Aprendizado de Máquina

Introdução

Thiago A. S. Pardo
Daniel Honorato
Solange O. Rezende
Ronaldo C. Prati

1

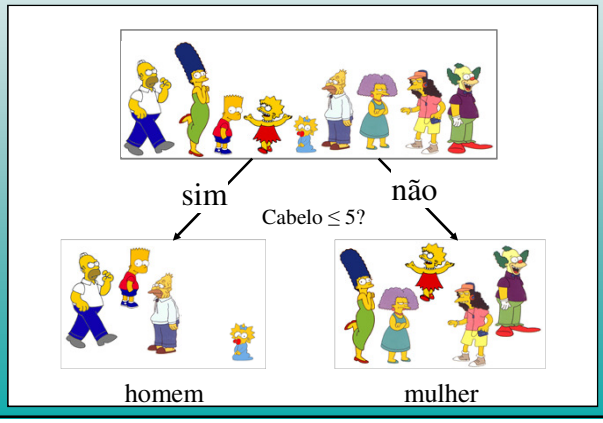
Relembrando: Simpsons

Pessoa	Comprimento do Cabelo	Peso	Idade	Classe: Sexo
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

2

Pergunta

- Qual o erro e a acurácia da hipótese abaixo?



3

Erro e Precisão

- Regressão: distância entre valor real e predito
  - Duas medidas usualmente utilizadas
    - mse: *mean squared error*
    - mad: *mean absolute distance*

$$\text{mse - err}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h(x_i))^2$$

$$\text{mad - err}(h) = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - h(x_i)\right|$$

4

Erro e Precisão

- **Erro majoritário**
  - Erro pelo palpite da classe mais freqüente
  - Limiar máximo abaixo do qual o erro do classificador deve ficar

5

Erro majoritário = 14-9/14 = 5/14 = 35%

Dia	Tempo	Temperatura	Umidade	Vento	Jogou tênis?
1	Sol	Quente	Alta	Fraco	Não
2	Sol	Quente	Alta	Forte	Não
3	Nublado	Quente	Alta	Fraco	**Sim**
4	Chuva	Mediana	Alta	Fraco	**Sim**
5	Chuva	Frio	Normal	Fraco	**Sim**
6	Chuva	Frio	Normal	Forte	Não
7	Nublado	Frio	Normal	Forte	**Sim**
8	Sol	Mediana	Alta	Fraco	Não
9	Sol	Frio	Normal	Fraco	**Sim**
10	Chuva	Mediana	Normal	Fraco	**Sim**
11	Sol	Mediana	Normal	Forte	**Sim**
12	Nublado	Mediana	Alta	Forte	**Sim**
13	Nublado	Quente	Normal	Fraco	**Sim**
14	Chuva	Mediana	Alta	Forte	Não

6

Erro majoritário?

Pessoa	Comprimento do Cabelo	Peso	Idade	Classe: Sexo
 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**

7

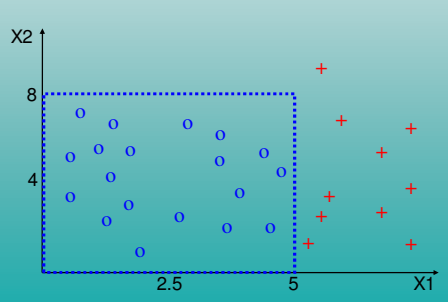
Espaço de Descrição

- *m* atributos podem ser vistos como um vetor
- Cada atributo corresponde a uma coordenada num espaço m-dimensional denominado espaço de descrição
- Cada ponto no espaço de descrição pode ser rotulado com a classe associada aos atributos

8

Espaço de Descrição

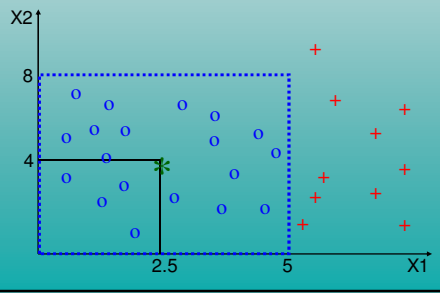
- Um indutor divide o espaço de descrição em regiões
- Cada região é rotulada com uma classe
- Exemplo: para 2 atributos X1 e X2, *if* X1 < 5 *and* X2 < 8 *then* classe **o** *else* classe **+,** divide o espaço em duas regiões



