Introdução ao Aprendizado de Máquina

Profª: Solange Oliveira Rezende

Aprendizado de Máquina

- Quantidade de conhecimento disponível pode ser muito grande para ser descrito (e portanto programado) por humanos.

- Ser humano não é capaz de executar algumas tarefas que demandam quantidades grandes de cálculos complexos, passíveis apenas de execução em computador:

Aprendizado de Máquina

- Definição formal (Mitchell, 1997):

É dito que um programa de computador

" **aprende"** a partir de **experiências** ***E*** *com respeito*

a alguma **classe de tarefas** ***T*** e ***medida de***

desempenho ***P,***

se seu desempenho em tarefas de ***T,*** *medido por* ***P,***

melhora com a experiência ***E***

Aplicações

- Algoritmos de AM têm sido bem sucedidos, por exemplo para:
  - Identificar genes associados a determinadas doenças
  - Discriminar tecidos (saudáveis e doentes), objetos celestiais, ...
  - Identificar nichos de mercado
  - Prever a vazão de rios e nível de represas
  - Detectar uso fraudulento de cartões de crédito
  - Otimizar ações de controle em processos de produção
  - Reconhecimento de faces, de voz, de assinaturas ...
  - ...

Características

- Três características devem ser identificadas para um problema ser bem definido:
  - A **classe de tarefas**
  - A **medida de desempenho a ser melhorada**
  - A **origem da experiência**
- Problema de Aprender Xadrez:
  - **Tarefa T: jogar xadrez**
  - **Medida de desempenho P: porcentagem** dos jogos vencidos contra adversários
  - **Experiência de treinamento E: praticar** jogando contra si próprio ou contra adversários humanos (p.ex. internet)

Sistema de Aprendizado

Modos de Aprendizado	Paradigmas de Aprendizado (Modelos representação)	Linguagens de Descrição	Formas de Aprendizado
■ Supervisionado ■ Não Supervisionado ■ Semi Supervisionado	■ Simbólico ■ Estatístico ■ Baseado em Exemplos (Instance-Based) ■ Conexionista ■ Genético	■ Instâncias ou Exemplos ■ Conceitos Aprendidos ou Hipóteses ■ Teoria de Domínio ou Conhecimento de Fundo	■ Incremental ■ Não Incremental

Aprendizado SUPERVISIONADO

- Guiado por um “professor” externo
- “Professor” possui conhecimento sobre o ambiente
- Representado por conjunto de pares (**x, d)**
- Modelo procura reproduzir comportamento do “professor”

Aprendizado por REFORÇO

- Guiado por um “crítico” externo
- Processo de tentativa e erro
- Procura maximizar sinal de reforço
- Se ação tomada por sistema é seguida por estado satisfatório, sistema é fortalecido, caso contrário, sistema é enfraquecido
- Tipos de reforço
  - Positivo = recompensa
  - Negativo = punição
  - Nulo

Aprendizado NÃO-SUPERVISIONADO

- Não tem crítico ou professor externo
- Extração de propriedades estatisticamente relevantes
- Exemplos:
  - *Clustering*: descobre categorias automaticamente
  - Associação
  - Sumarização

Aprendizado SEMI-SUPERVISIONADO

- Tem um professor externo apenas para parte dos exemplos de treinamento
- Exemplo:
  - Web mining: usuários podem fornecer alguns exemplos de páginas similares, pertencentes a uma determinada categoria, mas uma parcela ínfima de *web pages* teria essa informação associada

Modelos de Representação

- Modelos Matemáticos
  - Regressão linear,
  - Redes neurais (paradigma **conexionista / bioinspirado),**
  - Máquinas de vetores de suporte, ...
- Modelos Simbólicos
  - Árvores de decisão,
  - Regras em lógica proposicional ou de 1ª ordem,
  - Redes semânticas, ...

Modelos de Representação

- Modelos “Lazy” (paradigma baseado em **instâncias**)
  - K-NN,
  - Raciocínio Baseado em Casos (CBR), ...
- Modelos Probabilísticos (paradigma **probabilístico**)
  - Naive Bayes,
  - Redes Bayesianas,
  - Misturas de Gaussianas,
  - Modelos de Markov Escondidos (HMMs), ...

Aprendizado de Máquina

- Representatividade dos exemplos
  - Aprendizado é mais confiável quando exemplos de treinamento seguem uma distribuição representativa (semelhante) à da população

Aprendizado de Máquina

- Um sistema de AM deve ter:
  - Tipo exato de conhecimento a ser aprendido
    - **Função alvo**
  - Uma representação para o conhecimento adquirido
    - **Modelo**
  - Um mecanismo de aprendizado
    - **Técnica de aprendizado**

Função ALVO

- Estabelece qual conhecimento será aprendido e permite verificar quão bem ele foi aprendido
- Exemplos:
  - Função discriminante entre classes
  - Função de similaridade intra grupos
  - ...

Técnicas de Aprendizado

- Dado um tipo de **modelo**, uma **função alvo** e um **conjunto de exemplos de treinamento**, é preciso algum mecanismo para obter, a partir dos exemplos, um modelo específico daquele tipo que represente bem a função alvo.
- Esse mecanismo, denominado mecanismo de aprendizado, consiste fundamentalmente de uma **técnica de busca.**
- Busca-se no espaço dos modelos plausíveis por aquele modelo específico que melhor represente a função alvo.

