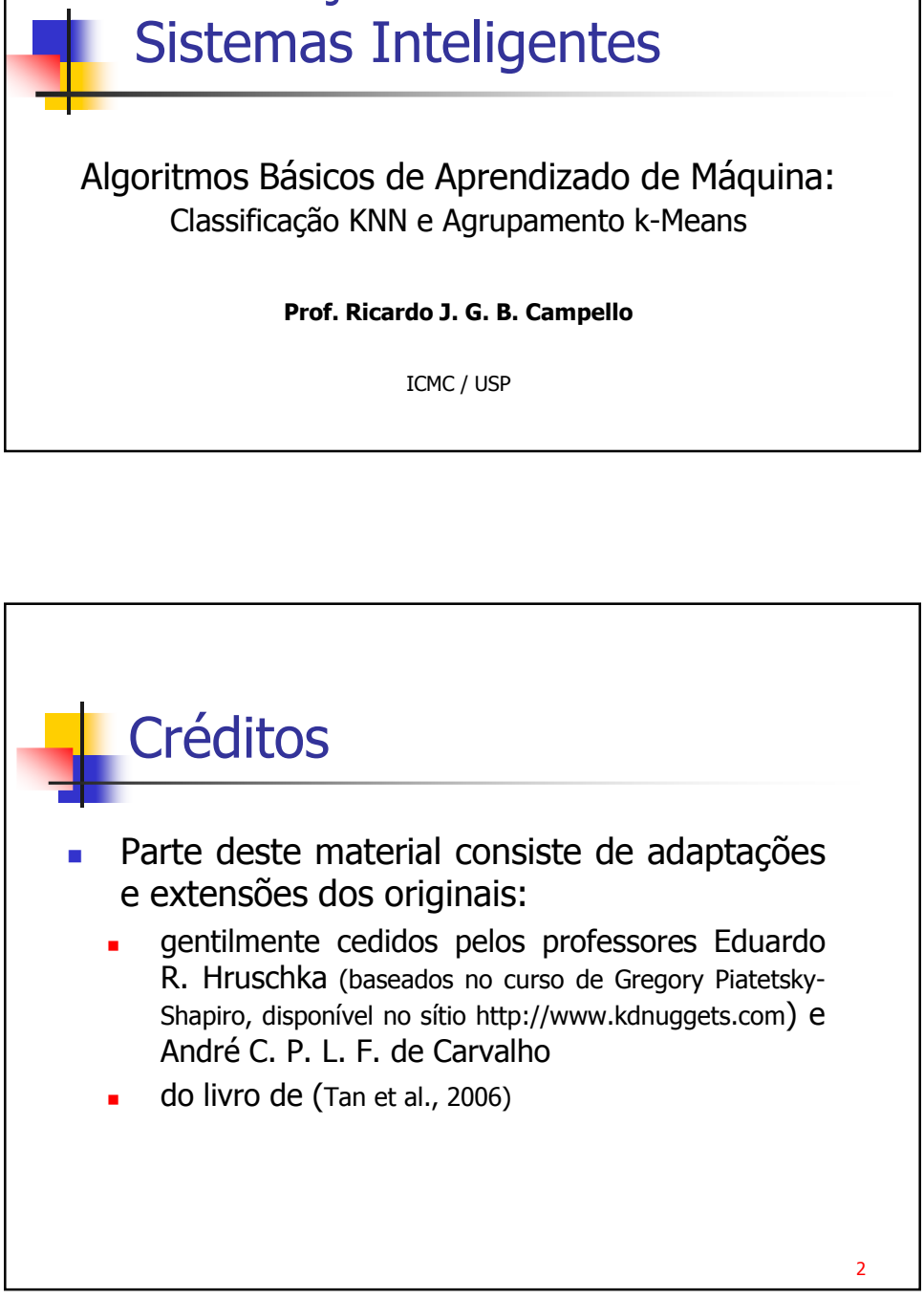


Introdução a Sistemas Inteligentes

Algoritmos Básicos de Aprendizado de Máquina:
Classificação KNN e Agrupamento k-Means

Prof. Ricardo J. G. B. Campello

ICMC / USP

Créditos

- Parte deste material consiste de adaptações e extensões dos originais:
  - gentilmente cedidos pelos professores Eduardo R. Hruschka (baseados no curso de Gregory Piatetsky-Shapiro, disponível no sítio http://www.kdnuggets.com) e André C. P. L. F. de Carvalho
  - do livro de (Tan et al., 2006)

2

Aula de Hoje

- Introdução
- Aprendizado de Máquina (AM) Supervisionado
  - Classificação
  - Algoritmo KNN
- AM Não Supervisionado
  - Agrupamento de Dados
  - Algoritmo das k-médias (*k-means*)

3

Relembrando AM...

