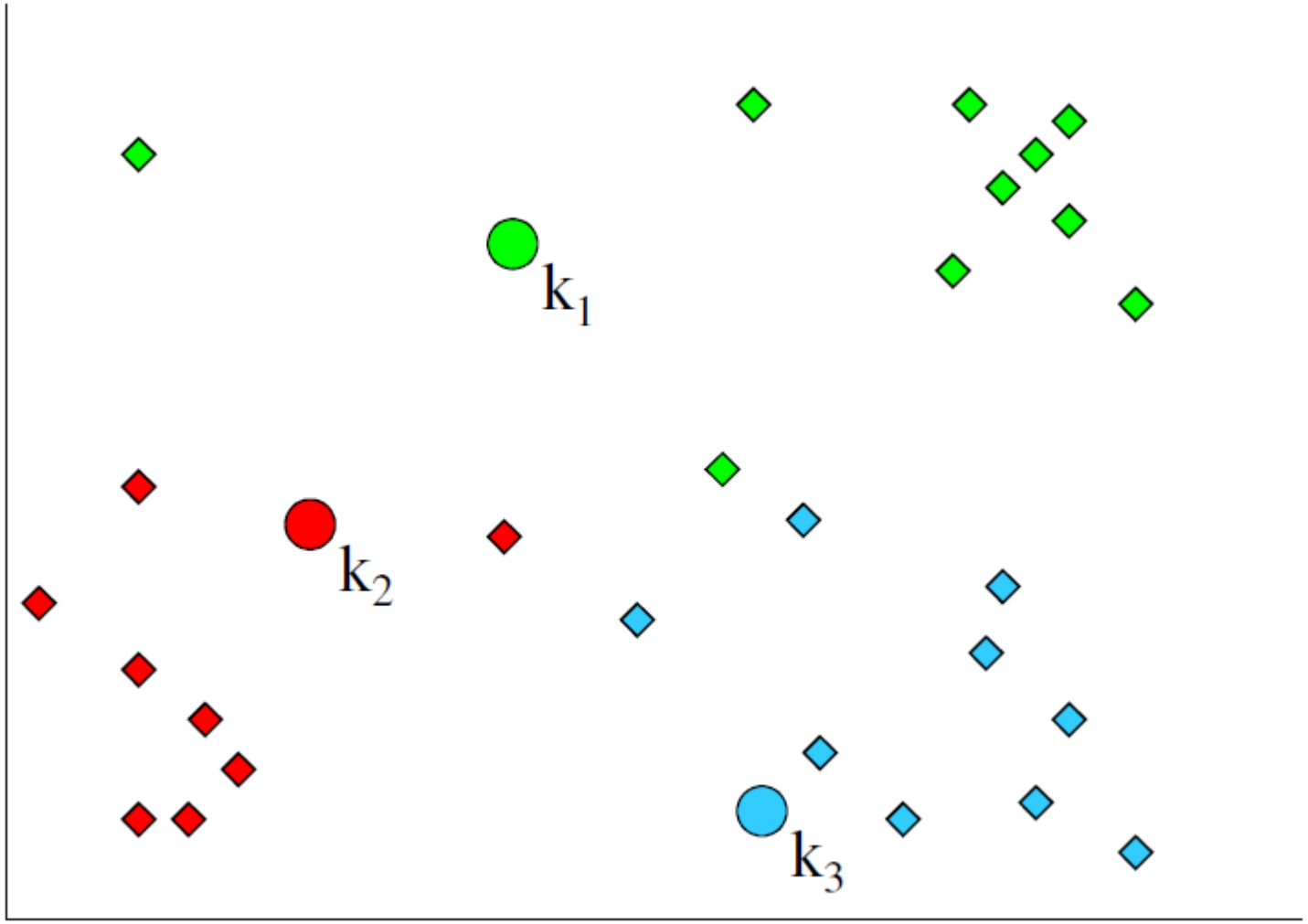
1. Escolher aleatoriamente um número k de protótipos (centros) para os clusters
2. Atribuir cada objeto para o cluster de centro mais próximo (segundo alguma distância, e.g. Euclidiana)
3. Mover cada centro para a média (centróide) dos objetos do cluster correspondente
4. Repetir os passos 2 e 3 até que algum critério de convergência seja obtido:
   - número máximo de iterações
   - limiar mínimo de mudanças nos centróides