9

Espaço de Descrição

- Para classificar um novo exemplo com (X1,X2) = (2.5, 4), basta verificar em qual região ela se localiza e atribuir a classe associada àquela região (neste caso, classe **o**)



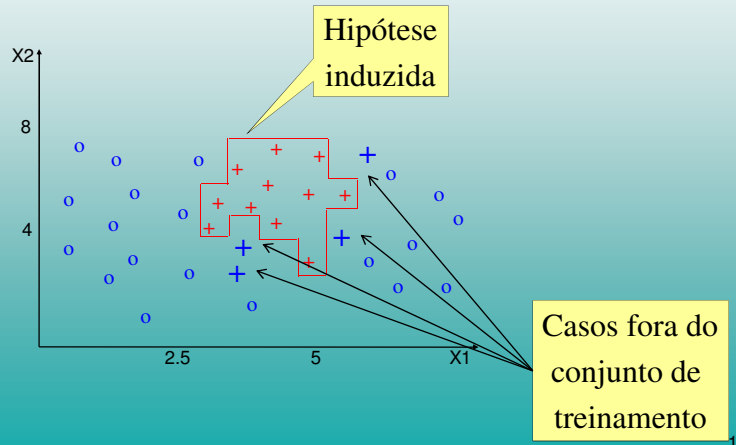
10

Overfitting

- Ocorre quando a hipótese extraída a partir dos dados é muito específica para o conjunto de treinamento
  - A hipótese apresenta uma boa performance para o conjunto de treinamento, mas uma performance ruim para os casos fora desse conjunto

11

Overfitting - Exemplo



12

Underfitting

- A hipótese induzida apresenta um desempenho ruim tanto no conjunto de treinamento como de teste. Por quê ?

13

Underfitting

- A hipótese induzida apresenta um desempenho ruim tanto no conjunto de treinamento como de teste. Por quê ?
  - poucas exemplos representativos foram dadas ao sistema de aprendizado
  - o usuário pré-definiu um tamanho muito pequeno para o classificador (por exemplo, o usuário definiu um alto valor de poda para árvores de decisão)

14

Overtuning

- Ajuste excessivo do algoritmo de aprendizado
  - Causa problemas similares ao overfitting

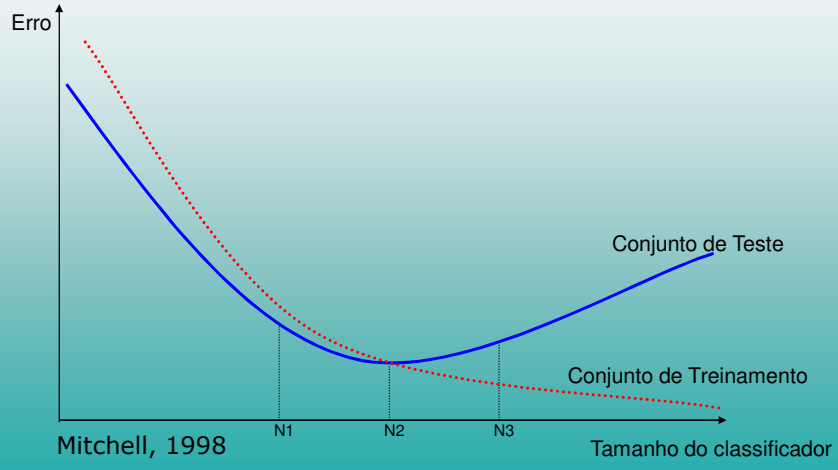
15

Poda

- Técnica para lidar com ruído, overfitting e overtuning
  - Generalização das hipóteses aprendida pelo corte (“poda”) de parte das hipóteses

