SCC0173 – Mineração de Dados Biológicos

Classificação I: Algoritmos 1Rule e KNN

Prof. Ricardo J. G. B. Campello

SCC / ICMC / USP

1

Créditos

- O material a seguir consiste de adaptações e extensões dos originais:
  - gentilmente cedidos pelo Prof. Eduardo R. Hruschka
  - de Tan et al., *Introduction to Data Mining*, Addison-Wesley, 2006
  - de Witten & Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, M. Kaufmann, 2005

2

Aula de Hoje

- Introdução
  - Classificação
- Classificação via Aprendizado de Máquina (AM) Simbólico
  - Algoritmo 1R (One Rule)
- Classificação via AM Baseado em Instâncias
  - Algoritmo KNN

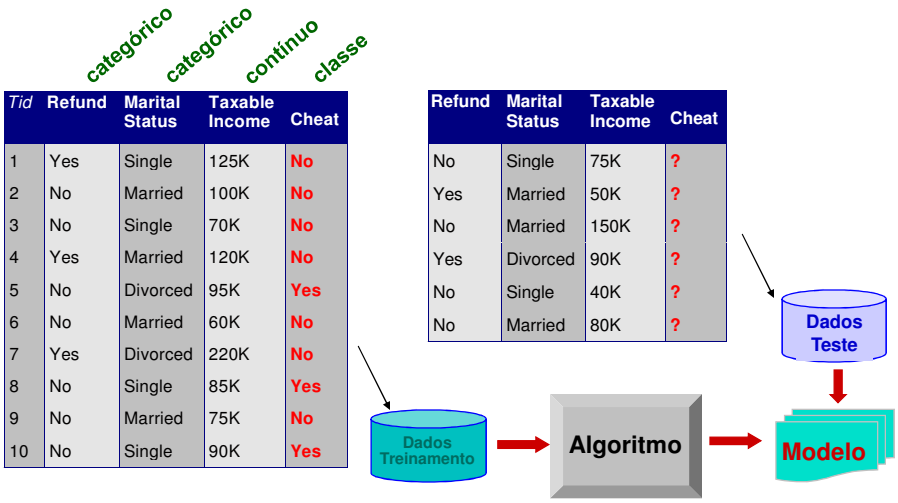
3

Classificação

- Técnica classifica novas entradas (instâncias) em uma ou mais dentre diferentes classes discretas
  - Número definido de classes
  - Freqüentemente apenas duas
    - **classificação binária**
- Exemplos
  - Diagnóstico, Análise de crédito, ...

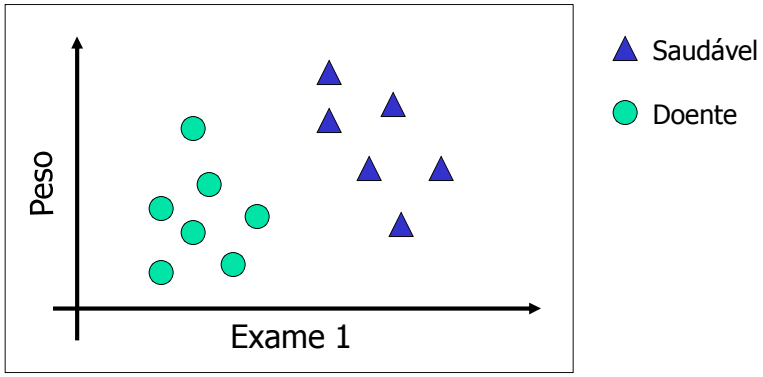
4

Exemplo de Classificação



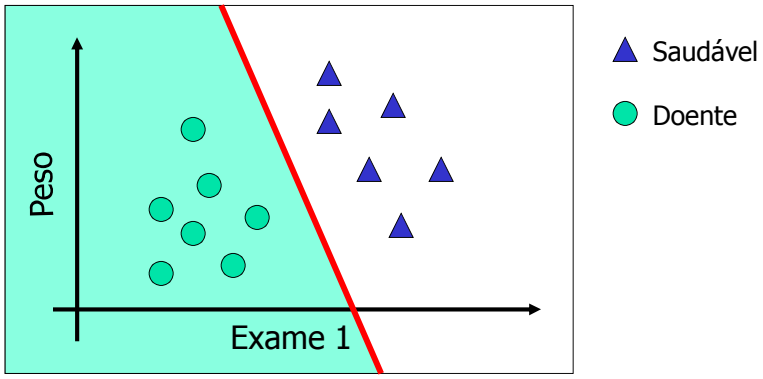
Classificação

- Como classificar?