k–Means : passo 1



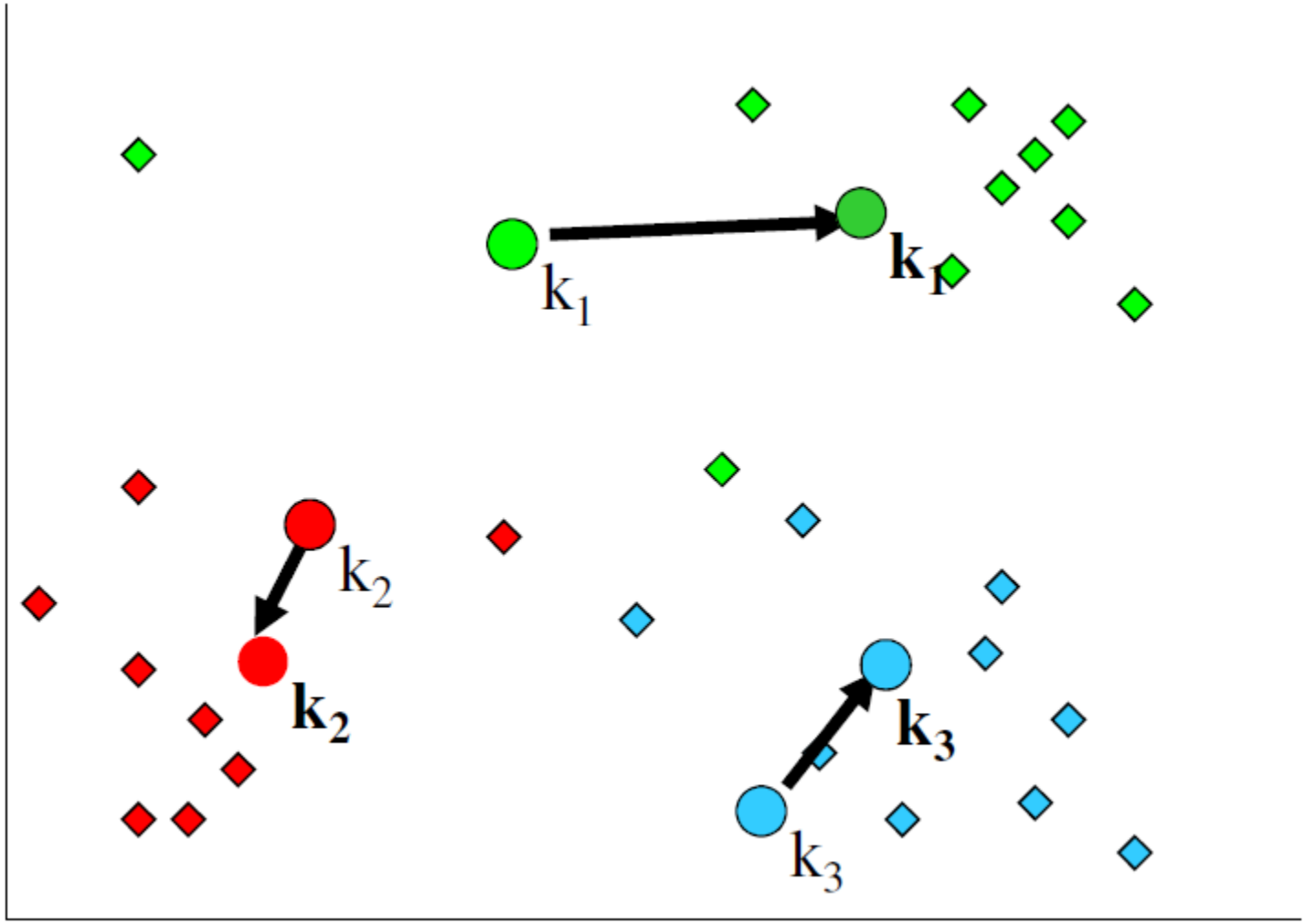
Escolher 3 centros iniciais

k–Means : passo 2



