Classificação: Árvores de Decisão e k-NN

Eduardo Raul Hruschka

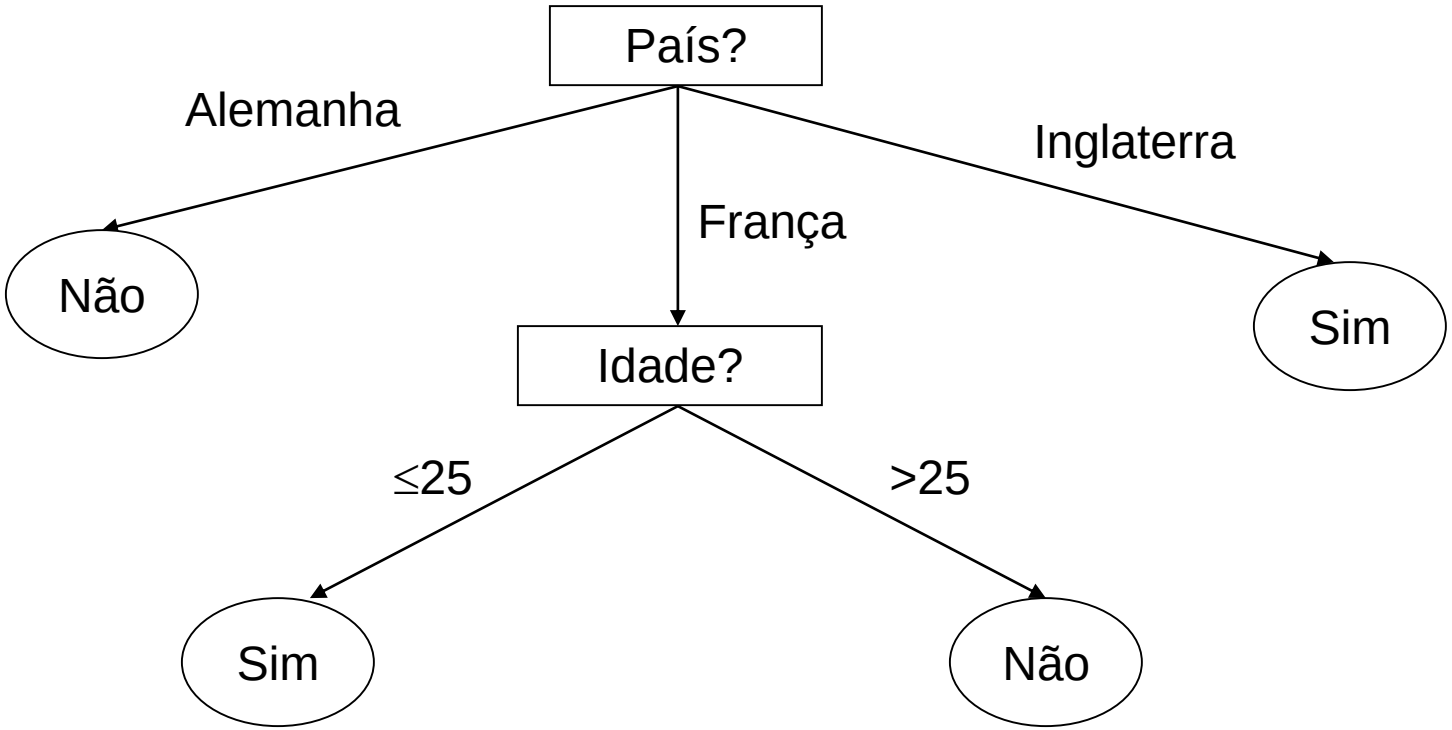
Agenda:

- Conceitos de Classificação
- Técnicas de Classificação
  - *One Rule* (1R)
  - Naive Bayes (com seleção de atributos)
  - Árvores de Decisão
  - K-Vizinhos Mais Próximos (K-NN)
- Super-ajuste e validação cruzada
- Combinação de Modelos

Árvores de Decisão

- Métodos para aproximar funções discretas, representadas por meio de uma árvore de decisão;
- Árvores de decisão podem ser representadas por conjuntos de regras "*se...então*";
  - *compreensibilidade*;
- Muito utilizadas em aplicações práticas.

