Aprendizado de Máquina

Sistemas de Informação Inteligente

Prof. Leandro C. Fernandes

Autores do material:
Thiago A. S. Pardo
Solange O. Rezende

SISTEMAS DE APRENDIZADO DE MÁQUINA INDUTIVO

Sistemas de AM Indutivo

- O paradigma de aprendizado indutivo busca **aprender conceitos através de instâncias** destes conceitos.

Sistemas de AM Indutivo

- O classificador utiliza os conceitos aprendidos para **classificar novos exemplos.**

Exercício de Classificação

- Atributos:
  - Comprimento do cabelo
  - Peso
  - Idade
  - Atributo de Classe: Sexo

- Conjunto de exemplos
  - Veja tabela a seguir ...

Induza a hipótese em 10 min!

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
Cmic	8	290	38	?

Preparação de Dados

- Fase que antecede o processo de aprendizagem, para facilitar ou melhorar o processo.

- Exemplos
  - Remover exemplos incorretos
  - Transformar o formato dos dados para que possam ser usados com um determinado indutor
  - Selecionar atributos relevantes (Seleção de Atributos)

Ruído

- Dados imperfeitos que podem ser derivados do processo de aquisição, transformação ou rotulação das classes.
- Exemplos com os mesmos atributos, mas com classes diferentes ...

X1	X2	X3	X4	Y
Overcast	19	65	yes	dont_go
Rain	19	70	yes	dont_go
Rain	23	80	yes	dont_go
Sunny	23	95	no	dont_go
Sunny	28	91	yes	dont_go
Sunny	30	85	no	dont_go
Overcast	19	65	yes	go
Rain	21	80	no	go
Rain	22	95	no	go
Sunny	22	70	no	go
Overcast	23	90	yes	go
Rain	25	81	no	go
Sunny	25	72	yes	go
Overcast	26	75	no	go
Overcast	29	78	no	Go

Classificador

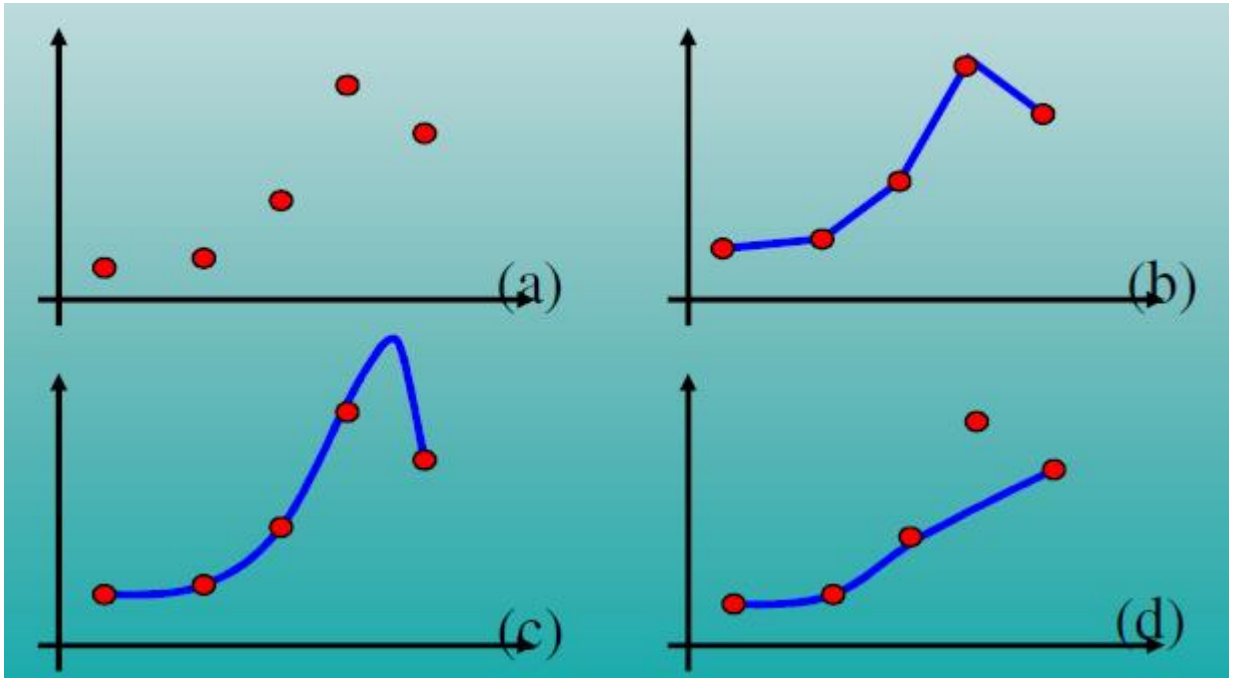
- Um exemplo pode ser representado pelo par:

$$(x, y) = (x, f(x))$$

- Onde:
  - x é a entrada;
  - $f(x)$ é a saída (f desconhecida!)
  - Indução ou inferência indutiva: dada uma coleção de exemplos de f, retornar uma **função *h* que aproxima *f***
  - h é denominada uma **hipótese** sobre f