Técnicas de Aprendizado

- Cada tipo de **modelo** é mais **apropriado** para uma determinada **classe de problemas**
- Assim como cada **técnica** de aprendizado é **mais apropriada** para um **tipo de modelo**
- É parte importante do estudo de AM **aprender a identificar os cenários** mais apropriados **para cada modelo e técnica** de aprendizado

Avaliação de AM

- Uma vez obtido um modelo a partir de exemplos de treinamento e de uma técnica de aprendizado, é preciso **avaliar a eficácia / eficiência deste modelo / técnica** para resolver a tarefa em questão
  - Em outras palavras, é preciso **validar ou não o modelo obtido**

Avaliação de AM

- **Avaliação Experimental**
  - Conduzir experimentos controlados
  - Dados reais representativos em aplicações práticas
  - Dados **benchmark em estudos acadêmicos e comparações**
  - Extrair resultados de desempenho
    - Ex.: Acurácia de teste, tempos de treinamento e teste, etc.
  - Analisar resultados e diferenças com rigor estatístico

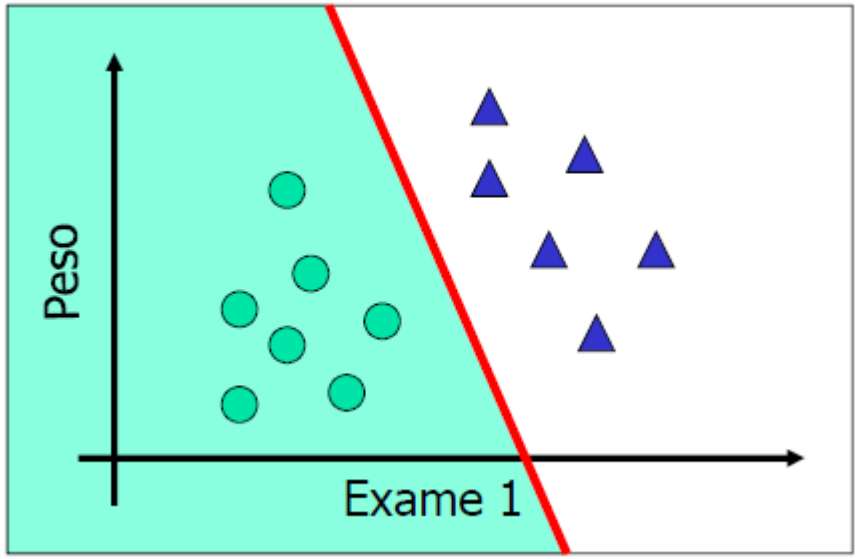
Avaliação de AM

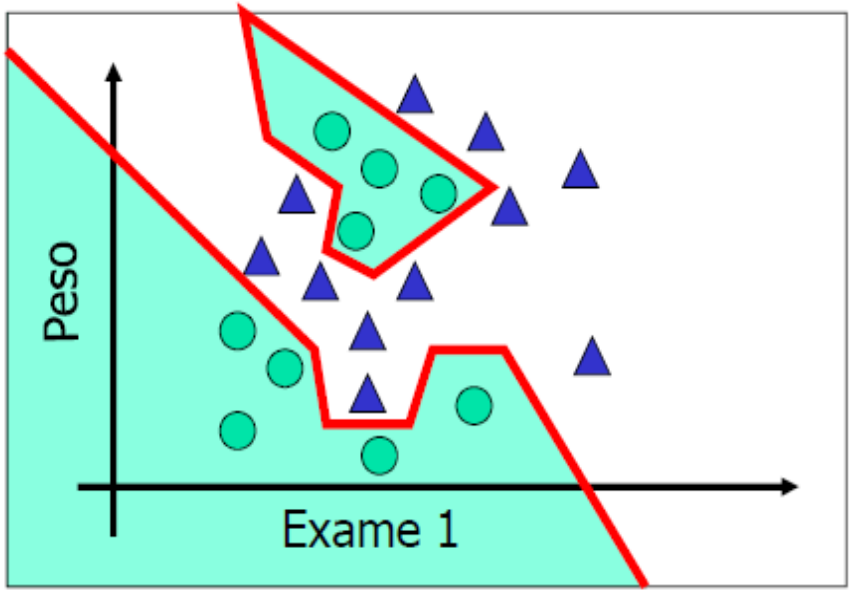
- **Avaliação Teórica**
  - Analisar algoritmos matematicamente e provar:
    - Complexidade computacional
    - Habilidade para ajustar dados de treinamento
    - Habilidade para generalizar dados de treinamento
    - Complexidade da amostra
    - Ordem de grandeza do no. de exemplos de treinamento necessários para aprender uma função com dada acurácia