- Nós internos □: teste em atributos *previsores*;
- Nós externos ○: valor previsto para o atributo meta;
- Como classificar um(a) novo(a) exemplo/registro/tupla/amostra?



Ross Quinlan

➢ Abordaremos conceitos do ID3 (Quinlan, 1986) e do C4.5 (Quinlan, 1993).

Construindo uma árvore de decisão:

1. Árvore é construída de maneira *top-down*, *recursivamente* e usando a idéia de *dividir para conquistar;*
2. Inicialmente, todos os exemplos (tuplas) de treinamento são *posicionados* na raiz da árvore;
3. Exemplos são *particionados* recursivamente com base em atributos selecionados, objetivando-se separar as tuplas por classes;
4. Condições de parada:
   - Todos os exemplos para um dado nó pertencem à mesma classe;
   - Não existem mais atributos para continuar o *particionamento*;
   - Todos os exemplos de treinamento estão classificados;

Noção Intuitiva:

- Como obter uma árvore de decisão para a seguinte base de dados?
  (Freitas & Lavington, Mining Very Large Databases with Parallel Processing, Kluwer, 1998)

SEXO	PAÍS	IDADE	COMPRAR
M	França	25	Sim
M	Inglaterra	21	Sim
F	França	23	Sim
F	Inglaterra	34	Sim
F	França	30	Não
M	Alemanha	21	Não
M	Alemanha	20	Não
F	Alemanha	18	Não
F	França	34	Não
M	França	55	Não

  - Assumindo que "COMPRAR" é nosso atributo-meta (classe) ...

Para cada atributo previsor (SEXO, PAÍS, IDADE), montar uma tabela:

- as linhas contêm os valores do atributo previsor (variável independente);
- as colunas contêm os valores do atributo-meta (variável dependente);
- Cada célula contém o nº de tuplas para a combinação de atributo_classe.
- **Exemplo:** Qual atributo discrimina melhor – SEXO OU PAÍS?

Classe

SEXO	Sim	Não
M	2	3
F	2	3

Classe

PAÍS	Sim	Não
França	2	3
Inglaterra	2	0
Alemanha	0	3

Se SEXO=M então Não; Senão Sim. ⇒ Acurácia (A) = 50%.

- Mas *Regra default* (Atribuir sempre CLASSE=Não) retorna A=60%!

Se PAÍS=Inglaterra então Sim; Senão Não. ⇒ A = 80%.

⇒ O que dizer sobre o atributo IDADE? Como medir a informação?

Entropia:

- Permite medir a *informação* fornecida por cada atributo;
- Caracteriza a impureza de um conjunto de exemplos;
- Para um nó da árvore que contém *p* exemplos *positivos* e *n* exemplos *negativos*:

$$entropia = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

Observação: a definição de *p* e *n* é arbitrária.

- Considerando o atributo meta “comprar” e 10 exemplos no nó raiz (4+,6-) :
  - Entropia (E) = 0,97 --- note que P(+)=0,4 e P(-)=0,6.
  - Para cada atributo teremos um valor de E. Para o atributo SEXO:

	Classe (COMPRAR)	
SEXO	Sim	Não
M	2	3
F	2	3

Ponderando o valor de E pelo número de exemplos que apresentam determinado valor para o atributo temos:

$E(SEXO) = 5/10 (-2/5 \log_2 2/5 - 3/5 \log_2 3/5) +$ (M)
$5/10 (-2/5 \log_2 2/5 – 3/5 \log_2 3/5)$ (F)
$E(SEXO) = 0,97.$

Ganho de Informação (GI) = 0,97 – 0,97 = 0,00

Não há GI ao particionar com base no SEXO.

- Consideremos o atributo PAÍS:
- França: 2(+) / 3(-); Inglaterra: 2(+) / 0(-); Alemanha: 0(+) / 3(-);
- Tabela de Contingência:

PAÍS	COMPRAR		Total
	Sim (+)	Não (-)	
França	2	3	5
Inglaterra	2	0	2
Alemanha	0	3	3
Total	4	6	10

E(PAÍS) = 5/10.Info(França) + 2/10.Info(Inglaterra) + 3/10.Info(Alemanha) =

$+ \ 5/10\ (-2/5 \log_2 2/5 - 3/5 \log_2 3/5)$ (França)