- Dentre os principais paradigmas de treinamento em AM tem-se:
  - **Supervisionado**
  - **Não supervisionado**

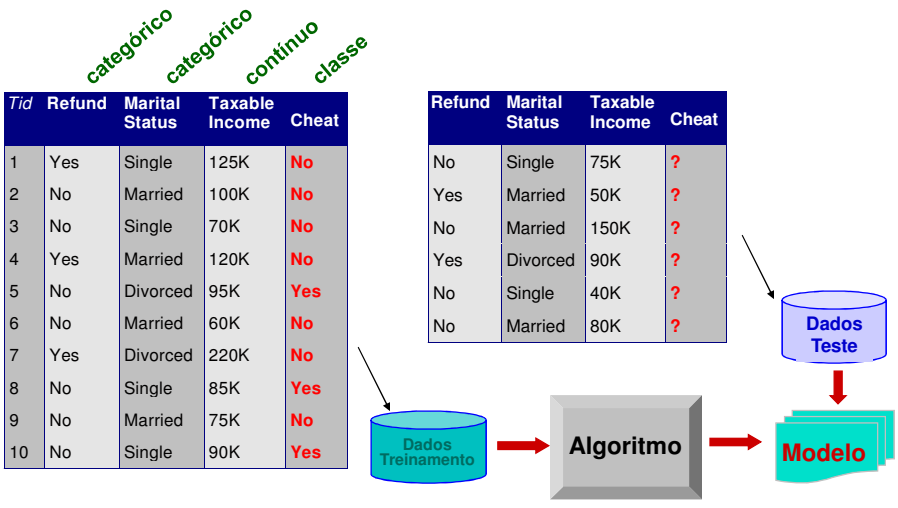
4

Classificação

- Técnica **supervisionada** que classifica novas instâncias em uma ou mais classes conhecidas
  - Número definido de classes
  - Freqüentemente apenas duas (**classificação binária**)
- Exemplos
  - Diagnóstico, Análise de crédito, ...

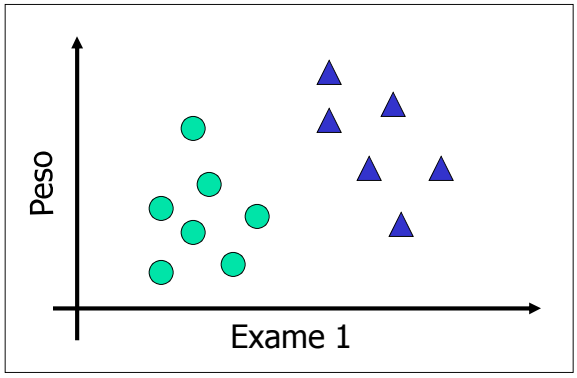
5

Exemplo de Classificação



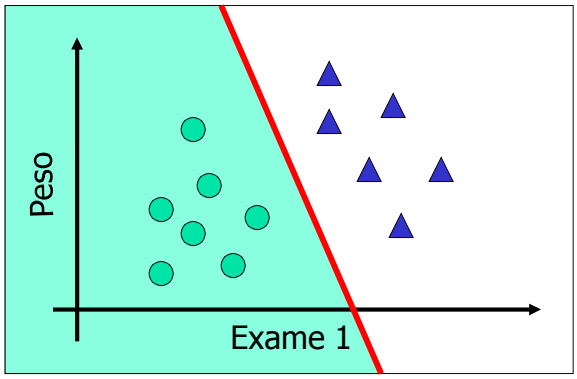
Classificação

- Como classificar?