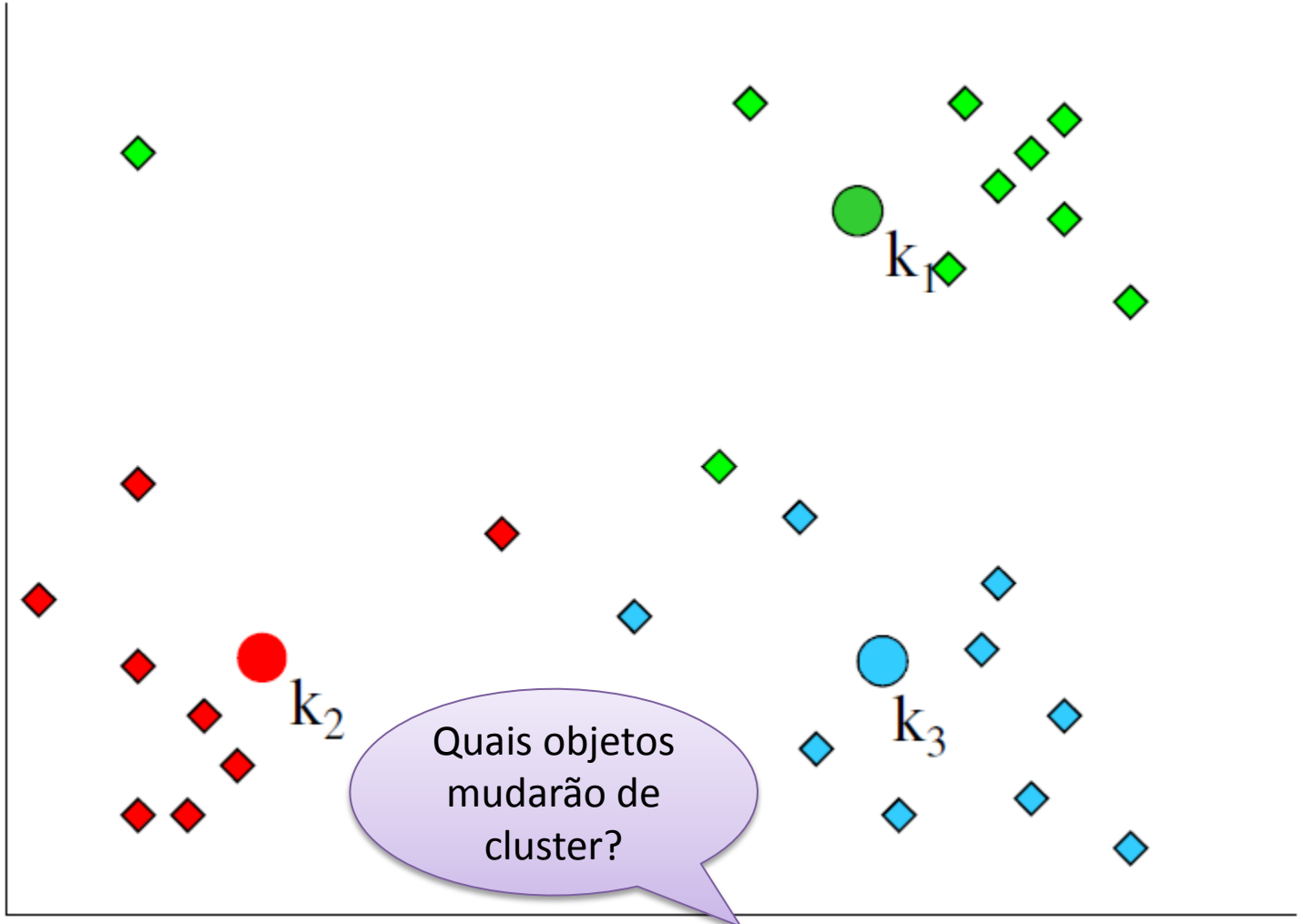
Atribuir cada objeto ao cluster de centro + próximo