$+ \ 2/10\ (-2/2 \log_2 2/2 - 0/2 \log_2 0/2)$ (Inglaterra)

$+ \ 3/10\ (-0/3 \log_2 0/3 - 3/3 \log_2 3/3)$ (Alemanha)

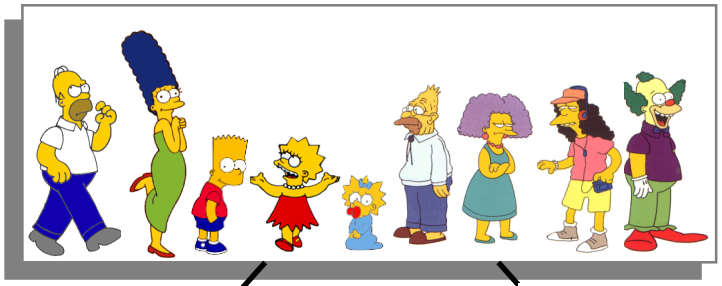
E(PAÍS) = 0,485

- Ganho de Informação (GI) / Redução na Entropia: (0,97 - 0,485) = 0,485.
- PAÍS (A=80%) é um atributo melhor do que SEXO (A=50%).
- Consideremos agora a seguinte base de dados:

ID Registro	Cabelo	Peso	Idade	Classe
 Homer	0”	250	36	**M**
 Marge	10”	150	34	**F**
 Bart	2”	90	10	**M**
 Lisa	6”	78	8	**F**
 Maggie	4”	20	1	**F**
 Abe	1”	170	70	**M**
 Selma	8”	160	41	**F**
 Otto	10”	180	38	**M**
 Krusty	6”	200	45	**M**

- **Considerar “ID Registro”? E para os demais atributos?**

Fonte: Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

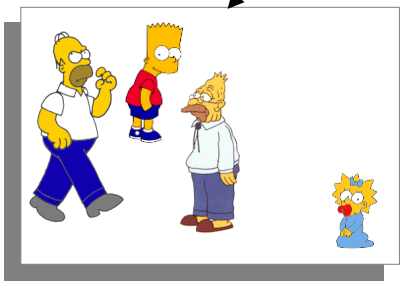


$Entropia(4\mathbf{F},5\mathbf{M}) = -(4/9)\log_2(4/9) - (5/9)\log_2(5/9)$
$= \mathbf{0.9911}$

sim

não

Cabelo <= 5?

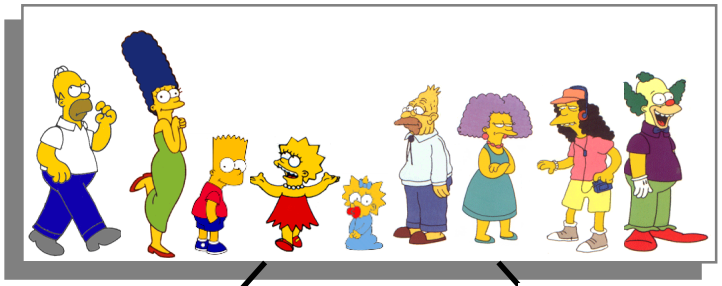


$Entropia(1\mathbf{F},3\mathbf{M}) = -(1/4)\log_2(1/4) - (3/4)\log_2(3/4)$
$= \mathbf{0.8113}$

$Entropia(3\mathbf{F},2\mathbf{M}) = -(3/5)\log_2(3/5) - (2/5)\log_2(2/5)$
$= \mathbf{0.9710}$

GI (Cabelo <= 5) = **0.9911** – (4/9 * **0.8113** + 5/9 * **0.9710**) = **0.0911**

Fonte: Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

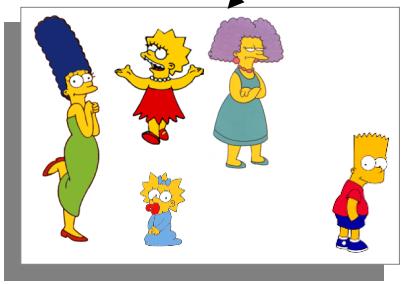


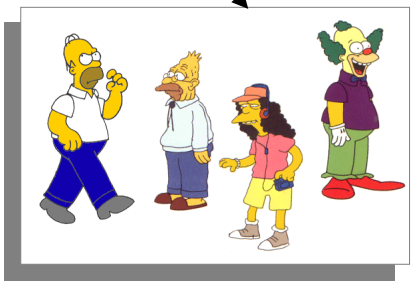
$Entropia(4\mathbf{F},5\mathbf{M}) = -(4/9)\log_2(4/9) - (5/9)\log_2(5/9)$
$= \mathbf{0.9911}$

sim

não

peso <= 160?