Áreas de Aplicação de AM

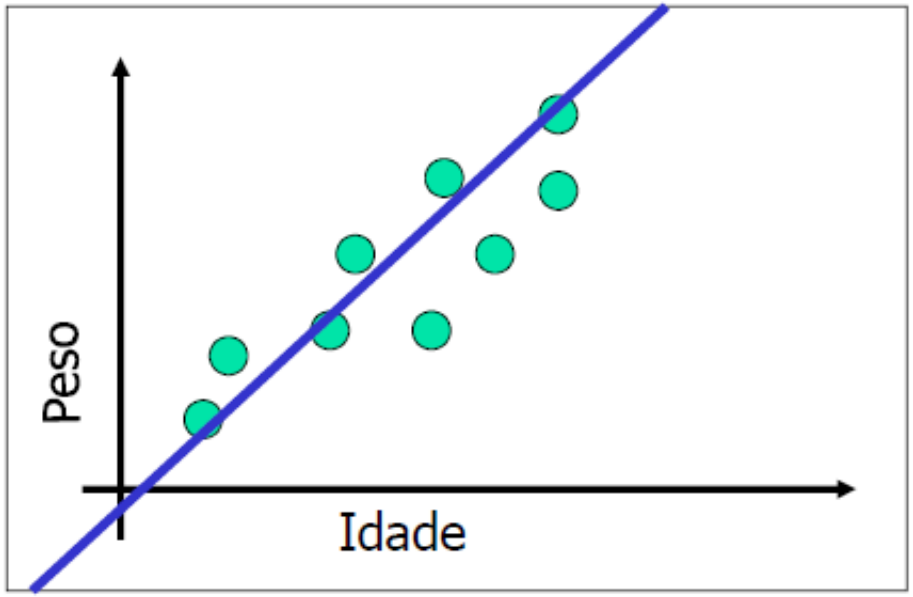
- Predição
  - Classificação de Padrões
  - Regressão
  - Detecção de Anomalias
- Descrição
  - Análise de Agrupamentos
  - Análise de Associação
  - Indução de Regras
- Otimização
- Automação e Processamento de Sinais