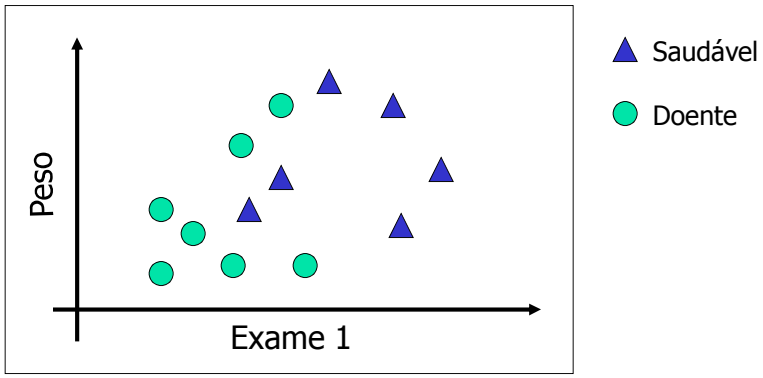
Classificação

- Problema linear



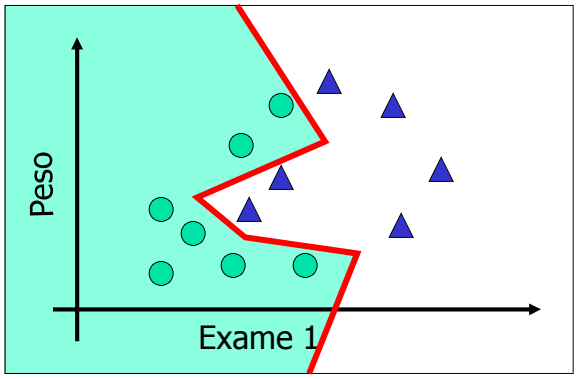
Classificação

- Como classificar?



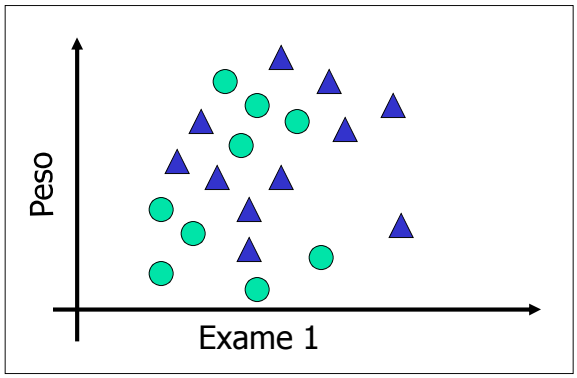
Classificação

- Problema não linear



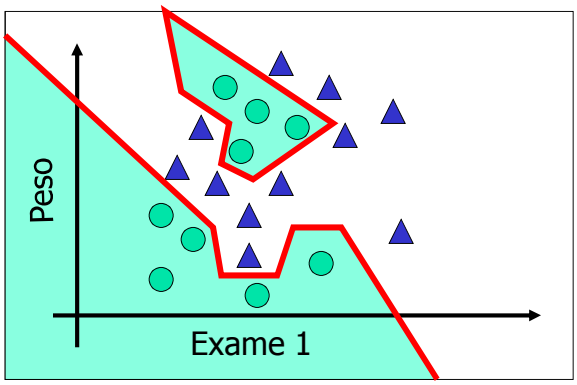
Classificação

- Como classificar?



Classificação

- Problema não linear



Exemplo: Problema *Weather*

(Witten & Frank, 2005)

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	**No**
Sunny	Hot	High	True	**No**
Overcast	Hot	High	False	**Yes**
Rainy	Mild	High	False	**Yes**
Rainy	Cool	Normal	False	**Yes**
Rainy	Cool	Normal	True	**No**
Overcast	Cool	Normal	True	**Yes**
Sunny	Mild	High	False	**No**
Sunny	Cool	Normal	False	**Yes**
Rainy	Mild	Normal	False	**Yes**
Sunny	Mild	Normal	True	**Yes**
Overcast	Mild	High	True	**Yes**
Overcast	Hot	Normal	False	**Yes**
Rainy	Mild	High	True	**No**

Prof. Eduardo R. Hruschka

12

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** !