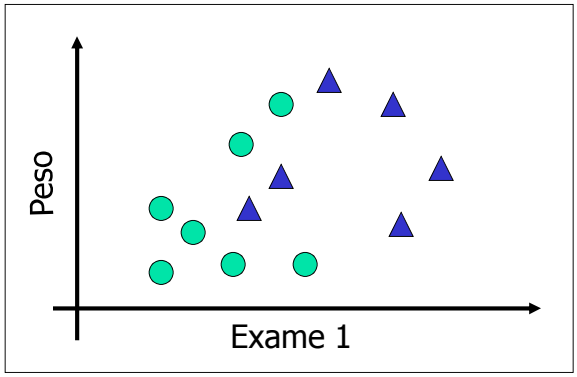
Classificação

- Problema linear



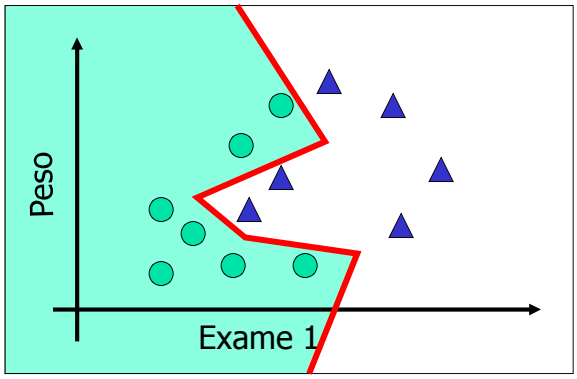
Classificação

- Como classificar?



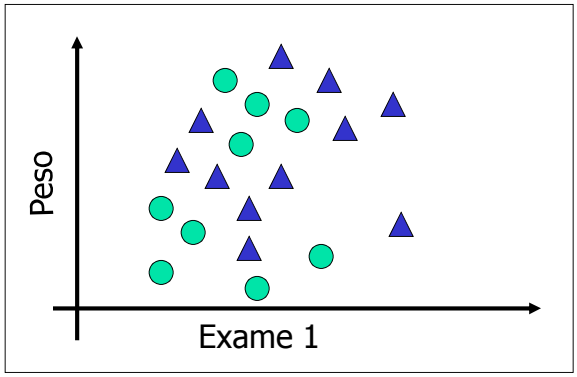
Classificação

- Problema não linear



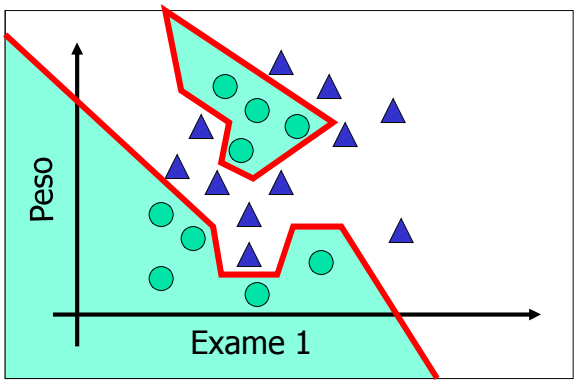
Classificação

- Como classificar?



Classificação

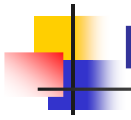
- Problema não linear



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** !

13



K-NN

- O Algoritmo K-NN (K-Vizinhos-Mais-Próximos ou K-Nearest-Neighbors do inglês) é um dos mais simples e bem difundidos algoritmos do **paradigma baseado em instâncias**

14

Classificadores Baseados em Instâncias

Conjunto de Instâncias Armazenadas

Atr1	.....	AtrN	Class
			A
			B
			B
			C
			A
			C
			B

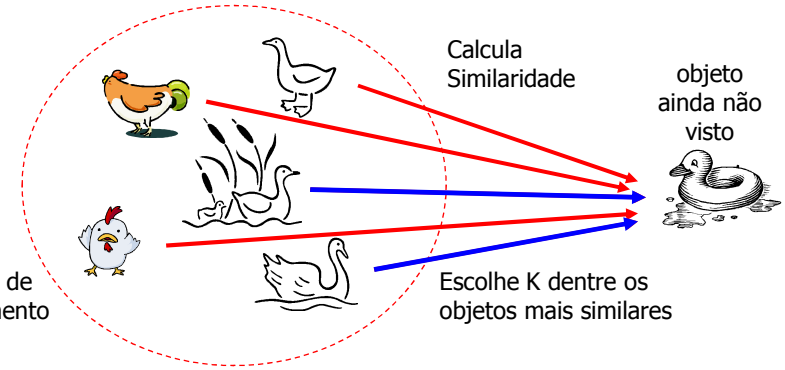
- Armazena dados de treinamento
- Usa os dados de treinamento para predizer os rótulos de classe das instâncias ainda não vistas