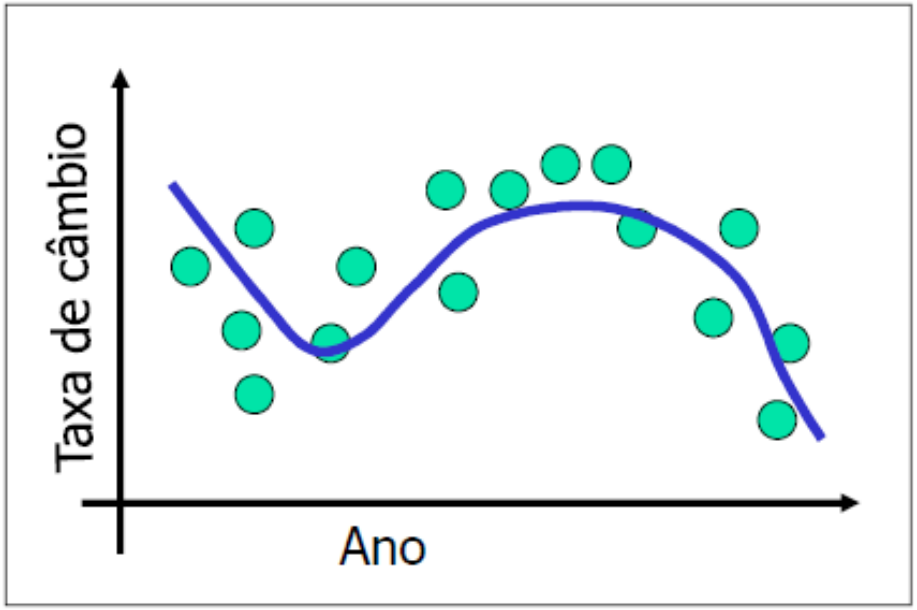
Classificação



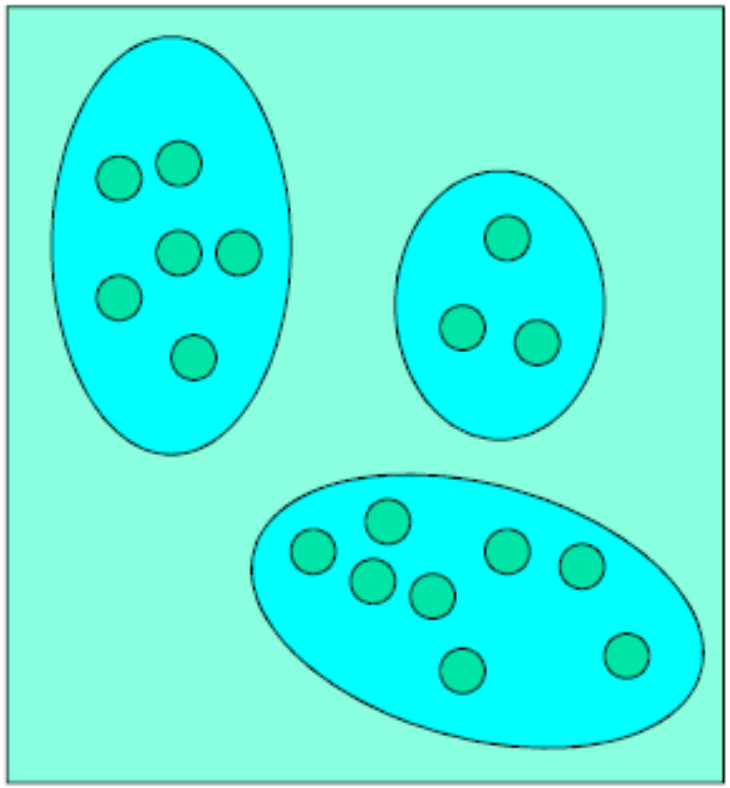


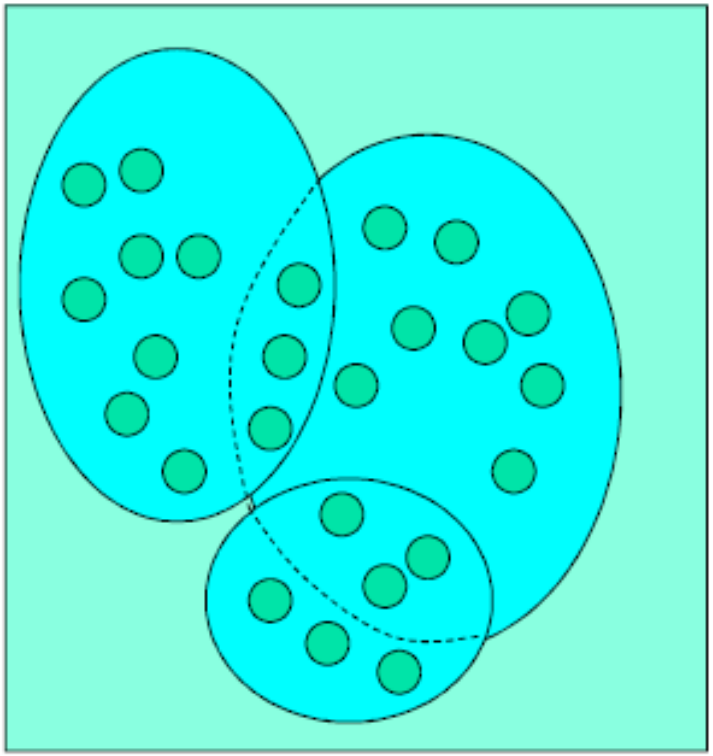
Regressão





Agrupamento





Regras de Associação

- Técnica descobre relações simétricas ou assimétricas entre conjuntos de padrões
- Exemplos de regras de associação
  - {Fraldas} ⇒ {Leite} (**útil, porém previsível)**
  - {Fraldas, Leite} ⇒ {Cerveja} (**útil e inovadora)**