k–Means : passo 3



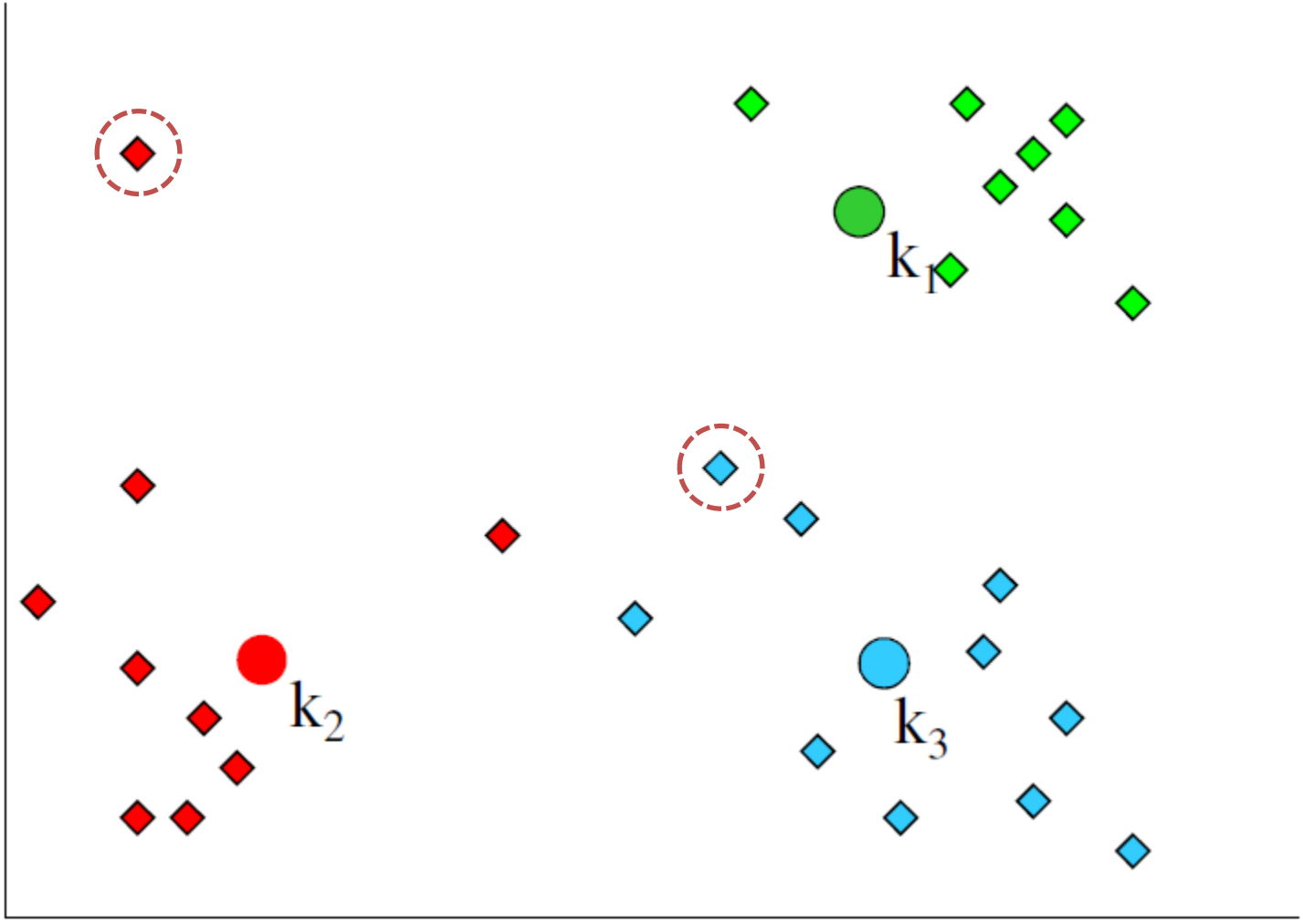
Mover cada centro para o vetor médio do cluster (centróide)