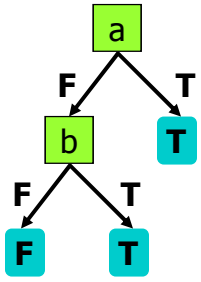
13

Exemplo

- Árvores de Decisão

a **OR** b

a	b	a v b
F	F	F
F	T	T
T	F	T
T	T	T



14

Algoritmo Rudimentar (1 Rule – 1R)

- 1R: Aprende uma árvore de decisão de um nível
  - Todas as regras usam somente um atributo
  - Atributo deve assumir valores categóricos
    - **Paradigma simbólico** de AM
- Versão Básica:
  - Um ramo para cada valor possível do atributo
  - Para cada ramo, atribuir a classe mais freqüente
  - Para cada ramo, calcular a taxa de erro de classificação
  - Escolher o atributo com a menor taxa de erro de classificação

15

Prof. Eduardo R. Hruschka

Pseudo-Código para o 1R:

Para cada atributo:
Para cada valor do atributo gerar uma regra como segue:
Contar a freqüência de cada classe;
Encontrar a classe mais freqüente*;
Formar uma regra que atribui a classe mais freqüente a este atributo-valor;
Calcular a taxa de erro de classificação das regras;
Escolher as regras com a menor taxa de erro de classificação.

*** Empates na classe mais freqüente podem ser decididos aleatoriamente.**

NOTA: Está implementado no software Weka

16

Prof. Eduardo R. Hruschka

Exemplo: Problema *Weather*

(Witten & Frank, 2005)

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Attribute	Rules	Errors	Total Errors
Outlook	Sunny → No	2/5	4/14
	Overcast → Yes	0/4	
	Rainy → Yes	2/5	
Temp	Hot → No*	2/4	5/14
	Mild → Yes	2/6	
	Cool → Yes	1/4	
Humidity	High → No	3/7	4/14
	Normal → Yes	1/7	
Windy	False → Yes	2/8	5/14
	True → No*	3/6	

1R seria composto ou das 3 regras para Outlook ou das 2 Regras para Humidity: decisão poderia ser feita, por ex., de acordo com o desempenho em um outro conjunto de dados (dados de teste)