$Entropia(4\mathbf{F},1\mathbf{M}) = -(4/5)\log_2(4/5) - (1/5)\log_2(1/5)$
$= \mathbf{0.7219}$

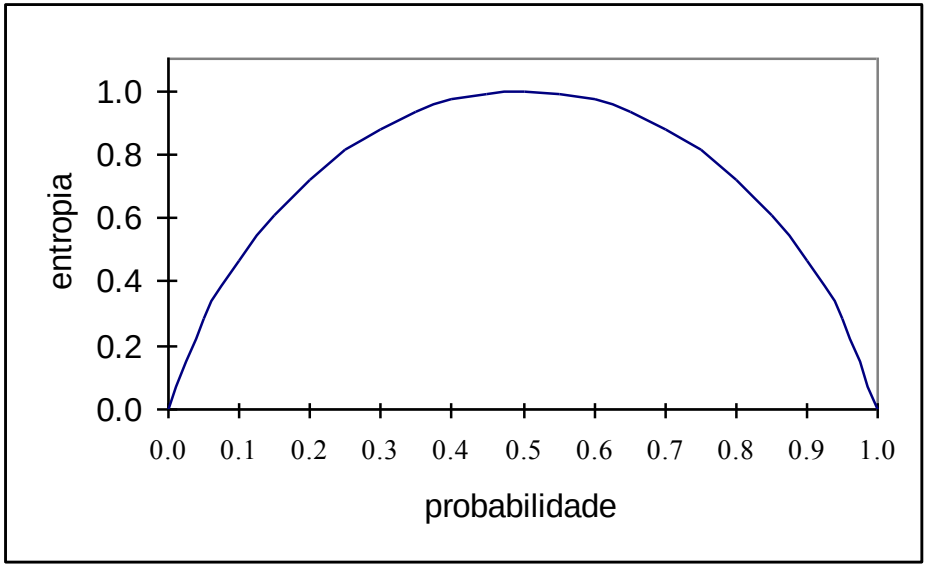
$Entropia(0\mathbf{F},4\mathbf{M}) = -(0/4)\log_2(0/4) - (4/4)\log_2(4/4)$
$= \mathbf{0}$

GI (peso <= 160) = **0.9911** – (5/9 * **0.7219** + 4/9 * **0**) = **0.5900**

- "Peso" discrimina melhor do que "Cabelo"...

Fonte: Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Problemas com duas classes equiprováveis:



- Lembrando que $\log_2 1=0$ e definindo $\log_2 0=0$;
- Para problemas em que há "c" classes temos:

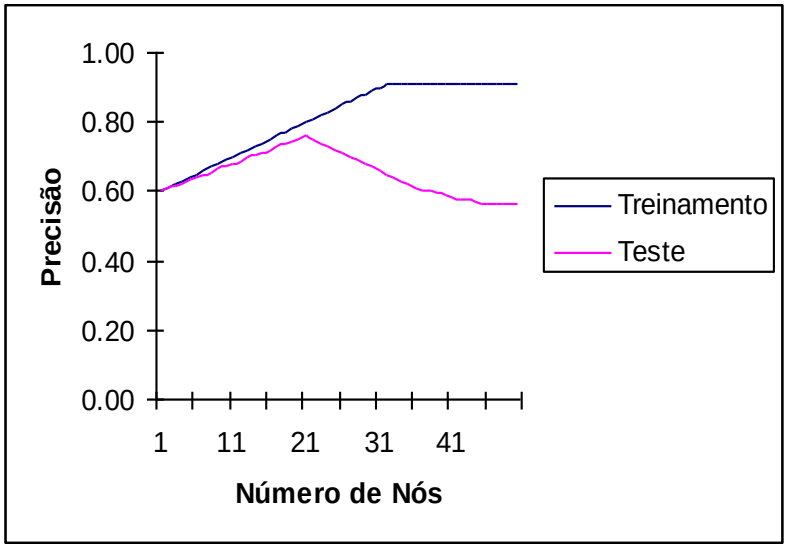
$$entropia = \sum_{i=1}^{c}(-p_i \log_2 p_i)$$

Sendo que p_i denota a probabilidade da classe "i".