16

Relação entre o tamanho do classificador e o erro



Mitchell, 1998

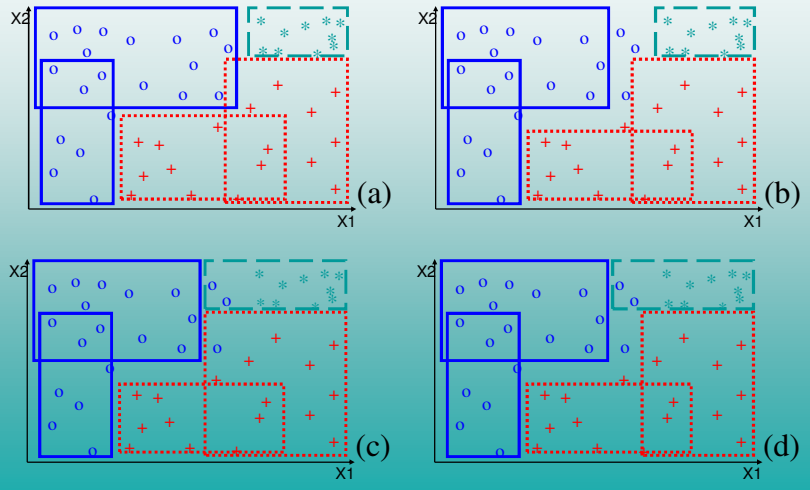
17

Consistência e Completude

- Depois de induzida, uma hipótese pode ser avaliada em relação aos critérios
  - **Consistência**: se classifica corretamente todos os exemplos
  - **Completude**: se classifica todos os exemplos