Instância Nova (desconhecida)

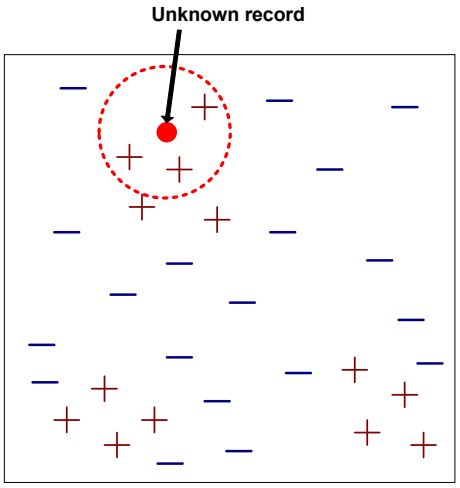
Atr1	.....	AtrN

K-NN

- Idéia Básica:
  - Se anda como um pato, "quacks" como um pato, então provavelmente é um pato

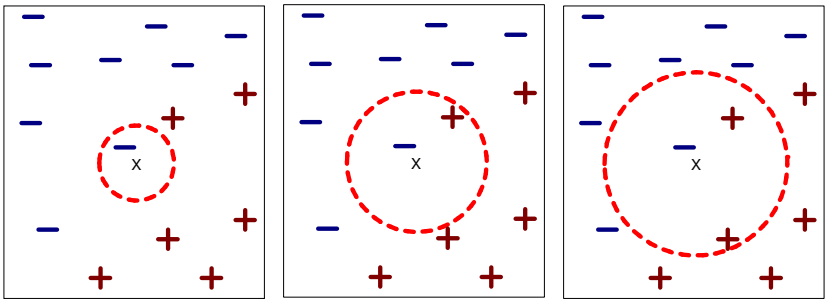


K-NN



- 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

K-NN

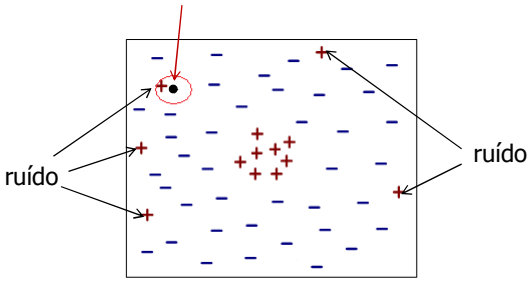


(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

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

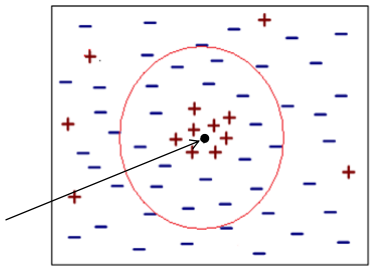
K-NN

- 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)



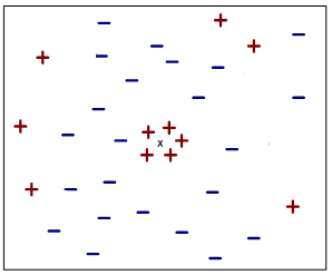
K-NN

- Escolha do Valor de K:
  - Muito grande:
    - mais robusto a ruído
    - porém, menor flexibilidade de discriminação entre classes
      - privilegia classe majoritária...



K-NN: Configuração

- Valor Ideal ?
  - Depende da aplicação
  - **Análise Exploratória de Dados** !



21

K-NN

- 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, Casamento Simples (Simple Matching), Jaccard, Cosseno, Pearson, ...

K-NN

- 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 classifique a última instância com KNN equipado com Distância Euclidiana e 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	???

24

K-NN 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	???

26

K-NN: Características

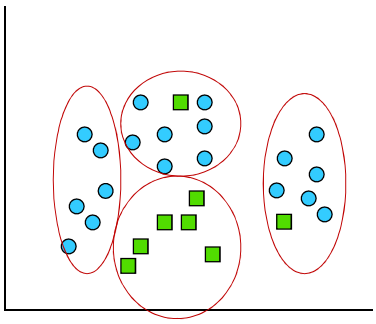
- K-NN não constrói explicitamente um modelo
  - Isso torna a classificação de novos objetos relativamente custosa computacionalmente
  - É necessário calcular as distâncias de cada um dos objetos a serem classificados a todos os objetos da base de instâncias rotuladas armazenada
    - Problema pode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