Observações:

a) Compreensibilidade / facilidade para gerar regras;

b) Possibilidade de *super-ajuste* (erros, ruído, poucos dados):



Definição: Diz-se que uma hipótese $h \in$ H *super-ajusta* os dados de treinamento se existe uma hipótese alternativa $h' \in$ H tal que h apresenta um erro menor do que h' no conjunto de treinamento mas um erro maior na distribuição completa de exemplos.

c) Procedimentos de poda:

- conjunto de validação;
- eliminar antecedentes das regras obtidas a partir da árvore;

d) GI tem um *bias* (tendência, preferência) que favorece a escolha de atributos com muitos valores;

e) Para minimizar/superar limitações:

- procedimentos de poda;
- critérios de escolha de atributos alternativos;
- seleção de atributos *a priori*, etc.

Exercício:

Para responder às perguntas 1-5, considere a seguinte base de dados:

A_1	A_2	A_3	A_4	Classe
S	H	H	W	N
S	H	H	S	N
O	H	H	W	Y
R	M	H	W	Y
R	C	N	W	Y
R	C	N	S	N
O	C	N	S	Y
S	M	H	W	N
S	C	N	W	Y
R	M	N	W	Y
S	M	N	S	Y
O	M	H	S	Y
O	H	N	W	Y
R	M	H	S	N

1) Obter a árvore de classificação pelo ganho de informação;

2) Quais são as regras de classificação obtidas por meio desta árvore?

3) Qual é a acurácia deste conjunto de regras para o conjunto de treinamento? Esta acurácia é uma estimativa adequada para o classificador?

4) Supondo-se que não se pode dispor adicionalmente de mais dados, descreva um procedimento que permita estimar melhor a acurácia do classificador em questão para dados *novos* (*e.g.*, ainda não observados e que serão classificados pela árvore de decisão).

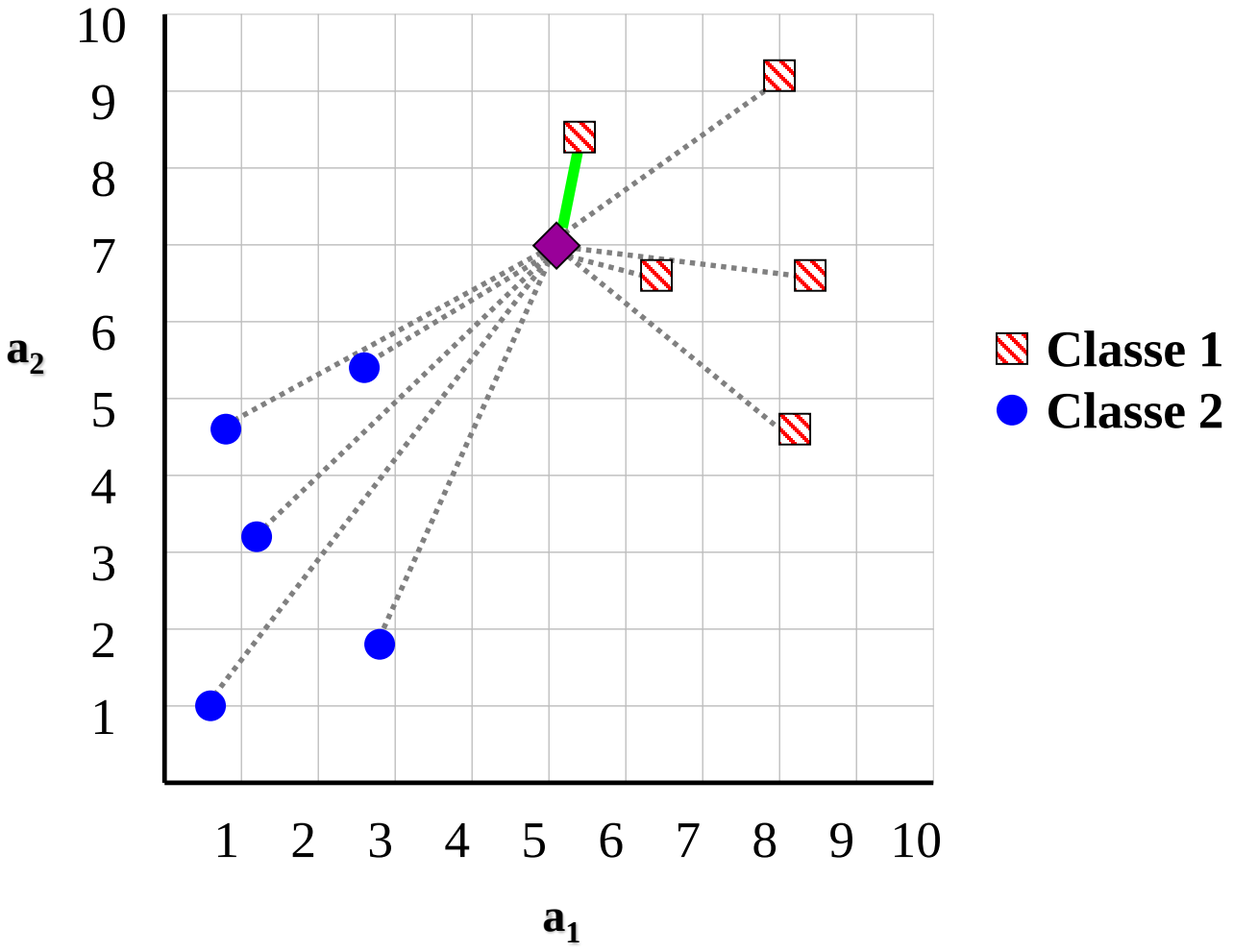
Agenda:

- Conceitos de Classificação
- Técnicas de Classificação
  - *One Rule* (1R)
  - Naive Bayes (com seleção de atributos)
  - Árvores de Decisão
  - K-Vizinhos Mais Próximos (K-NN)
- Super-ajuste e validação cruzada
- Combinação de Modelos

Aprendizado Baseado em Exemplos: *k-NN*

- Não constroem descrições gerais e explícitas (função alvo) a partir dos exemplos de treinamento;
- Generalização é *adiada* até o momento da classificação *(lazy methods);*
- Armazena-se uma base de exemplos (*instances*) que é usada para realizar a classificação de uma nova *query* (exemplo *não visto*);
- Inclui técnicas como KNN, CBR, métodos de regressão;
- Em muitos casos apresenta um alto custo computacional;
- Cálculo de distâncias pode ser problemático.

Noção Intuitiva:



Fonte: Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Método do(s) Vizinho(s) Mais Próximo(s):

- *K-Nearest Neighbors (KNN);*
- Exemplos correspondem a pontos no $\Re^n$;
- Vizinhos definidos em função de uma medida de distância;
- Por exemplo, considerando-se dois vetores $\mathbf{x}=[x_1,x_2,...,x_n]$ e $\mathbf{y}=[y_1,y_2,...,y_n]$, a distância Euclidiana entre estes é:

$$d_E(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- A distância Euclidiana é uma medida de dissimilaridade. Como obter, a partir desta, uma medida de similaridade?

Função alvo discreta:

- f: $\Re^n \rightarrow V$, $V=\{v_1, v_2,..., v_s\}$ /* s classes */
- Algoritmo:

  Dado um exemplo $\mathbf{x}_q$ a ser classificado e considerando que $\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_k$ representam os k exemplos mais próximos de $\mathbf{x}_q$, retornar:

$$f(\mathbf{x}_q) \leftarrow \underset{v \in V}{arg\,max} \sum_{i=1}^{k} \delta(v, f(\mathbf{x}_i)) \qquad \begin{cases} (a = b) \Rightarrow \delta(a,b) = 1 \\ (a \neq b) \Rightarrow \delta(a,b) = 0 \end{cases}$$

➢ **Classificação por meio da classe majoritária da vizinhança.**

Função alvo contínua:

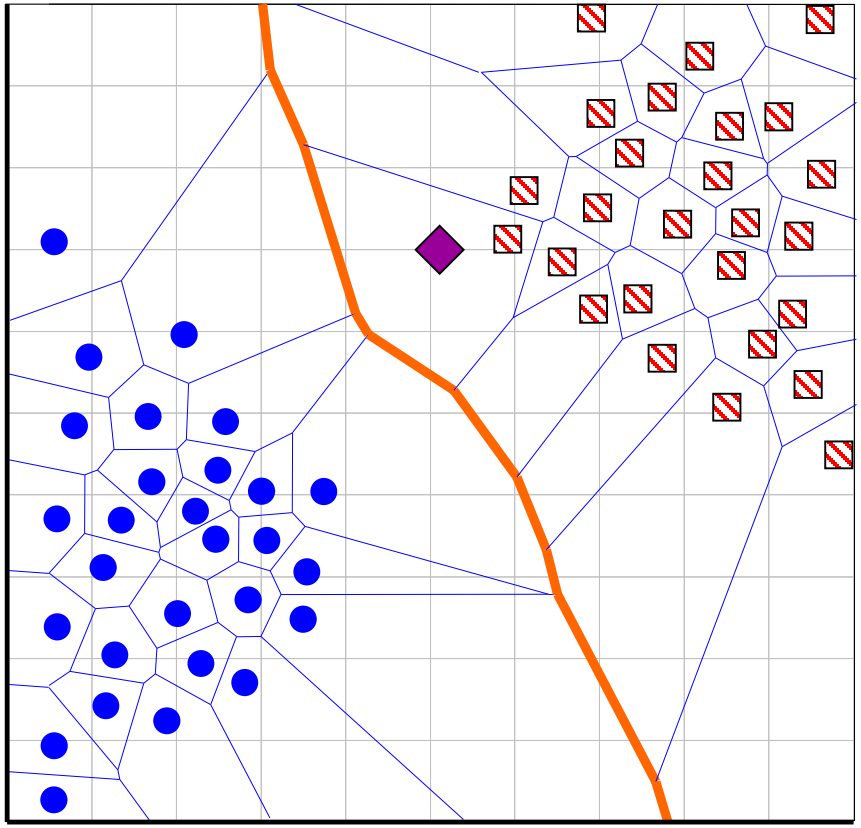
- $f: \Re^n \rightarrow \Re$
- Algoritmo:

  Dado um exemplo $\mathbf{x}_q$ cujo valor da variável dependente (y) se deseja estimar e considerando que $\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_k$ representam os k exemplos mais próximos de $\mathbf{x}_q$, retornar:

$$y = f(\mathbf{x}_q) = \frac{\sum_{i=1}^{k} f(\mathbf{x}_i)}{k}$$

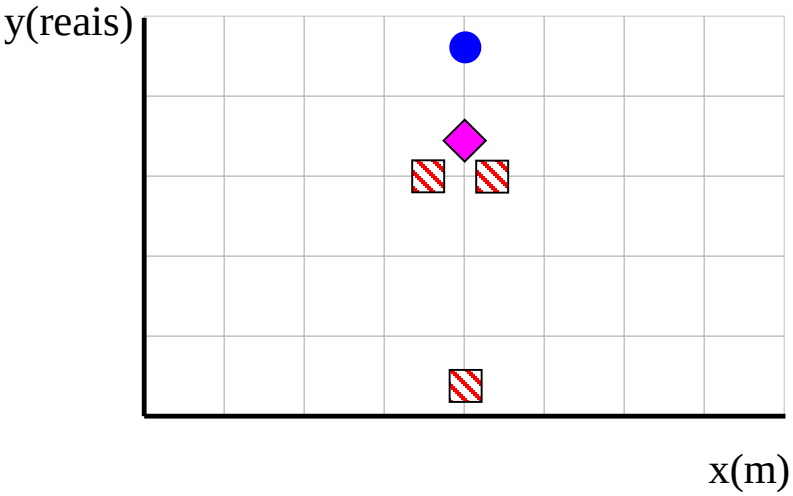
➢ **Predição por meio da média da vizinhança.**