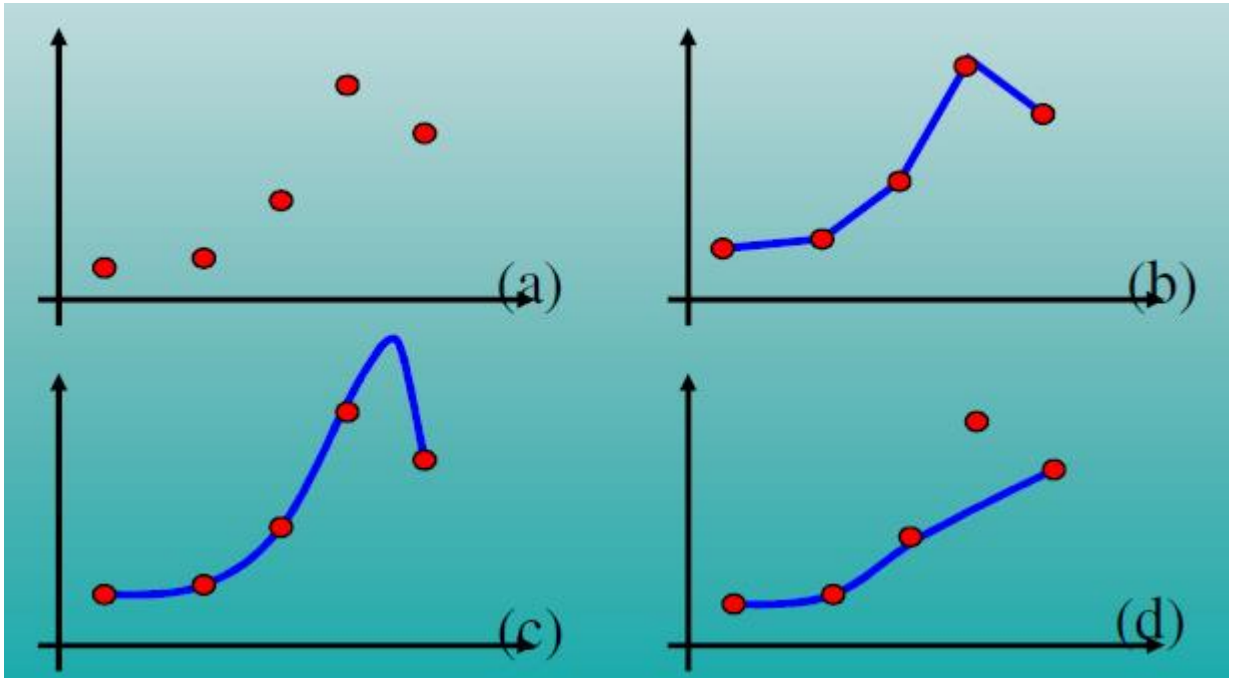
Exemplo de Hipóteses

- (a) dados originais
- (b), (c), (d) possíveis hipóteses



Bias

- Qualquer critério de **preferência** de uma hipótese sobre outra (além da consistência com os dados)



Classificação e Regressão

- Qual a diferença entre esses dois conceitos?
- Em problemas de Regressão a variável de saída *y* assume valores contínuos, enquanto que em problemas de classificação *y* é estritamente categórica.

Hierarquia de Aprendizado

Sistemas de Aprendizado de Máquina

Modo de Aprendizado	Paradigmas de Aprendizado	Linguagens de Descrição	Formas de Aprendizado
- Supervisionado - Não-supervisionado	- Simbólico - Estatístico - Baseado em exemplos - Conexionista - Evolutivo	- Instâncias ou Exemplos - Conceitos Aprendidos ou Hipóteses - Teoria de Domínio ou Conhecimento de Fundo	- Incremental - Não-Incremental

Como construir o melhor Classificador

- Qual o algoritmo (A_i) para se construir o melhor Classificador (C_i)?
- **Estudos experimentais** são necessários, uma vez que não existe uma análise matemática que possa determinar se um algoritmo de aprendizado irá desempenhar bem em um conjunto de exemplos.

Erro e Precisão

- Relembrando a notação que foi adotada:
  - Exemplo (x, y) = (x, f(x))
  - Atributos: x
  - Classe (rotulada): y = f(x)
  - Classe (classificada): h(x)
  - *n* é o número de exemplos

Erro e Precisão

- Classificação

$$err(h) = \frac{1}{n}\sum_{i=1}^{n} \| y_i \neq h(x_i) \| \quad \text{(erro)}$$

$$acc(h) = 1 - err(h) \quad \text{(precisão)}$$

- O operador $\|E\|$ retorna:
  - 1 se *E* é verdadeiro
  - 0 se *E* é falso

Erro e Precisão

- Regressão: Distância entre valor real e predito.
- Duas medidas usualmente utilizadas
  - mse: *mean squared error*