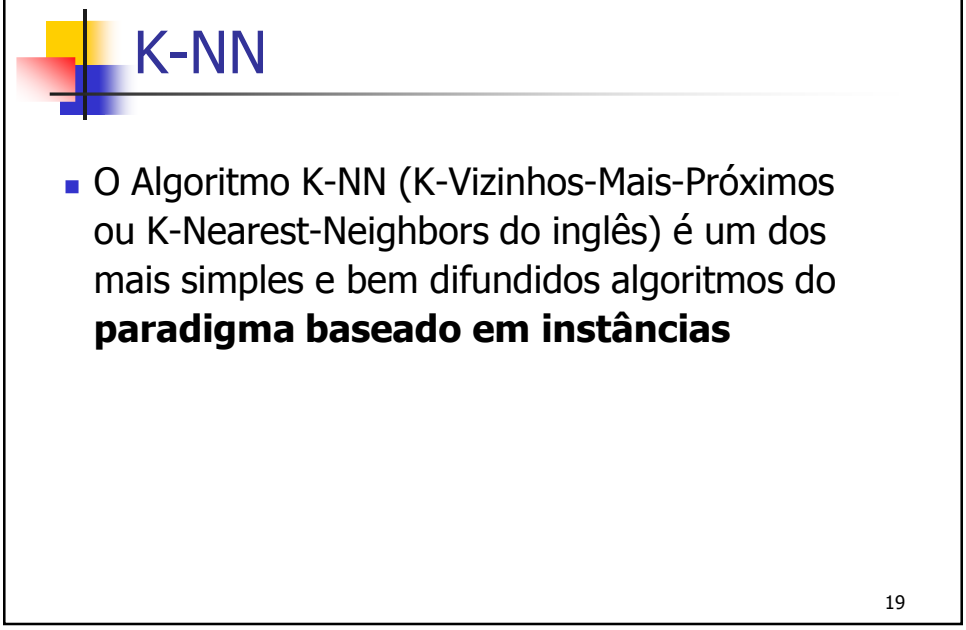
Prof. Eduardo R. Hruschka

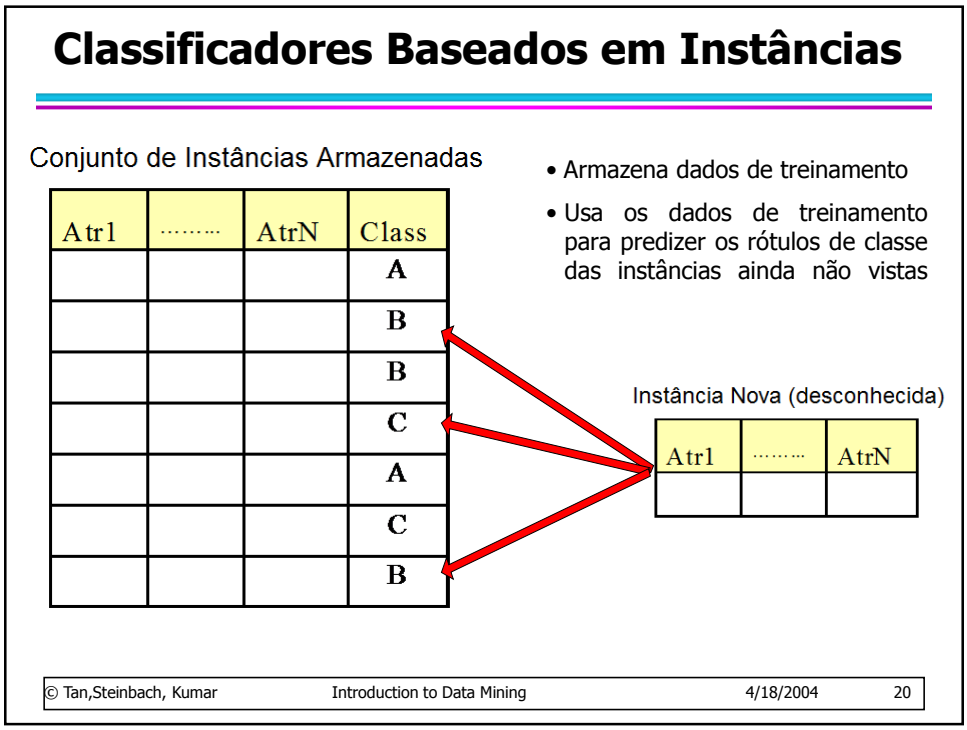
Exercício

- Obter um classificador 1R para os dados:

Febre	Enjôo	Mancha	Dor	Diagnóstico
Sim	Sim	Não	Sim	Não
Não	Sim	Não	Não	Sim
Sim	Sim	Sim	Não	Sim
Sim	Não	Não	Sim	Não
Sim	Não	Sim	Sim	Sim
Não	Não	Sim	Sim	Não