27

K-NN: 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)

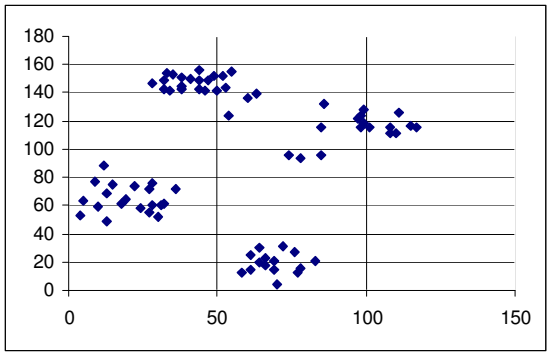
Agrupamento de Dados (*Clustering*):

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

Slide baseado no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com

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

31

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** !

32

k-Means

- O Algoritmo das k-médias (*k-means* em inglês) é um dos mais simples e populares algoritmos de agrupamento de dados
  - Minimiza as distâncias **intra-grupos**
    - indiretamente maximiza as distâncias **inter-grupos**

33

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:

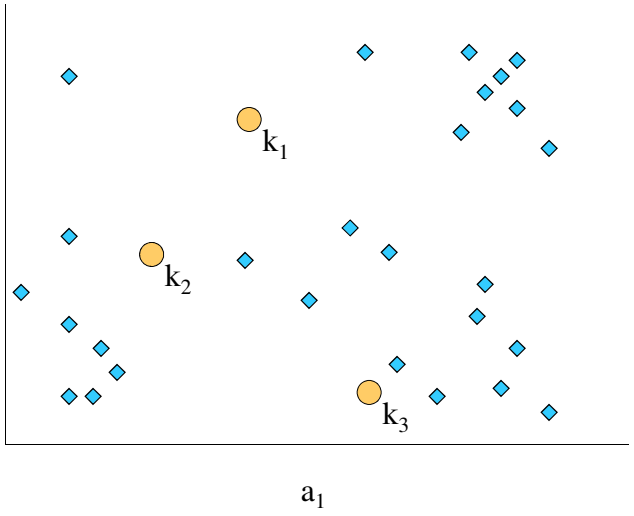
   - número máximo de iterações

   - limiar mínimo de mudanças nos centróides

34

Prof. Eduardo R. Hruschka

k-means - passo 1:

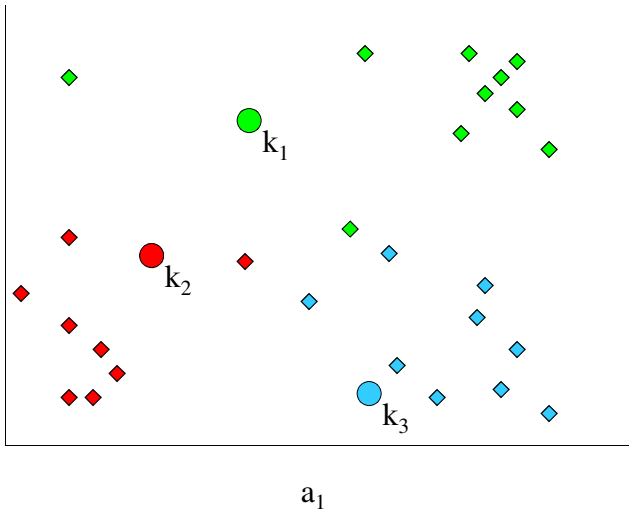


Escolher 3 centros iniciais

Slide baseado no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com

35

k-means - passo 2:



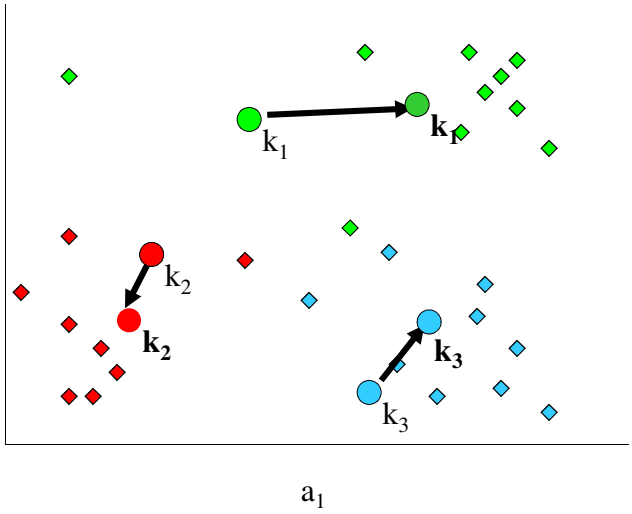
Atribuir cada objeto ao cluster de centro + próximo