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

  - mad: *mean absolute distance*

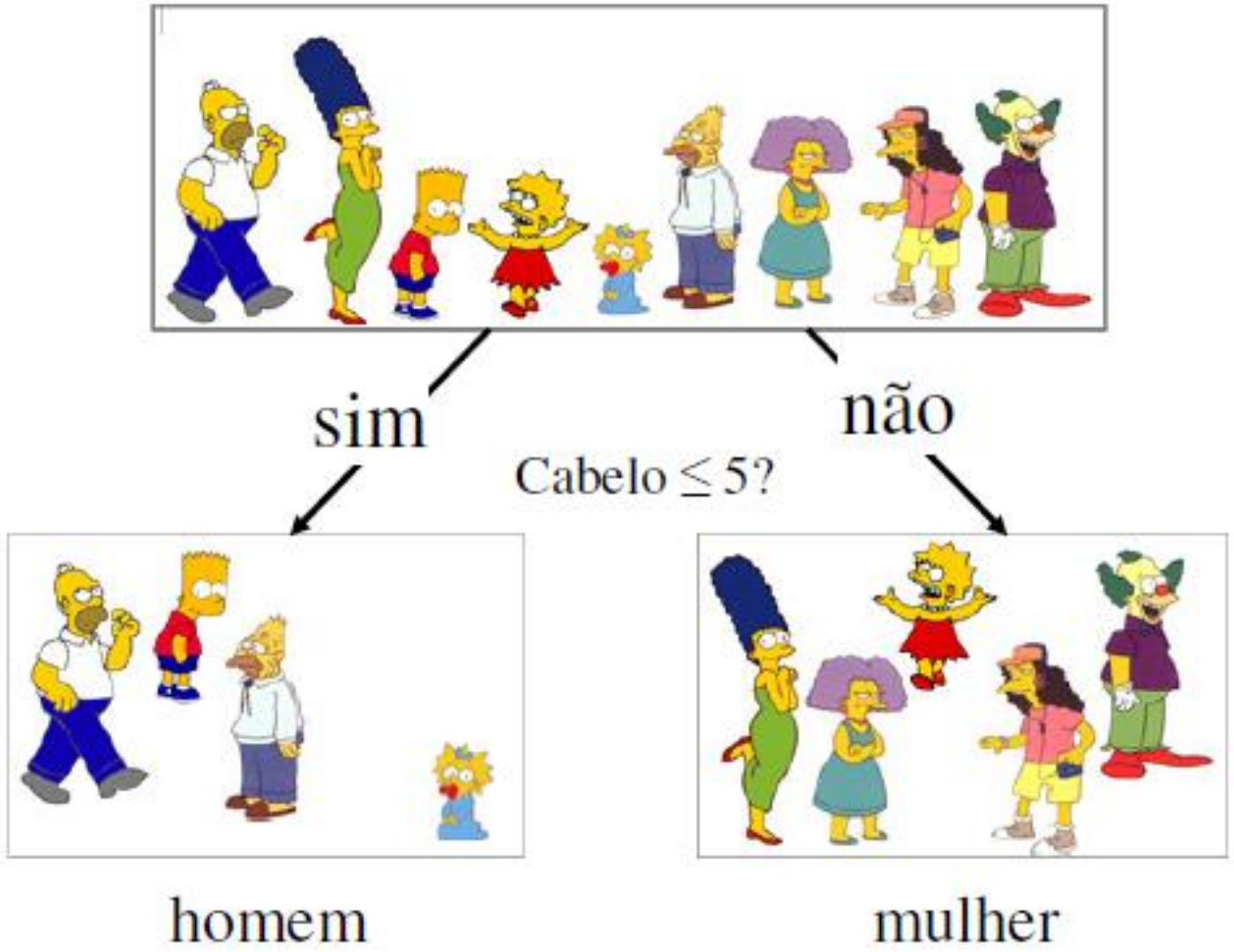
$$err(h) = \frac{1}{n}\sum_{i=1}^{n}|y_i - h(x_i)|$$

Pergunta

- Qual o erro de sua(s) hipótese(s)?

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

Qual o erro e a acurácia da hipótese?



Erro e Precisão

Erro majoritário:

- Erro pelo palpite da classe mais frequente.
- Limiar máximo abaixo do qual o erro do classificador deve ficar

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

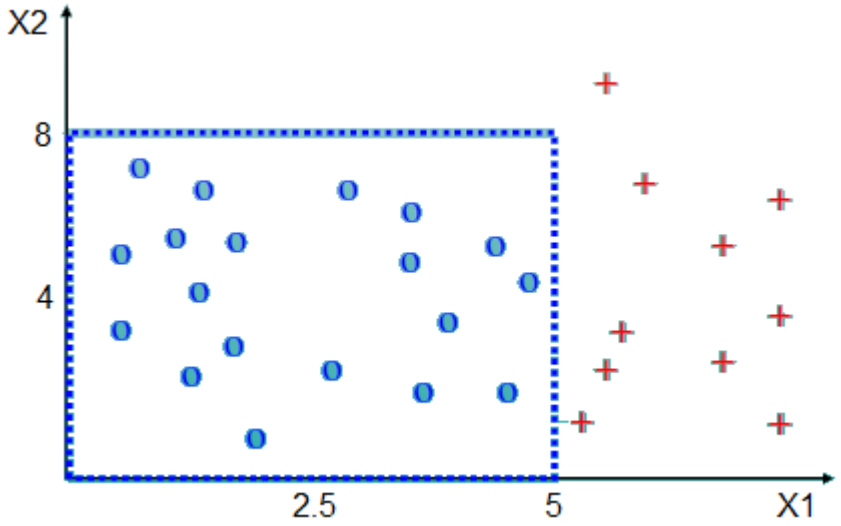
Qual o erro majoritário?