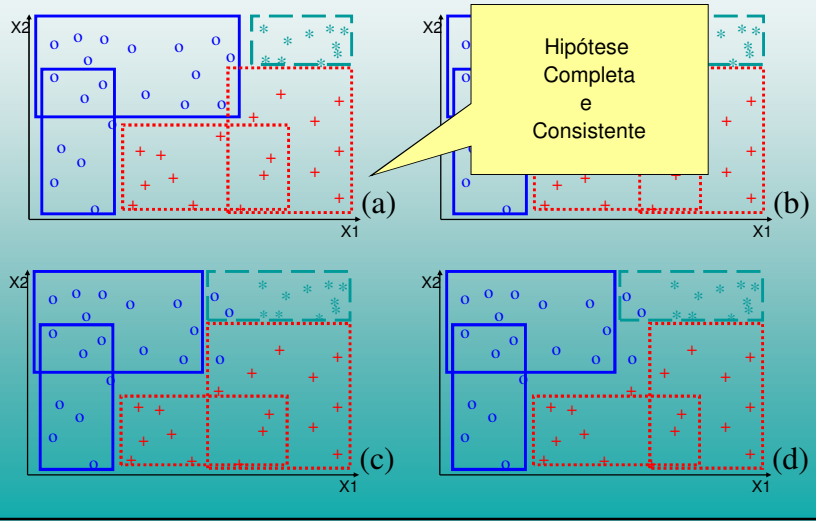
18

Relação entre Completude e Consistência



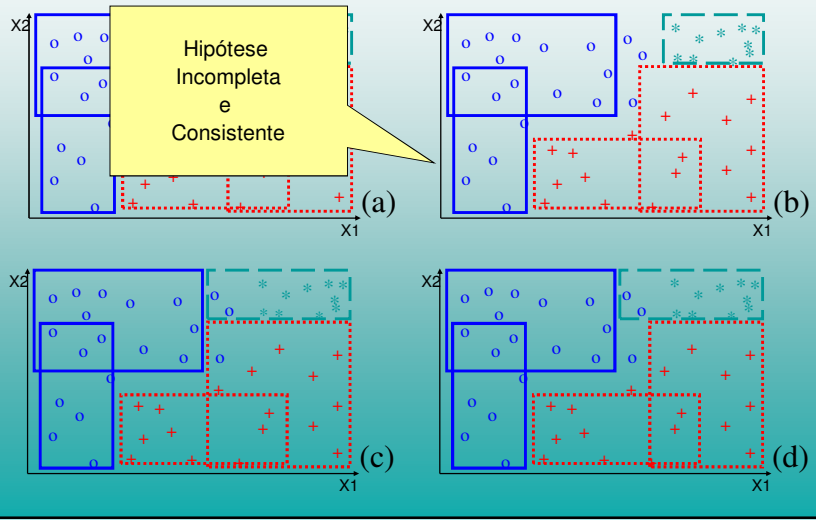
19

Relação entre Completude e Consistência



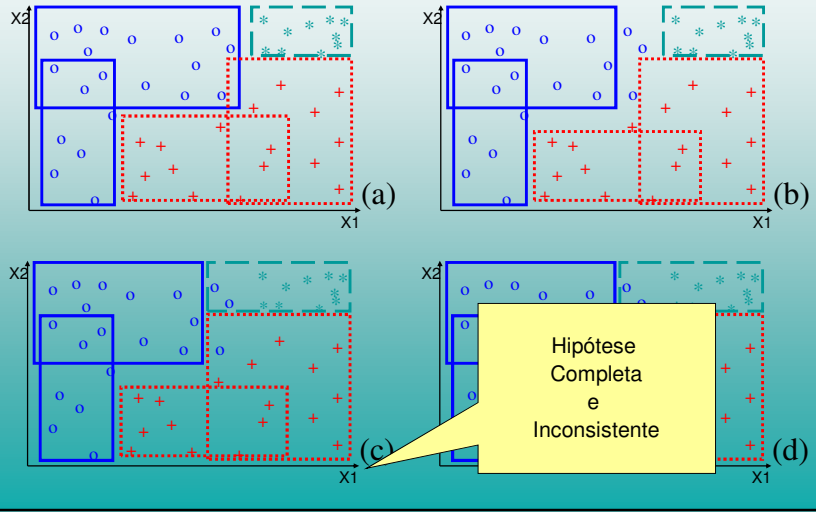
20

Relação entre Completude e Consistência



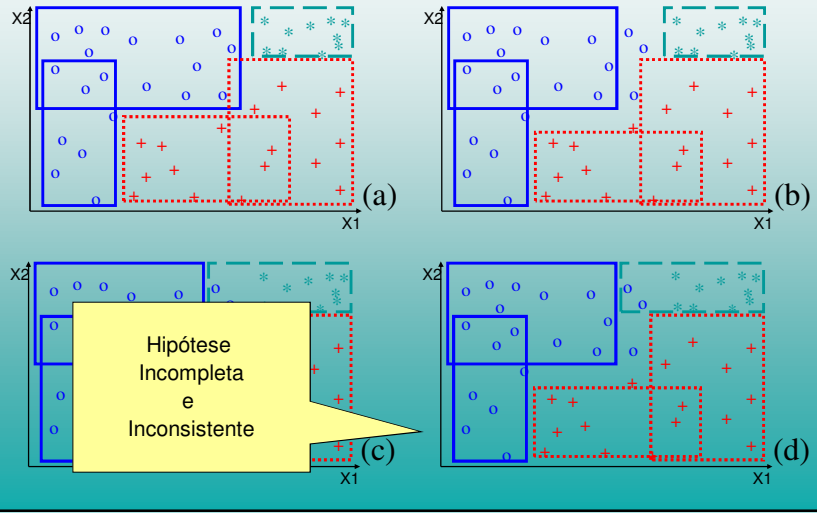
21

Relação entre Completude e Consistência



22

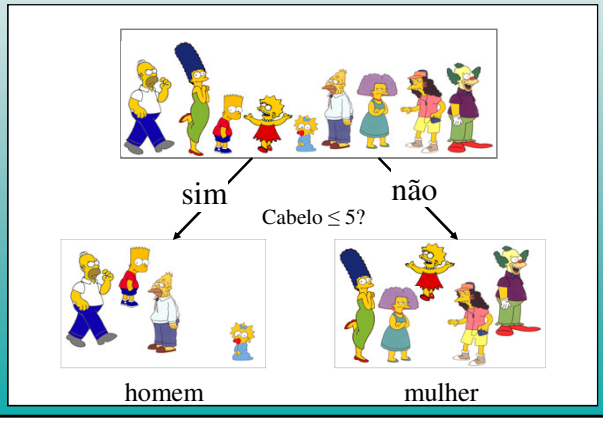
Relação entre Completude e Consistência



23

Pergunta

- Como classificar a hipótese abaixo?



24

Matriz de Confusão

- Oferece uma medida da eficácia do modelo de classificação, mostrando o número de classificações corretas *versus* o número de classificação prevista para cada classe

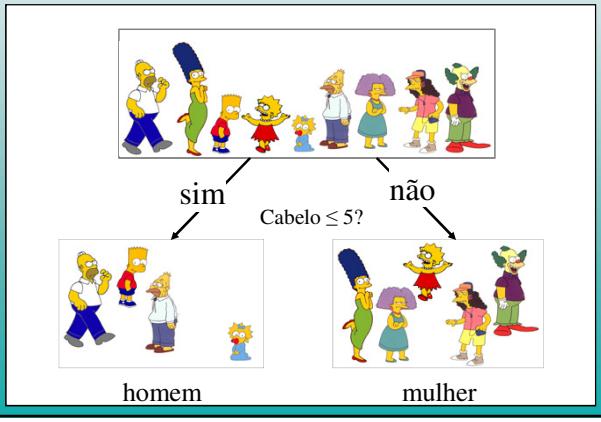
Classe	prevista C_1	prevista C_2	$\cdots$	prevista C_k
real C_1	$M(C_1,C_1)$	$M(C_1,C_2)$	$\cdots$	$M(C_1,C_k)$
real C_2	$M(C_2,C_1)$	$M(C_2,C_2)$	$\cdots$	$M(C_2,C_k)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
real C_k	$M(C_k,C_1)$	$M(C_k,C_2)$	$\cdots$	$M(C_k,C_k)$