Conceitos Básicos de Aprendizado de Máquina Supervisionado

Erro e Precisão

Recordando a notação 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

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

$$ca(h) = 1 - ce(h) \qquad \text{(precisão)}$$

Erro e Precisão

- Regressão: Erro quadrático médio (MSE) e distância absoluta média (MAD)

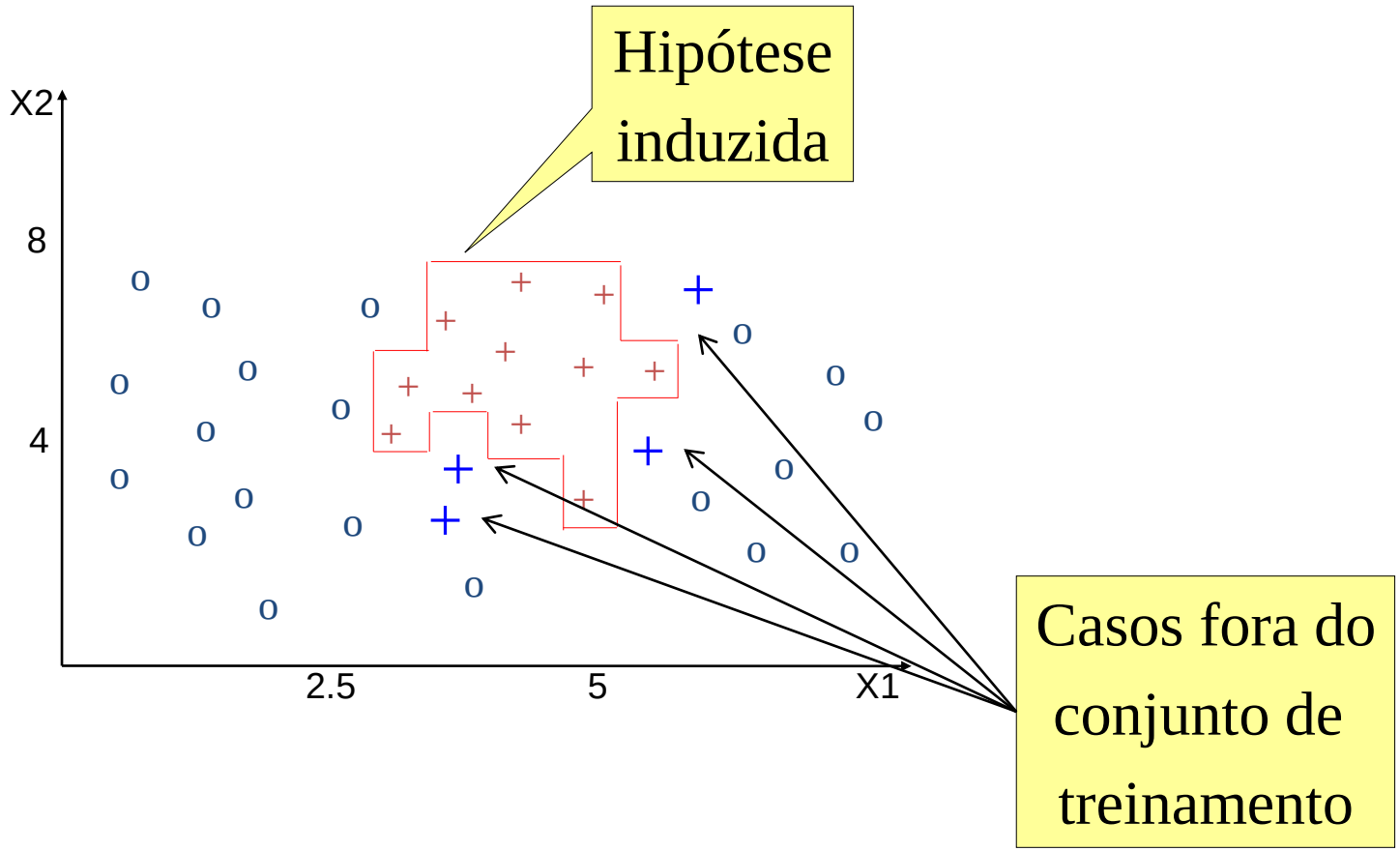
$$\text{pe-mse}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h(x_i))^2$$