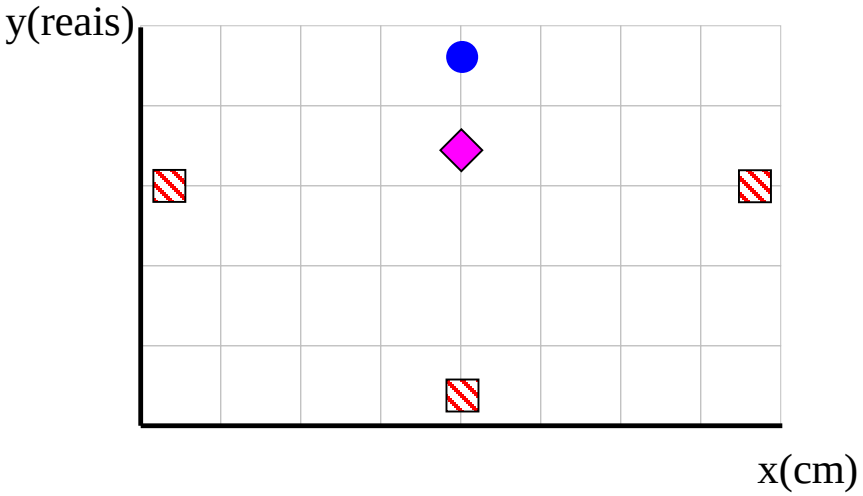
Superfície de Decisão:



Fonte: Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Sensibilidade em relação à escala:





Como diminuir este problema?

- Normalização linear;
- *Score z;*

Fonte: Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Como lidar com atributos nominais?

- Mudar a função de distância;
- Por exemplo, pode-se usar o procedimento chamado de casamento simples (*simple matching*):

$$d_{SM}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{i=n} s_i \qquad \begin{cases} (x_i = y_i) \Rightarrow s_i = 0; \\ (x_i \neq y_i) \Rightarrow s_i = 1; \end{cases}$$

- Várias outras medidas de distância propostas na literatura (e.g., ver Kaufman & Rousseeuw, Finding Groups in Data, 1990);
- Como lidar com bases de dados formadas por diferentes tipos de atributos (ordinais, contínuos, nominais, binários)?

Exemplo:

Instância	a_1	a_2	a_3	Classe
1	0	250	36	A
2	10	150	34	B
3	2	90	10	A
4	6	78	8	B
5	4	20	1	A
6	1	170	70	B
7	8	160	41	A
8	10	180	38	B
9	6	200	45	?

Perguntas:

a) Qual é a função de distância a ser empregada?

b) Classificar o objeto #9 considerando k=1,2,3,4,5.

c) Como escolher k ?

d) Qual problema a função de distância empregada apresenta? Como corrigi-lo?

k-NN ponderado por uma função de distância:

Função alvo discreta:

$$f(\mathbf{x}_q) \leftarrow \underset{v \in V}{arg\,max} \sum_{i=1}^{k} w_i \delta(v, f(\mathbf{x}_i)) \qquad \begin{cases} (a = b) \Rightarrow \delta(a,b) = 1 \\ (a \neq b) \Rightarrow \delta(a,b) = 0 \end{cases}$$

Função alvo contínua:

$$y = f(\mathbf{x}_q) = \frac{\sum_{i=1}^{k} w_i f(\mathbf{x}_i)}{\sum_{i=1}^{k} w_i}$$

Ponderação: $w_i = \frac{1}{d(\mathbf{x}_q, \mathbf{x}_i)^2}$

Alguns problemas do *k*-NN:

- Custo computacional:
  - Métodos globais X métodos locais;
  - Sumarização via protótipos (*clustering*).

- O que pode acontecer se, em 20 atributos, somente dois são relevantes pra classificação?
  - Maldição da dimensionalidade (*curse of dimensionality*);
  - Ponderação e/ou seleção de atributos.

Exercícios:

1. Considerando-se a base de dados previamente apresentada (Slide 28) responda à pergunta b) para um *k*-NN ponderado pela distância.

2. Suponha que naquela mesma base de dados seja inserido um atributo nominal a_4 cujos valores para os registros 1,...,9 sejam respectivamente {M,F,C,M,F,C,M,F,C}. Proponha uma medida de distância e, com base nela, classifique o registro 9 para *k*=1.

Agenda:

- Conceitos de Classificação
- Técnicas de Classificação
  - *One Rule* (1R)
  - Naive Bayes (com seleção de atributos)
  - Árvores de Decisão
  - K-Vizinhos Mais Próximos (K-NN)
- Super-ajuste e validação cruzada
- Combinação de Modelos