$$M(C_i,C_j) = \sum_{\{\forall(x,y)\in T: y=C_i\}} \|h(x)=C_j\|$$

25

Exercício

- Faça a matriz de confusão



26

Matriz de Confusão para 2 Classes

Classe	prevista C_+	prevista C_-	Taxa de erro da classe	Taxa de erro total
real C_+	T_P	F_N	$\frac{F_N}{T_P + F_N}$	$\frac{F_P + F_N}{n}$
real C_-	F_P	T_N	$\frac{F_P}{F_P + T_N}$	

T_P = True Positive (verdadeiro positivo)
F_N = False Negative (falso negativo)
F_P = False Positive (falso positivo)
T_N = True Negative (verdadeiro negativo)
$n = (T_P+F_N+F_P+T_N)$

27

Avaliação do classificador

- Para se estimar o erro verdadeiro de um classificador, a **amostra** para teste deve ser aleatoriamente escolhida
- Amostras não devem ser pré-selecionadas de nenhuma maneira
- Para problemas reais, tem-se uma amostra de uma única população, de tamanho *n,* e a tarefa é estimar o erro verdadeiro para essa população

28

Métodos para estimar o erro verdadeiro de um classificador

- Resubstitution
- Holdout
- Random
- r-fold cross-validation
- r-fold stratified cross-validation
- Leave-one-out

29

Resubstitution

- Gera o classificador e testa a sua performance com o **mesmo conjunto** de dados
  - Os desempenhos computados com este método são otimistas e tem grande bias
  - Desde que o bias da resubstitution foi descoberto, os métodos de cross-validation são usados

30

Holdout (Validação simples)

- Divide os dados em uma porcentagem fixa *p* para treinamento e *(1-p)* para teste
  - Geralmente ***p*=2/3** e ***(1-p)*=1/3**
  - Para que os resultados não dependam da divisão dos dados (exemplos), pode-se calcular a média de vários resultados de holdout

31

Random

- **I classificadores, I<<n**, são induzidos de cada conjunto de treinamento
- O erro é a média dos erros dos classificadores medidos por conjuntos de treinamentos gerados aleatória e independentemente
- Pode produzir estimativas melhores que o holdout

32

r-fold cross-validation

- Os exemplos são aleatoriamente divididos em *r* **partições** (folds) de tamanho aproximadamente igual (*n/r*)
- Os exemplos de (*r-1*) folds são independentemente usados no treinamento e os classificadores obtidos são testados com o fold remanescente
- O processo é repetido *r* vezes, e a cada repetição um fold diferente é usado para tese. O erro do cross-validation é a média dos erros dos *r* folds

33

r-fold stratified cross-validation

- É similar ao cross-validation, mas no processo de geração dos folds a **distribuição das classes** no conjunto de exemplos é levada em consideração durante à amostragem
- Por exemplo, se o conjunto de exemplos tiver duas classes com uma distribuição de 80% para uma classe e 20% para outra, cada fold também terá essa proporção

34

Leave-one-out

- Para um exemplo de tamanho *n*, um classificador é gerado usando ***n-1*** **exemplos**, e **testado no exemplo remanescente**
- O processo é repetido *n* vezes, utilizando cada um dos *n* exemplos para teste. O erro é a soma dos erros dos testes para cada exemplo divido por *n*
- Caso especial de cross-validation
- Computacionalmente caro e usado apenas quando o conjunto de exemplos é pequeno

35

Avaliando Classificadores

- **Não há um único bom algoritmo** de AM para todas as tarefas
- É importante conhecer o poder a as limitações de indutores diferentes
- Na prática, devemos testar algoritmos diferentes, estimar sua precisão e escolher entre os algoritmos aquele que apresentar maior precisão, por exemplo, para um domínio específico

36

Metodologia de Avaliação (Russel e Norvig, 2003)