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

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

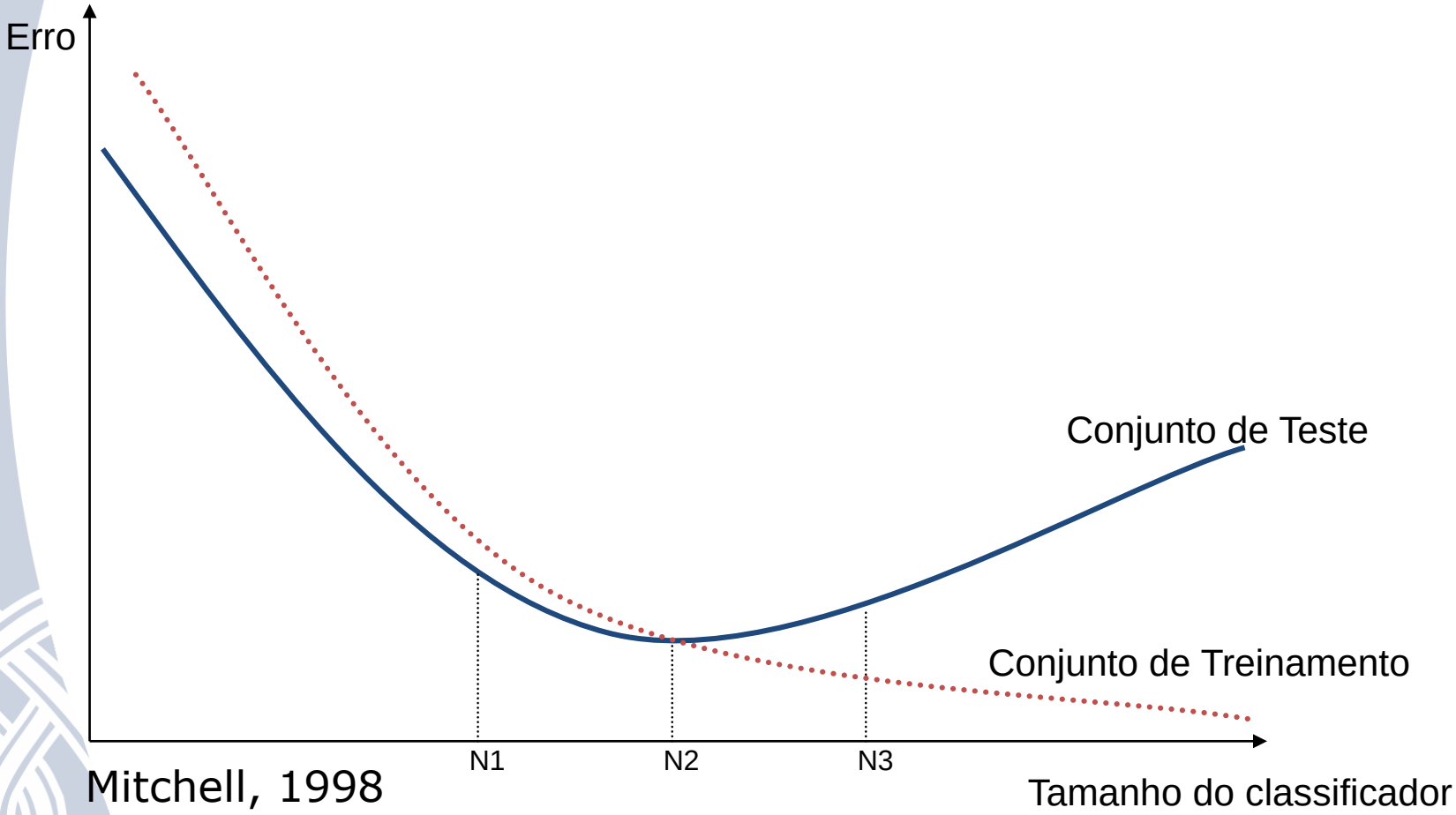
Overfitting - Exemplo



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

Relação entre o tamanho do previsor e o erro

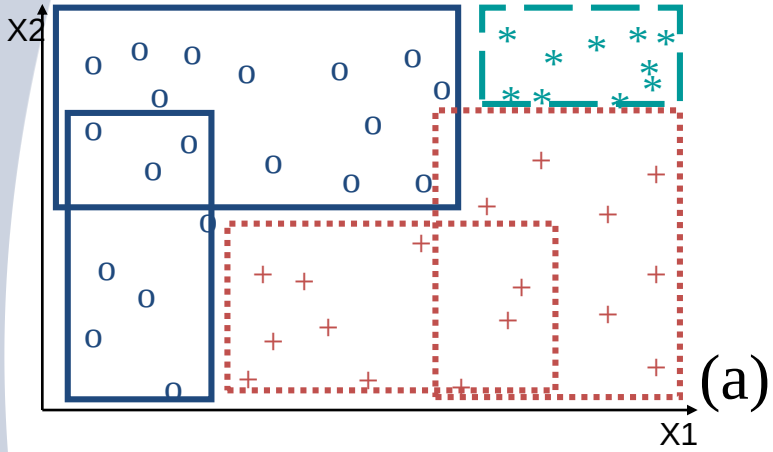


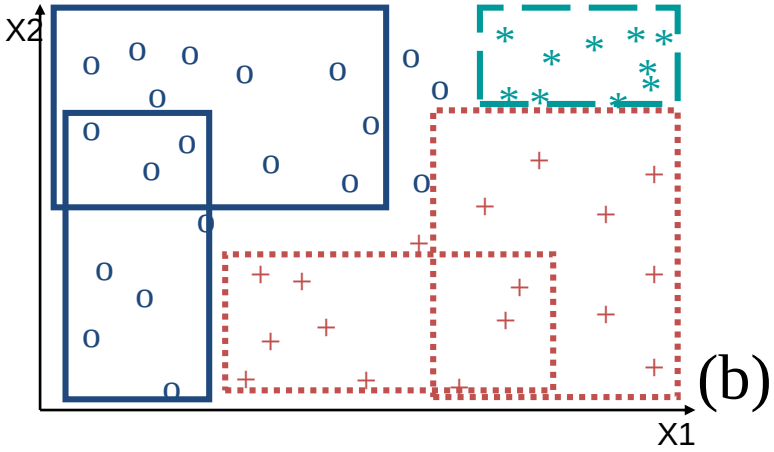
Mitchell, 1998

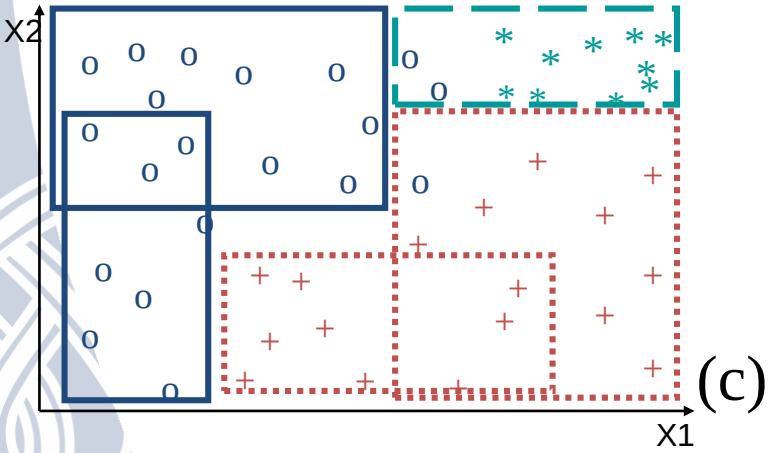
Consistência e Completude

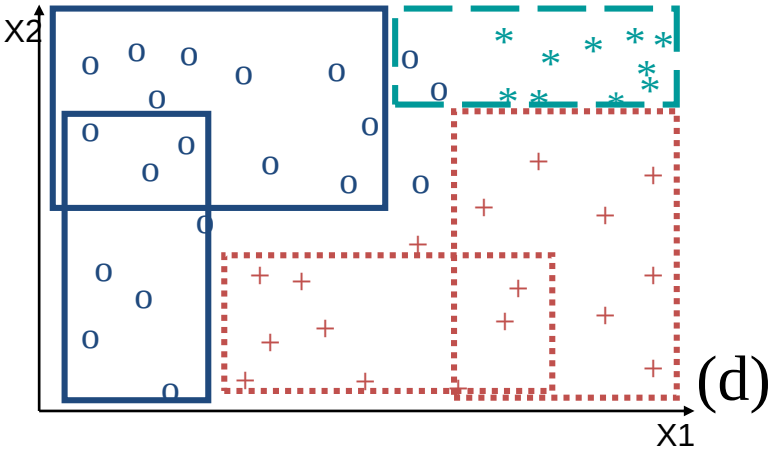
- Depois de induzida, uma hipótese pode ser avaliada sobre
  - consistência, se classifica corretamente todos os exemplos
  - completude, se classifica todos os exemplos