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

ESPAÇO DE DESCRIÇÃO

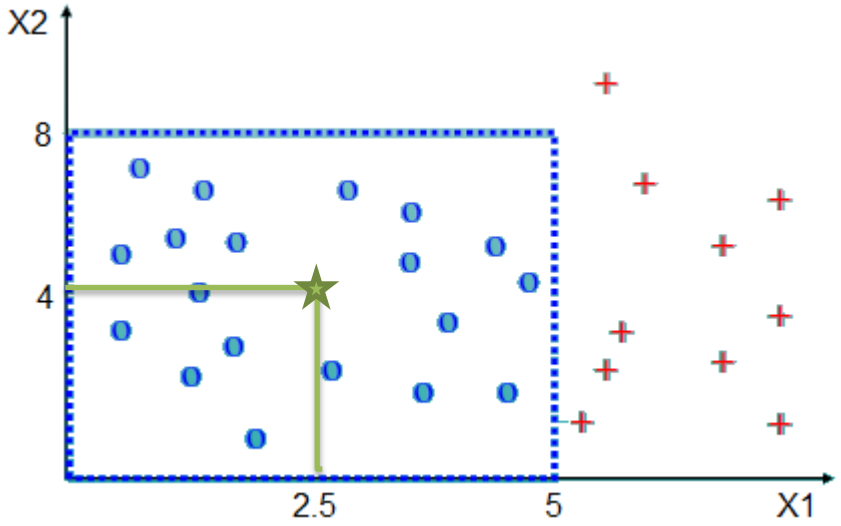
Espaço de Descrição

- *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.



Espaço de Descrição

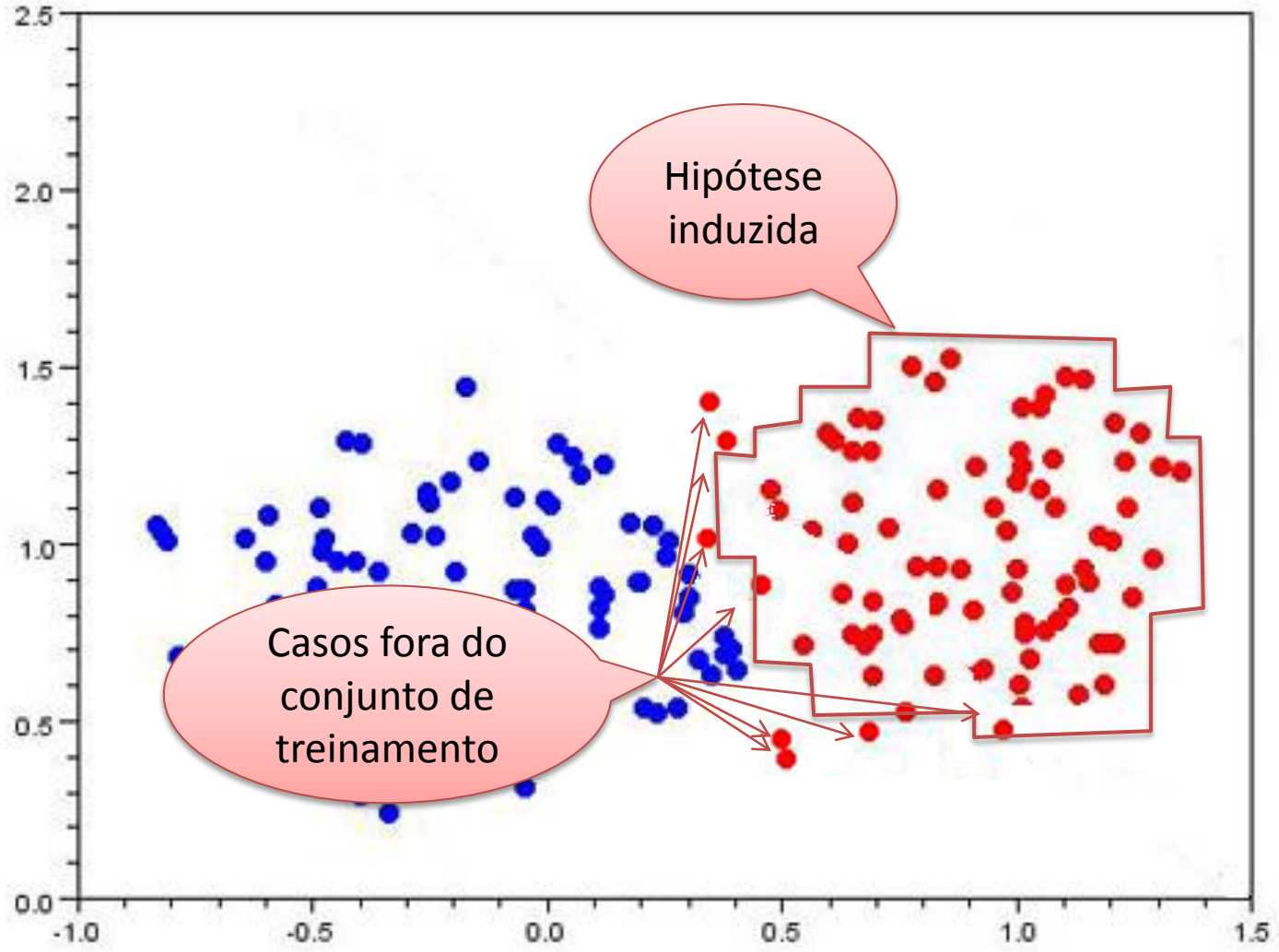
- Para classificar um novo exemplo com (X1,X2) = (2.5, 4), basta verificar em qual região ela se localiza e atribuir a classe associada àquela região (neste caso, classe **o**)