1. Coletar um conjunto de exemplos, de preferência sem "ruído"
2. Dividir randomicamente o conjunto de exemplos em um conjunto de teste e um conjunto de treinamento.
3. Aplicar um ou mais indutores ao conjunto de treinamento, obtendo uma hipótese ***h*** para cada indutor
4. Medir a performance dos classificadores com o conjunto de teste
5. Estudar a eficiência e robustez de cada indutor, repetindo os passos 2 a 4 para diferentes conjuntos e tamanhos do conjunto de treinamento
6. Se estiver propondo um ajuste ao indutor, voltar ao passo 1

37

Calculando Média e Desvio Padrão usando Amostragem

Usando cross-validation: dado um algoritmo A, para cada fold *i*, calculamos o erro $err(h_i)$, i = 1, 2, ..., r, temos:

$$média(A) = \frac{1}{r}\sum_{i=1}^{r} err(h_i)$$

$$\text{variância} = \frac{1}{r}\left[\frac{1}{r-1}\sum_{i=1}^{r}\left(err(h_i) - média(A)\right)^2\right]$$

$$desvio\ \text{padrão} = \sqrt{\text{variância}(A)}$$

38

Calculando Média e Desvio Padrão usando Amostragem

- Exemplo: Considerando um exemplo de cross-validation 10-fold (r=10), para um algoritmo A que apresente os erros 5.5, 11.4, 12.7, 5.2, 5.9, 11.30, 10.9, 11.2, 4.9 e 11.0, então:

$$média(A) = \frac{90.0}{10} = 9.0$$

$$desvio\ padrão = \sqrt{\frac{1}{10(9)}90.3} = 1.0$$

39

Comparando dois Algoritmos

$A_S \Rightarrow$ algoritmo padrão

$Ap \Rightarrow$ algoritmo proposto

$$média(A_S - A_P) = média(A_S) - média(A_P)$$

$$desvio\ padrão(A_S - A_P) = \sqrt{\frac{dp(A_S)^2 + dp(A_p)^2}{2}}$$

$$diferença\ absoluta(A_S - A_P) = \frac{média(A_S - A_P)}{dp(A_S - A_P)}$$

40

Comparando dois Algoritmos

- Se da(A_S-A_P) > 0, A_P tem melhor performance que A_S
- Se da(A_S-A_P) >= 2, A_P tem melhor performance que A_S com um nível de confiança de 95%.
- Se da(A_S-A_P) <= 0, A_s tem melhor performance que A_p
- Se da(A_S-A_P) <= -2, A_s tem melhor performance que A_p com um nível de confiança de 95%.

41

Comparando dois Algoritmos

- Exemplo: considerando que $A_S = 9.00 \pm 1.00$ e $A_P = 7.50 \pm 0.80$

$$média(A_s - A_P) = 9.00 - 7.50 = 1.50$$

$$dp(A_S - A_P) = \sqrt{\frac{1.00^2 + 0.80^2}{2}} = 0.91$$

$$da(A_S - A_P) = \frac{1.50}{0.91} = 1.65$$

Como da(A_S-A_P) < 2, A_P não tem uma performance significativamente melhor que A_S, com um nível de confiança de 95%.

42

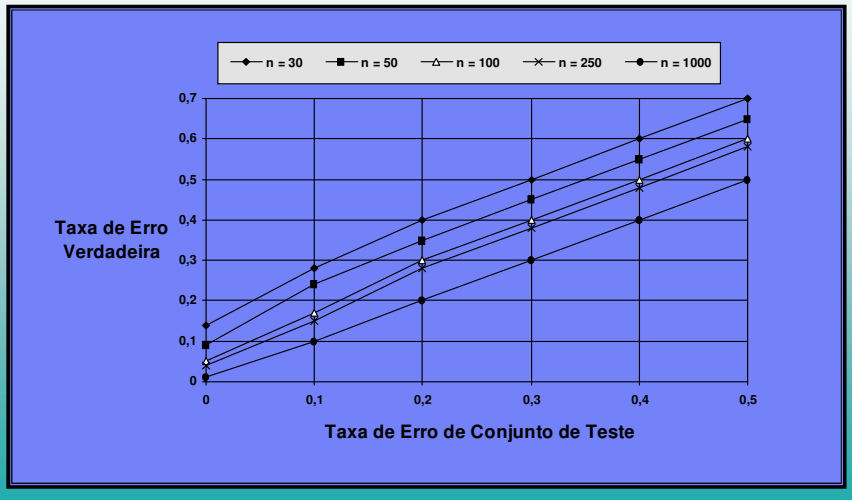
Métodos de Treinar-e-Testar

"Quantos casos de teste são necessários para uma estimativa precisa?"