Relação entre Completude e Consistência

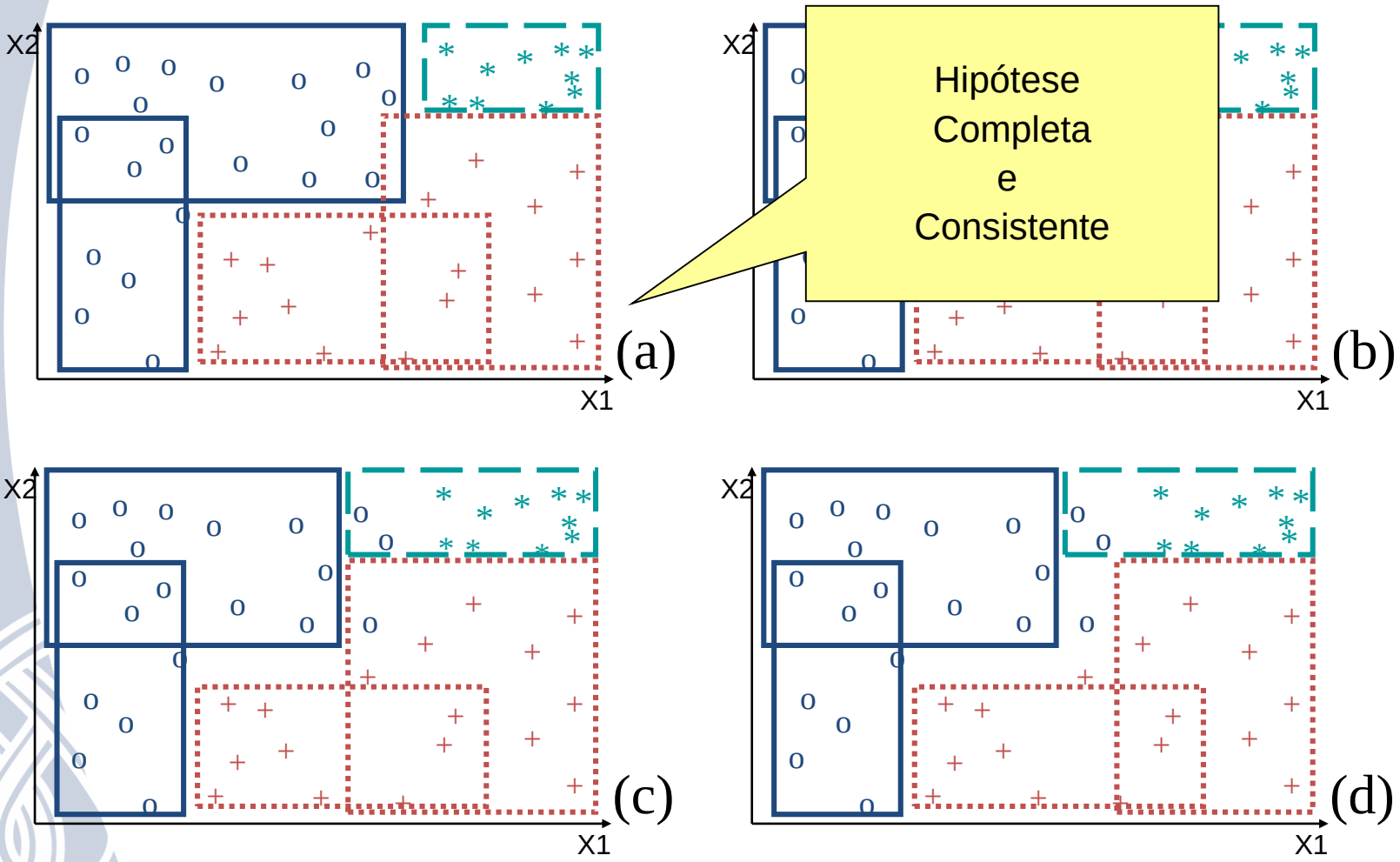




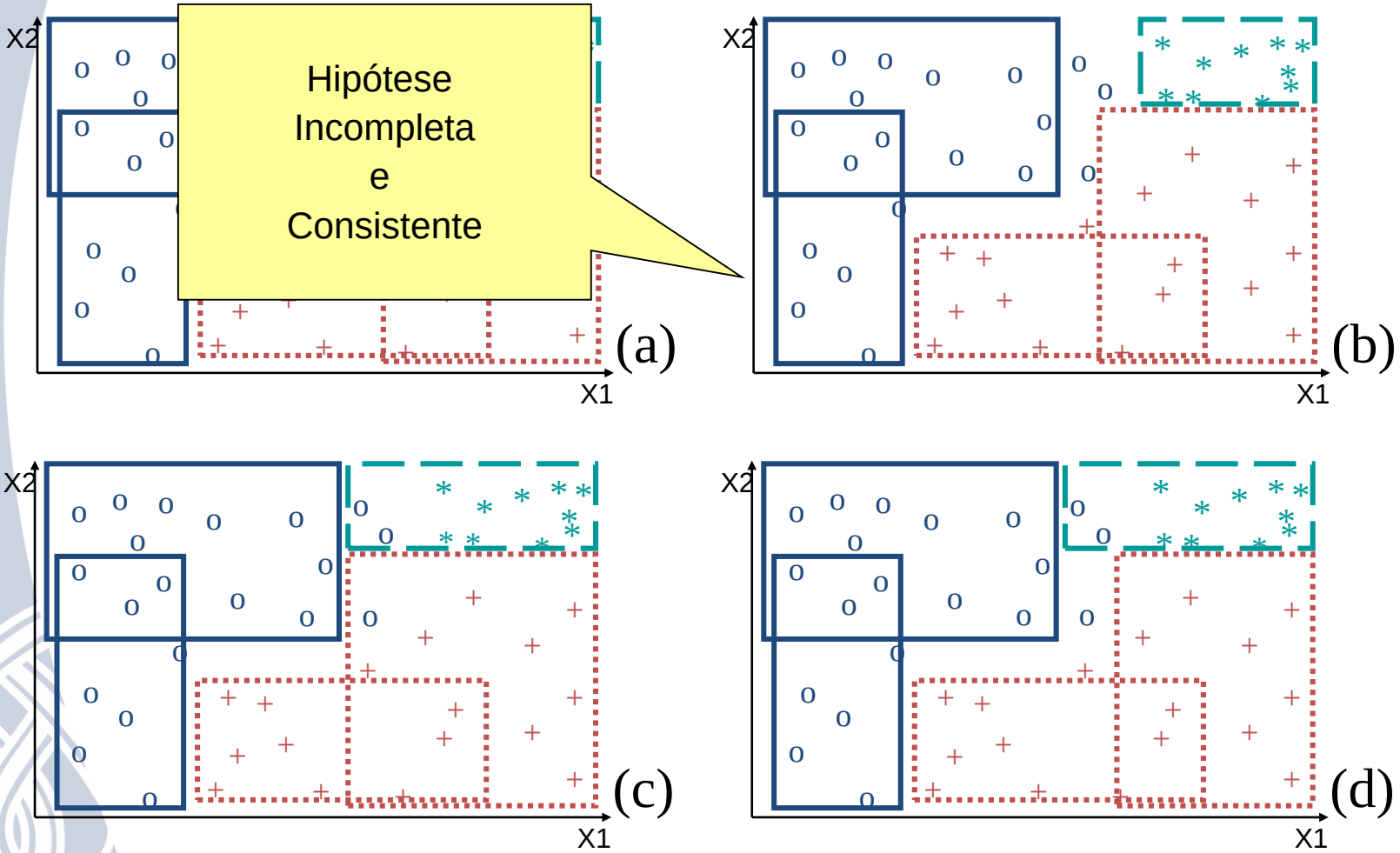