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.

Exemplo


Hipótese induzida
Casos fora do conjunto de treinamento
2.5
2.0
1.5
1.0
0.5
0.0
-1.0
-0.5
0.0
0.5
1.0
1.5

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)

Overtuning

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

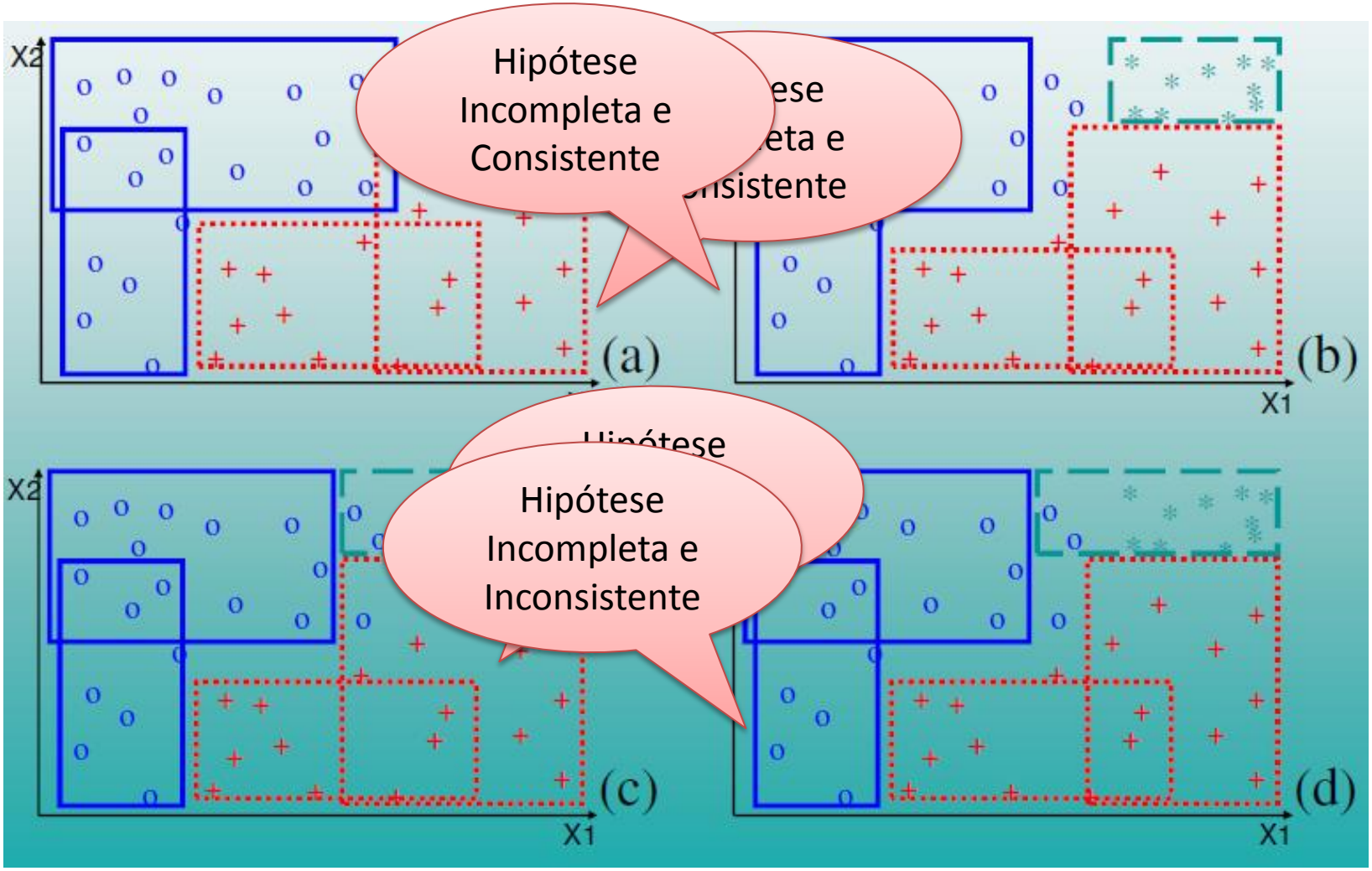
Poda

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

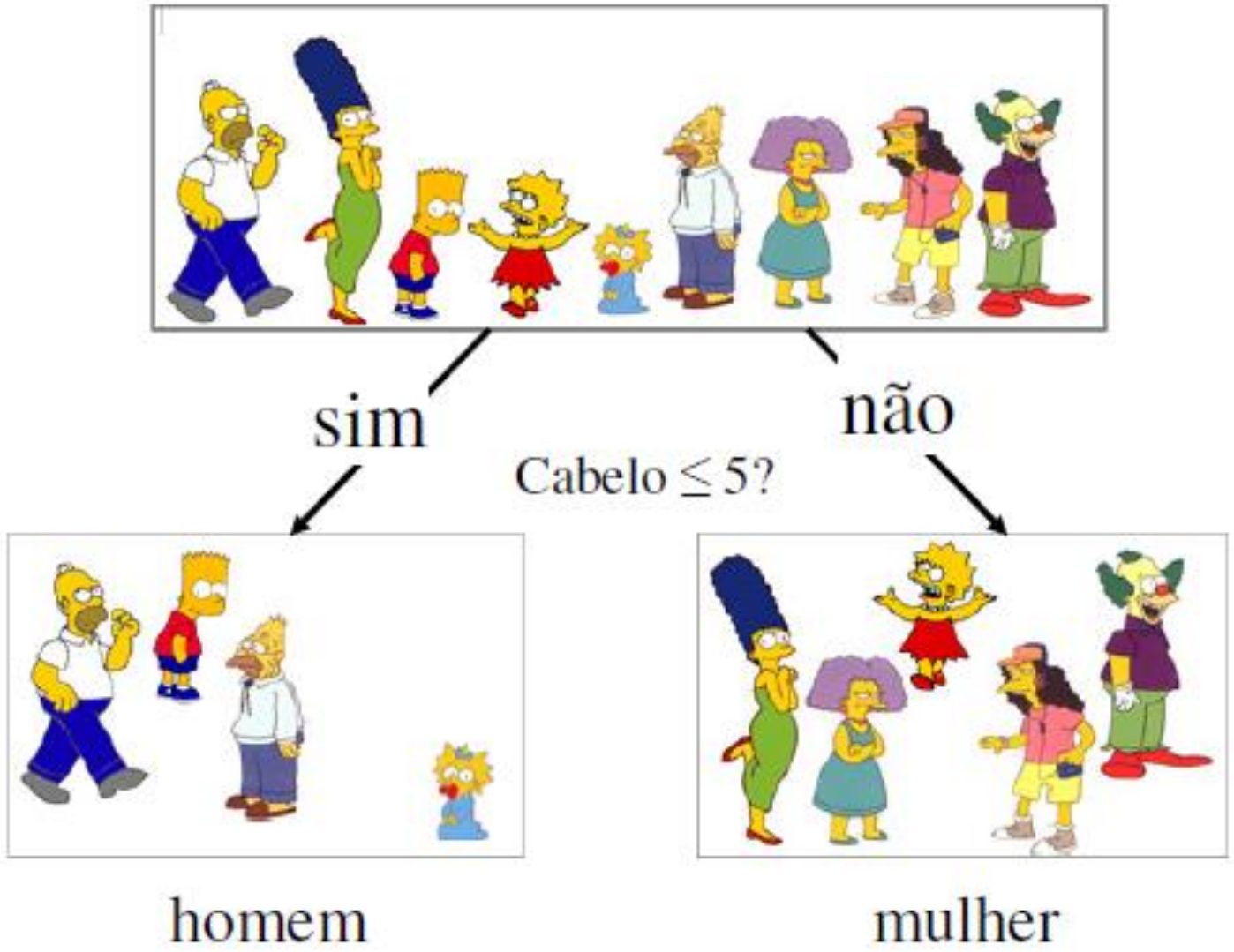
Consistência e Completude

- 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

Relação entre Completude e Consistência



Como classificar a hipótese abaixo?



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	...	Prevista C_k
Real C_1	$M(C_1, C_1)$	$M(C_1, C_2)$	...	$M(C_1, C_k)$
Real C_2	$M(C_2, C_1)$	$M(C_2, C_2)$	...	$M(C_1, C_k)$
⋮	⋮	⋮	⋱	⋮
Real C_k	$M(C_k, C_1)$	$M(C_k, C_2)$	...	$M(C_k, C_k)$