"Quantos casos deve conter cada conjunto de treinamento e teste?"

43

Número de Casos de Teste e Qualidade da Predição



44

Número de Casos de Teste e Qualidade da Predição (Cont.)

- Quando o tamanho do conjunto de teste atinge 1000 casos, a estimativa já é bastante precisa
- Com 5000 casos, a taxa de erro do conjunto de teste é virtualmente idêntica à taxa de erro verdadeira

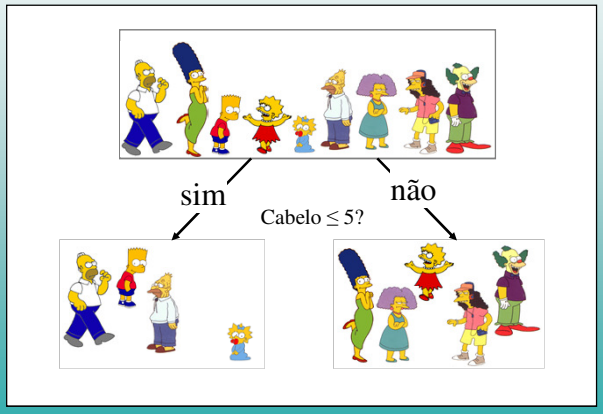
45

Classificação dos Simpsons

- Hipóteses?