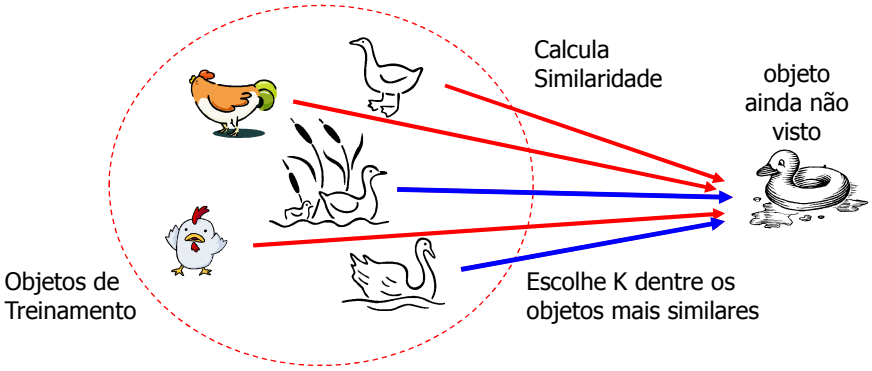
18


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**
19

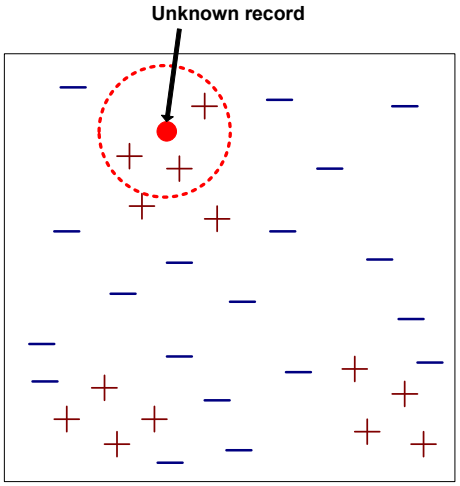

Classificadores Baseados em Instâncias
Conjunto de Instâncias Armazenadas
Atr1 ..... AtrN Class
A
B
B
C
A
C
B
• Armazena dados de treinamento
• Usa os dados de treinamento para predizer os rótulos de classe das instâncias ainda não vistas
Instância Nova (desconhecida)
Atr1 ..... AtrN
© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 20

K-NN

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

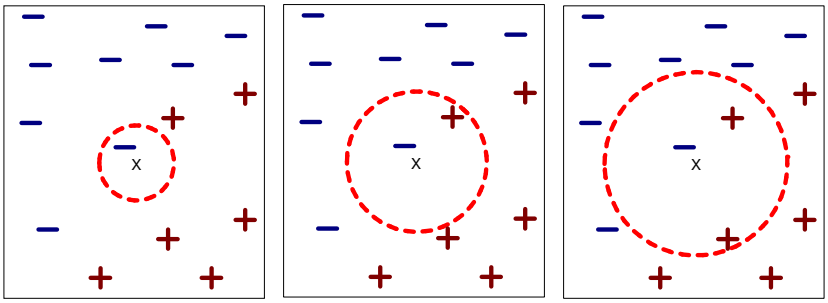


K-NN



- 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

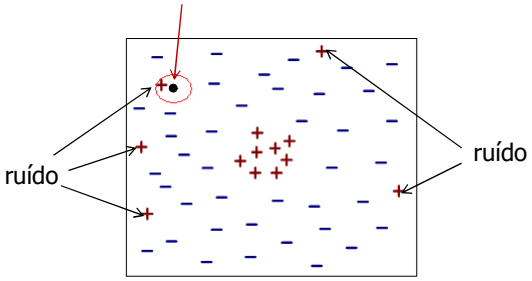


(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

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

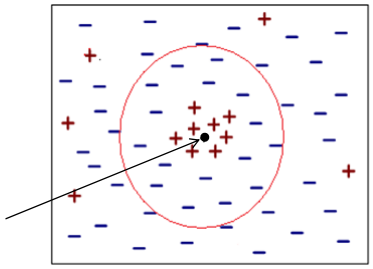
K-NN

- 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)



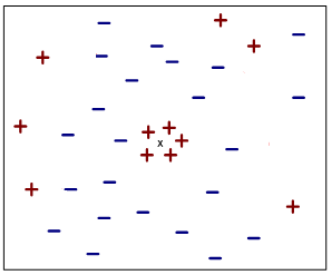
K-NN

- 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** !



K-NN

- Como calcular as (dis)similaridades... ?
  - Já vimos anteriormente no curso que a medida 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

- Converta os dados abaixo para valores numéricos em [0, 1] (sem aumentar o no. de atributos) e classifique a última instância com KNN equipado com Distância Euclidiana e K = 1, 3 e 5. Discuta.

Febre	Enjôo	Mancha	Diagnóstico
baixa	sim	grande	doente
média	não	média	saudável
alta	sim	grande	doente
alta	não	ausente	saudável
baixa	não	enorme	doente
média	não	pequena	**???**

29

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
    - Nota: está implementada no software **Weka**

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
baixa	sim	grande	doente
média	não	média	saudável
alta	sim	grande	doente
alta	não	ausente	saudável
baixa	não	enorme	doente
média	não	pequena	???

31

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)

32

K-NN: Características

- Sensíveis ao projeto
  - Escolha de K...
  - Escolha da medida de (dis)similaridade...
- Podem ter poder de classificação elevado
  - Função de discriminação muito flexível para K pequeno
- Podem ser sensíveis a ruído
  - Pouco robustos para K pequeno

33

K-NN: Características

- É sensível a atributos irrelevantes
  - distorcem o cálculo das distâncias
  - maldição da dimensionalidade...
    - demanda seleção de atributos (veremos depois no curso)
- Por outro lado, permitem atribuir importâncias distintas para diferentes atributos
  - estratégias de ponderação de atributos
  - geralmente levam a classificadores mais precisos
    - além do escopo deste curso

34

Perguntas