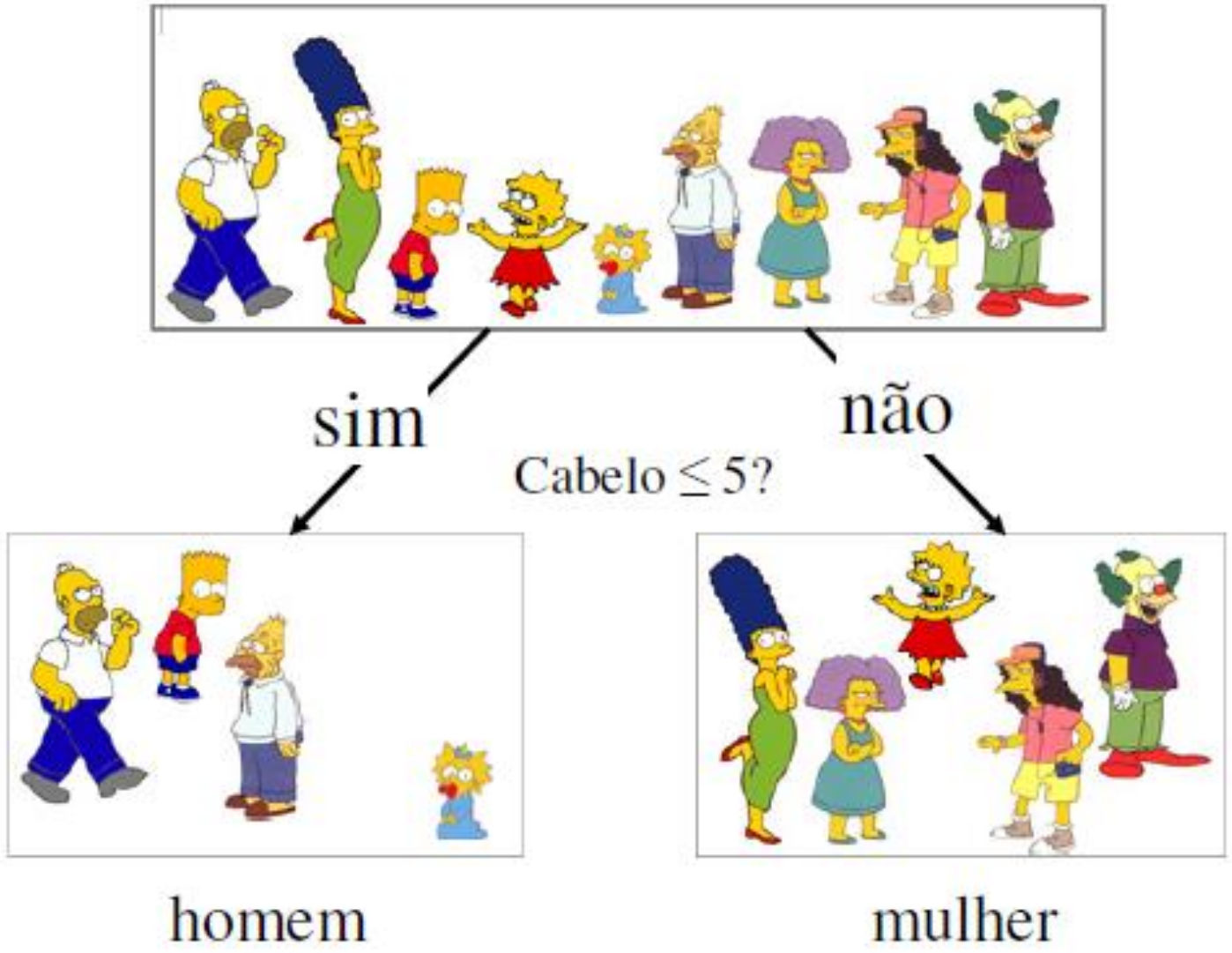
$$M(C_i, C_j) = \sum_{\forall (x,y) \in T : y = C_i} \| h(x) = C_j \|$$

Matriz de Confusão para 2 Classes

Classe	Prevista C+	Prevista C-	Taxa de erro da Classe	Taxa de erro da Total
Real C+	T_{Pos}	F_{Neg}	$\frac{F_{Neg}}{T_{Pos} + F_{Neg}}$	$\frac{F_{Pos} + F_{Neg}}{n}$
Real C-	F_{Pos}	T_{Neg}	$\frac{F_{Pos}}{F_{Pos} + T_{Neg}}$	

- T_{Pos} = *True Positive* (verdadeiro positivo)
- F_{Neg} = *False Negative* (falso negativo)
- F_{Pos} = *False Positive* (falso positivo)
- T_{Neg} = *True Negative* (verdadeiro negativo)
- $n = (T_{Pos} + F_{Neg} + F_{Pos} + T_{Neg})$

Exercício: monte a matriz de confusão!



AVALIAÇÃO DE UM CLASSIFICADOR

Avaliação do classificador

- Para se estimar o erro verdadeiro de um classificador, a **amostra** para teste deve ser escolhida de maneira aleatória.
- As 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.

Métodos para estimar o erro verdadeiro de um classificador

1. Resubstitution
2. Holdout (Validação Simples)
3. Random
4. r-fold cross-validation
5. r-fold stratified cross-validation
6. Leave-one-out

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.

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*.

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*

r-fold cross-validation

- Os exemplos são aleatoriamente divididos em *r* **partições** (folds) de tamanho aproximadamente iguais (*n*/*r*)
- Os exemplos de (*r*-1) folds são usados de modo independente 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 teste. O erro do cross-validation é a média dos erros dos *r* folds.

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.

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.*
- É um caso especial de cross-validation.
- Computacionalmente caro e usado apenas quando o conjunto de exemplos é pequeno.

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.

Metodologia para a Avaliação

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.