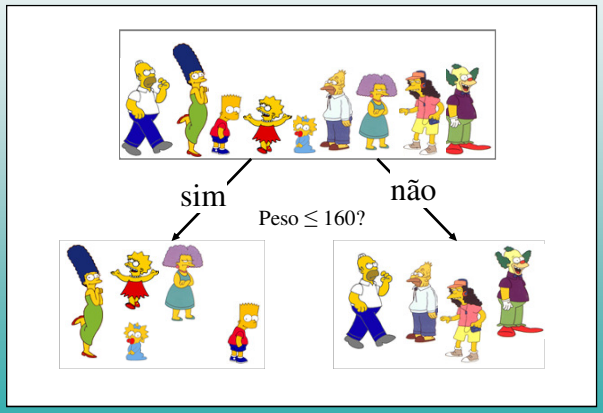
46

Classificação dos Simpsons
Possíveis indutores



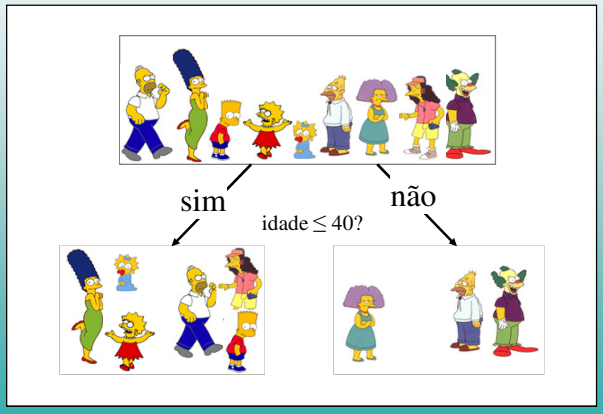
47

Classificação dos Simpsons
Possíveis indutores



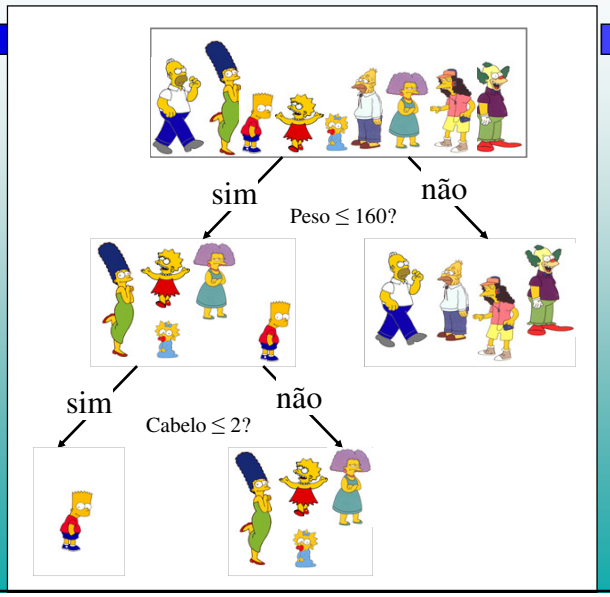
48

Classificação dos Simpsons
Possíveis indutores



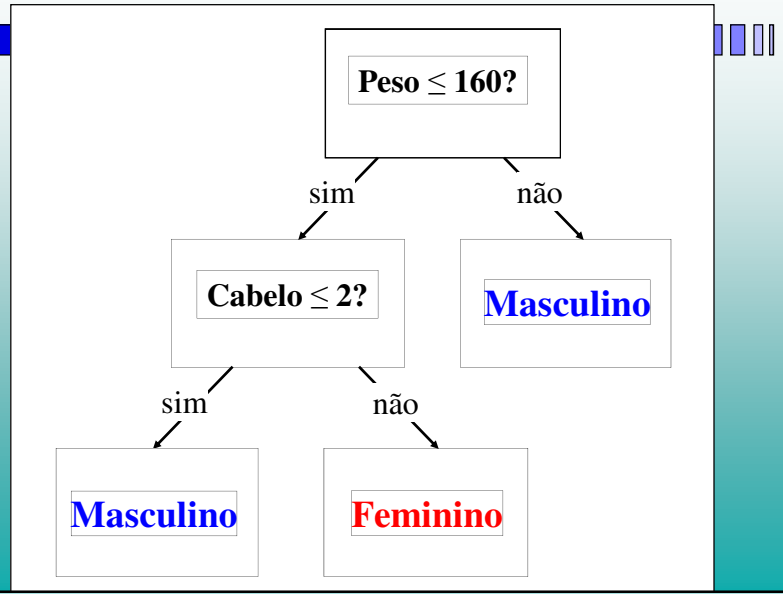
49

Classificação dos Simpsons
Possíveis indutores



50

Mais genericamente...



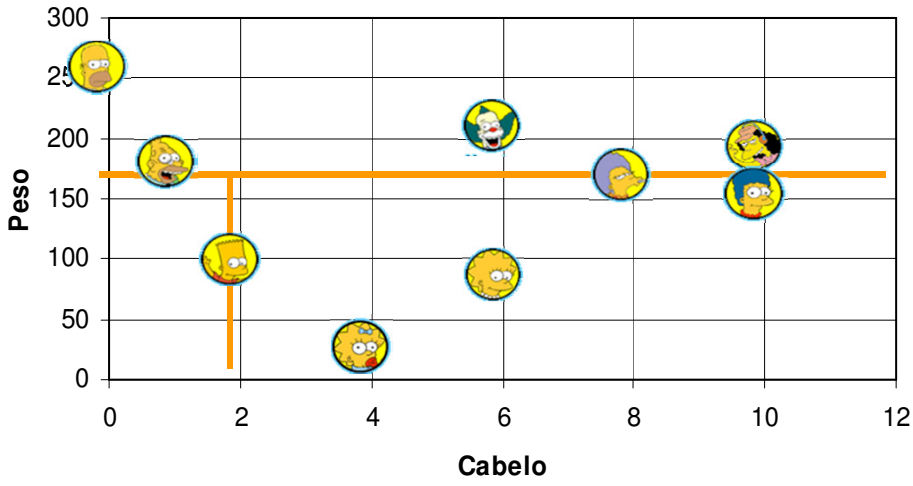
51

Ou...

Se *PESO* ≤ *160* então
    Se *CABELO* ≤ *2* então
        MASCULINO
    Senão
        FEMININO
Senão
    MASCULINO

52

Interpretação Geométrica



Exercício em duplas: hipótese, matriz de confusão (por *resubstitution*), taxa de acerto por classe, acurácia, erro majoritário, etc.

Personagem	Arma	Transformação	Idade	Classe
He-man	Lâmina	Sim	Adulto	Herói
Seiya	Magia	Não	Novo	Herói
Mun-ra	Magia	Sim	Velho	Vilão
Bob Esponja	Não usa	Não	Novo	Herói
Magneto	Não usa	Não	Adulto	Vilão
Voldemort	Magia	Não	Adulto	Vilão
Wolverine	Lâmina	Não	Velho	Herói
Lex Luthor	Não usa	Não	Velho	Vilão
T. Cristina	Lâmina	Sim	Adulto	Vilão
Superman	Não usa	Sim	Adulto	Herói
Ben 10	Não usa	Sim	Novo	Herói