k–Means : passo 4

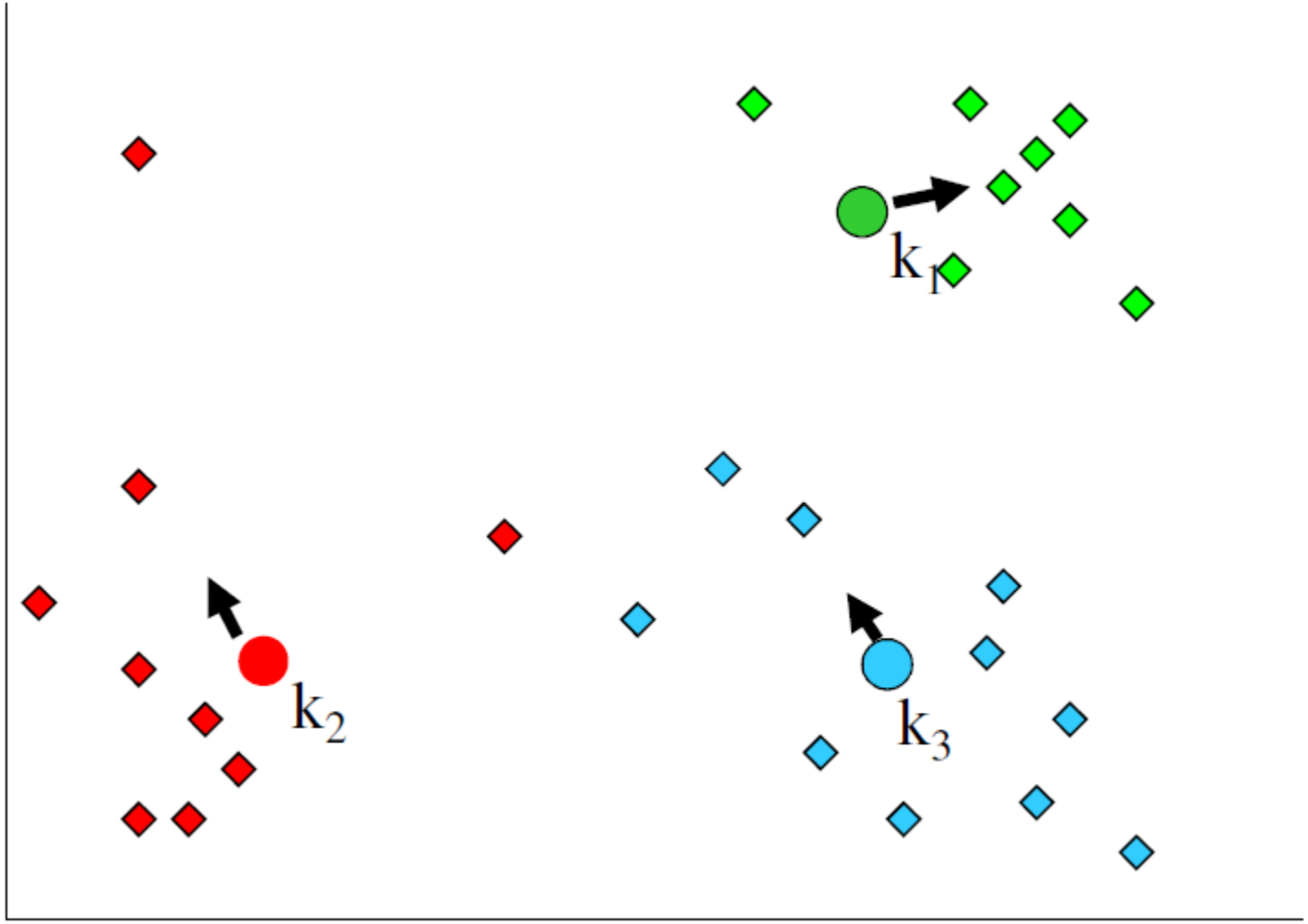


Re-atribuir objetos aos clusters de centróides mais próximos

k–Means : passo 5

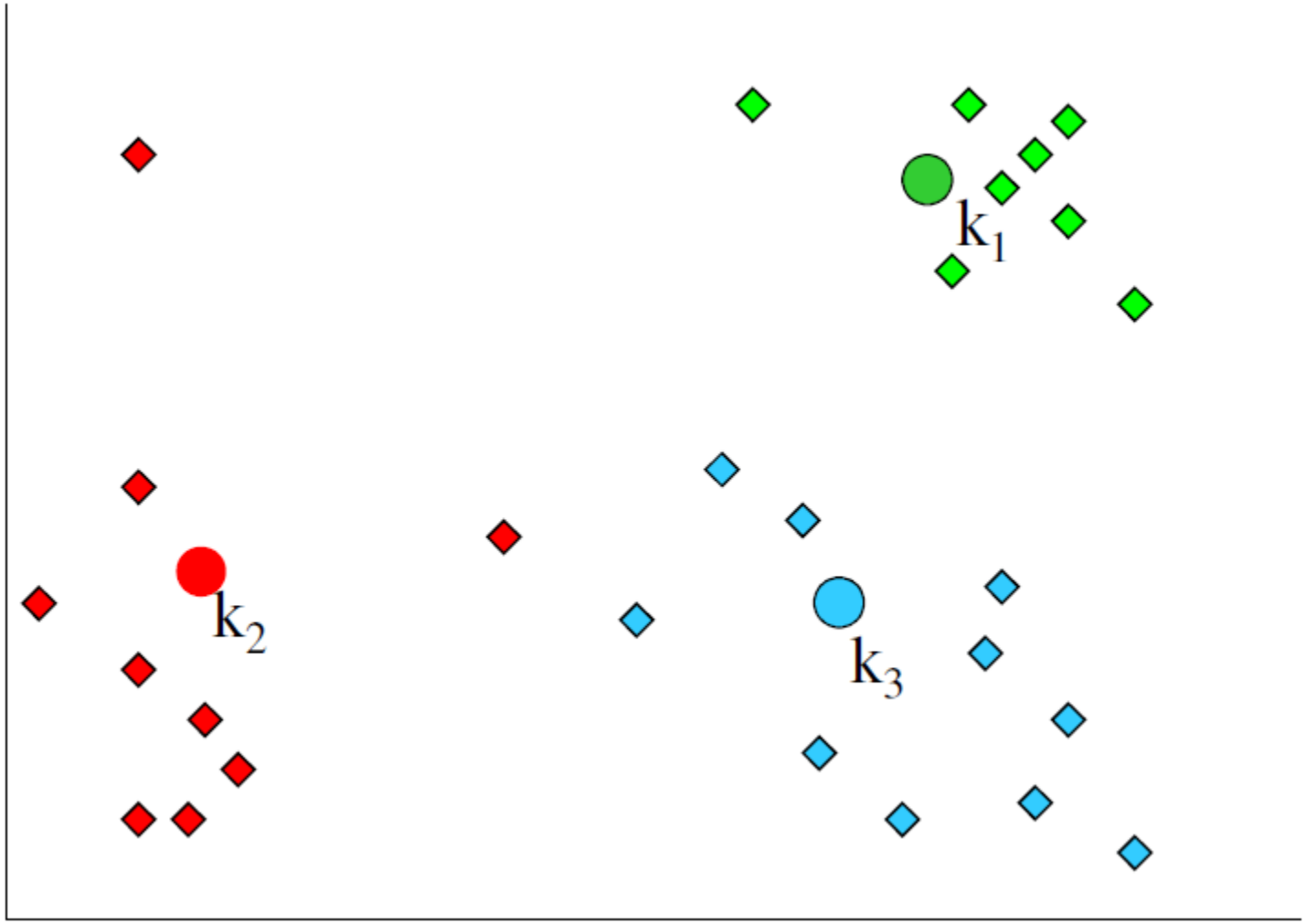


k–Means : passo 6



re-calcular vetores médios

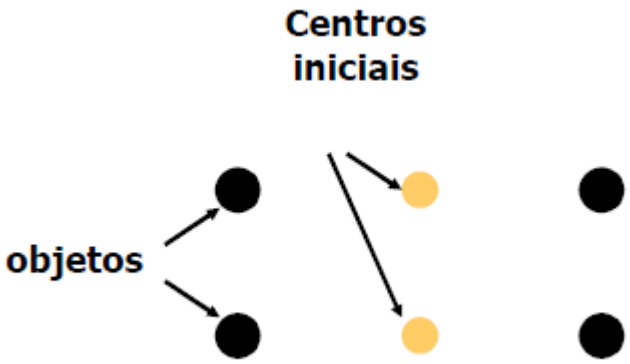
k–Means : passo 7



Mover centros dos clusters…

Aspectos importantes ...

- O resultado pode variar significativamente dependendo da escolha das sementes (protótipos) iniciais;
- *k*-means pode "ficar preso" em ótimos locais;
- Exemplo:
  - Como evitar?



Resumo do *k*-Means

Vantagens

- Simples e intuitivo
- Possui complexidade computacional linear em todas as variáveis críticas
- Eficaz em muitos cenários de aplicação e produz resultados de interpretação simples
- Considerado um dentre os 10 mais influentes algoritmos em mineração de dados

Desvantagens

- *k* = ?
- Sensível à inicialização dos protótipos e a *outliers*
- Limita-se a encontrar clusters volumétricos / globulares
- Cada item deve pertencer a um único cluster (partição rígida, ou seja, sem sobreposição)
- Limitado a atributos numéricos