(Russel e Norvig, 2003)

Calculo de 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:

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

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

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

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:

$$média(A) = \frac{1}{r}\sum_{i=1}^{r} err(h_i) = \frac{90.0}{10} = 9.0$$

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

Comparando dois Algoritmos

$$média(A_s - A_p) = média(A_s) - média(A_p)$$

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

$$diferença\ absoluta(A_s - A_p) = \frac{média(A_s - A_p)}{desvio(A_s - A_p)}$$

- Sendo: A_s algoritmo padrão; e A_p algoritmo proposto

Comparando dois Algoritmos

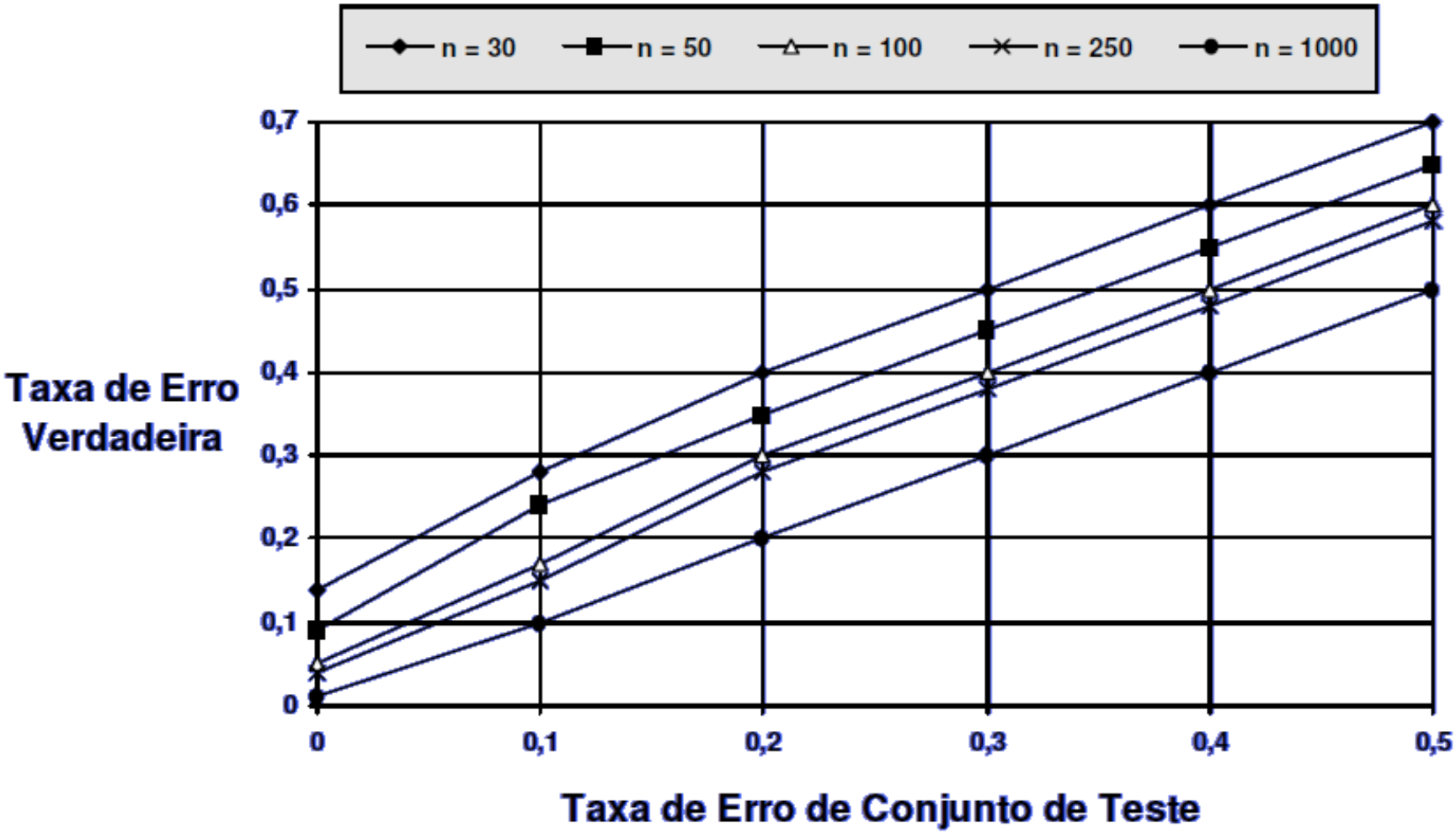
- Se $dif.absoluta(A_s - A_p) > 0$,
  - A_p tem melhor performance que A_s
- Se $dif.absoluta(AS - AP) >= 2$,
  - A_p tem melhor performance que A_s com um nível de confiança de 95%.
- Se $dif.absoluta(AS - AP) <= 0$,
  - A_s tem melhor performance que A_p
- Se $dif.absoluta(AS - AP) <= -2$,
  - A_s tem melhor performance que A_p com um nível de confiança de 95%.

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?”

Número de Casos de Teste e Qualidade da Predição



Número de Casos de Teste e Qualidade da Predição

- 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.