Slide baseado no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com

36

k-means - passo 3:

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



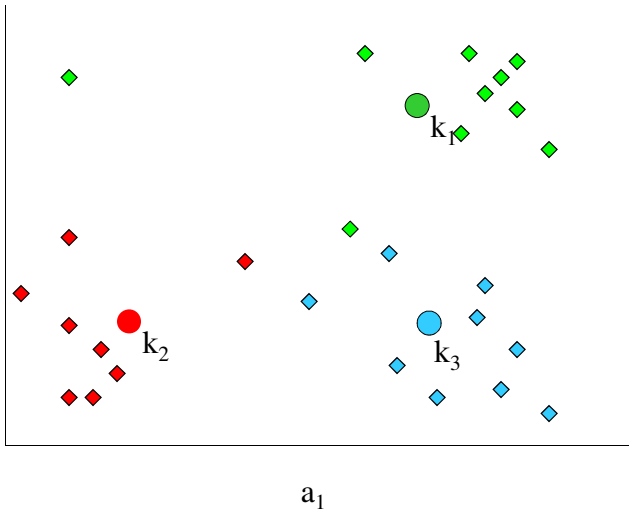
Slide baseado no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com

37

k-means:

Re-atribuir objetos aos clusters de centróides mais próximos

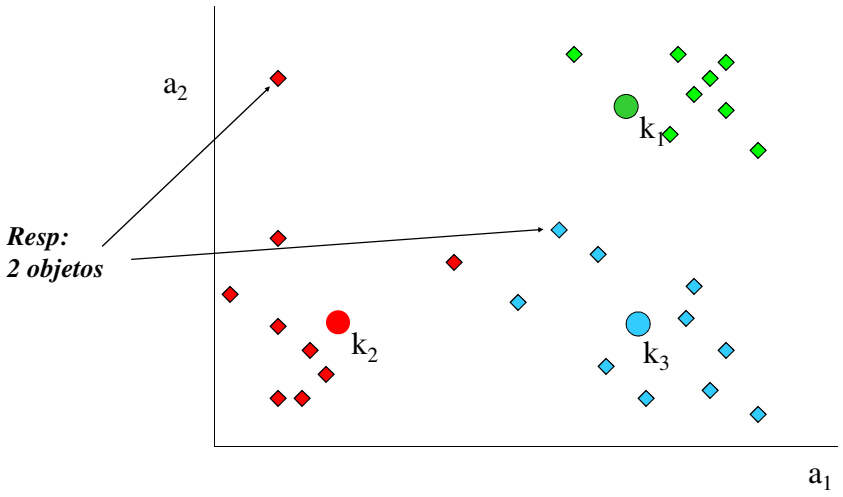
Quais objetos mudarão de cluster?



Slide baseado no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com

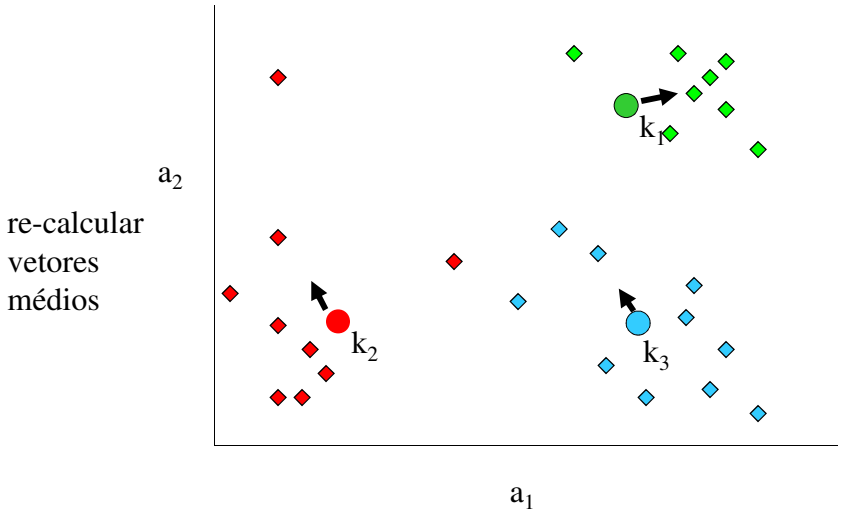
38

k-means:



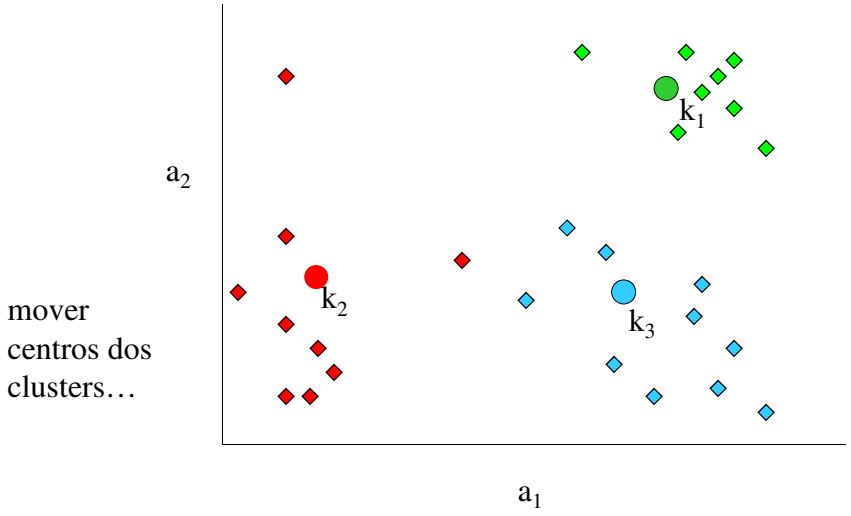
Slide baseado no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com 39

k-means:



Slide baseado no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com 40

k-means:



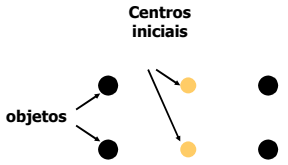
mover centros dos clusters…

Slide baseado no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com

41

Discussão

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

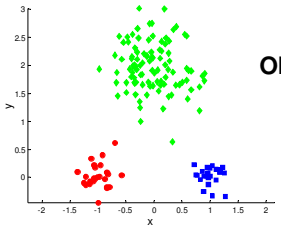


- Como evitar ... ?

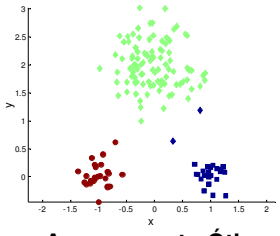
Slide baseado no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com

42

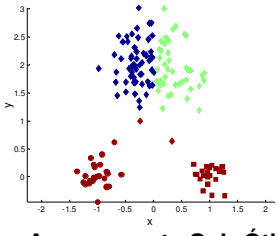
Exemplo (2 execuções diferentes)



Objetos Originais



Agrupamento Ótimo

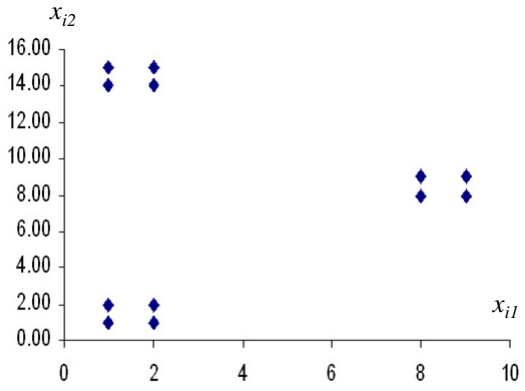


Agrupamento Sub-Ótimo

Exercício

Objeto $\mathbf{x}_i$	x_{i1}	x_{i2}
1	1	2
2	2	1
3	1	1
4	2	2
5	**8**	**9**
6	**9**	**8**
7	**9**	**9**
8	**8**	**8**
9	1	15
10	2	15
11	1	14
12	2	14

Prof. Eduardo R. Hruschka



- Executar k-means com k=3 nos dados acima a partir dos protótipos [6 6], [4 6] e [5 10] e outros a sua escolha

44

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

45

Referências

- P.-N. Tan, Steinbach, M., and Kumar, V., *Introduction to Data Mining*, Addison-Wesley, 2006

46