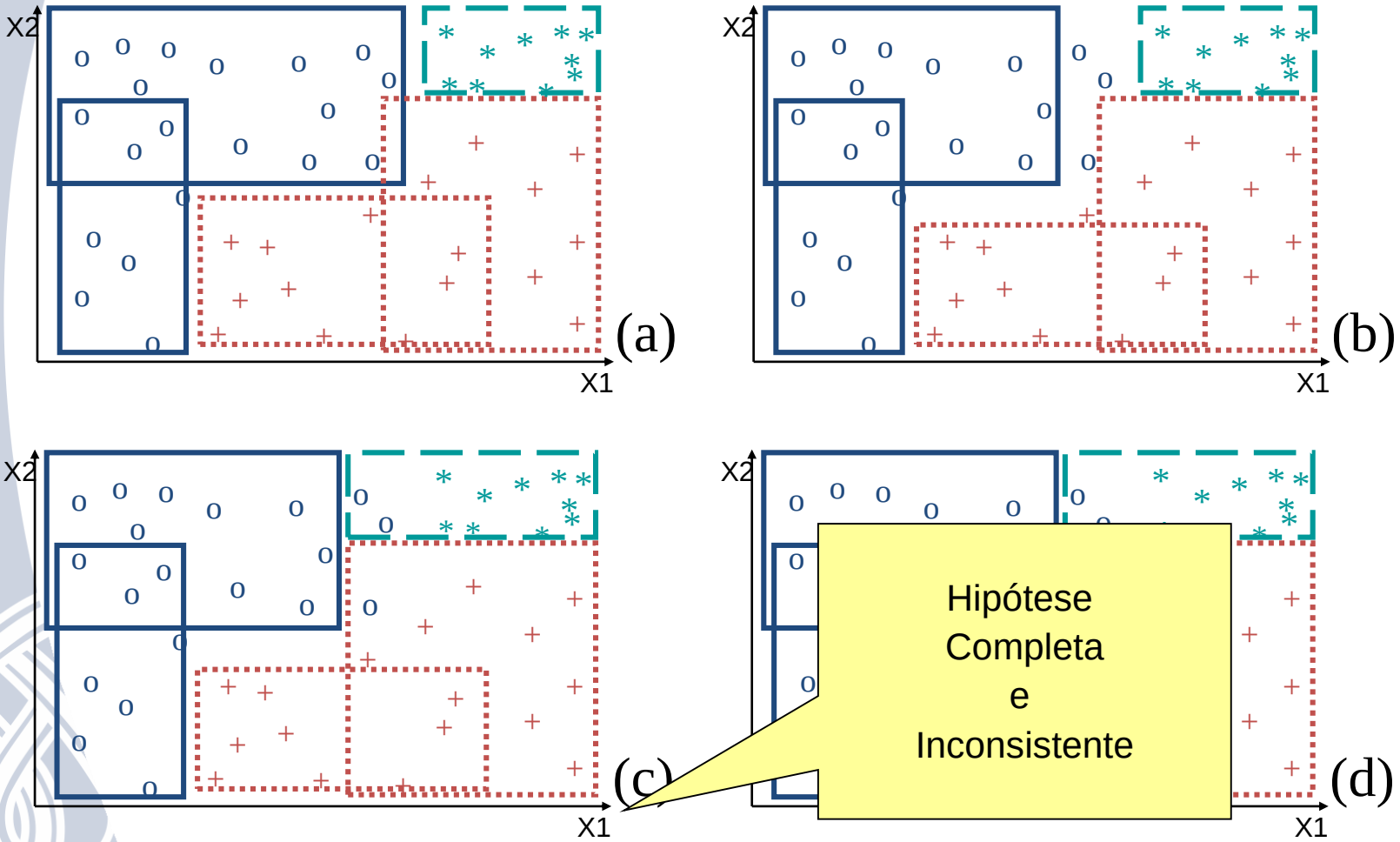
Relação entre Completude e Consistência



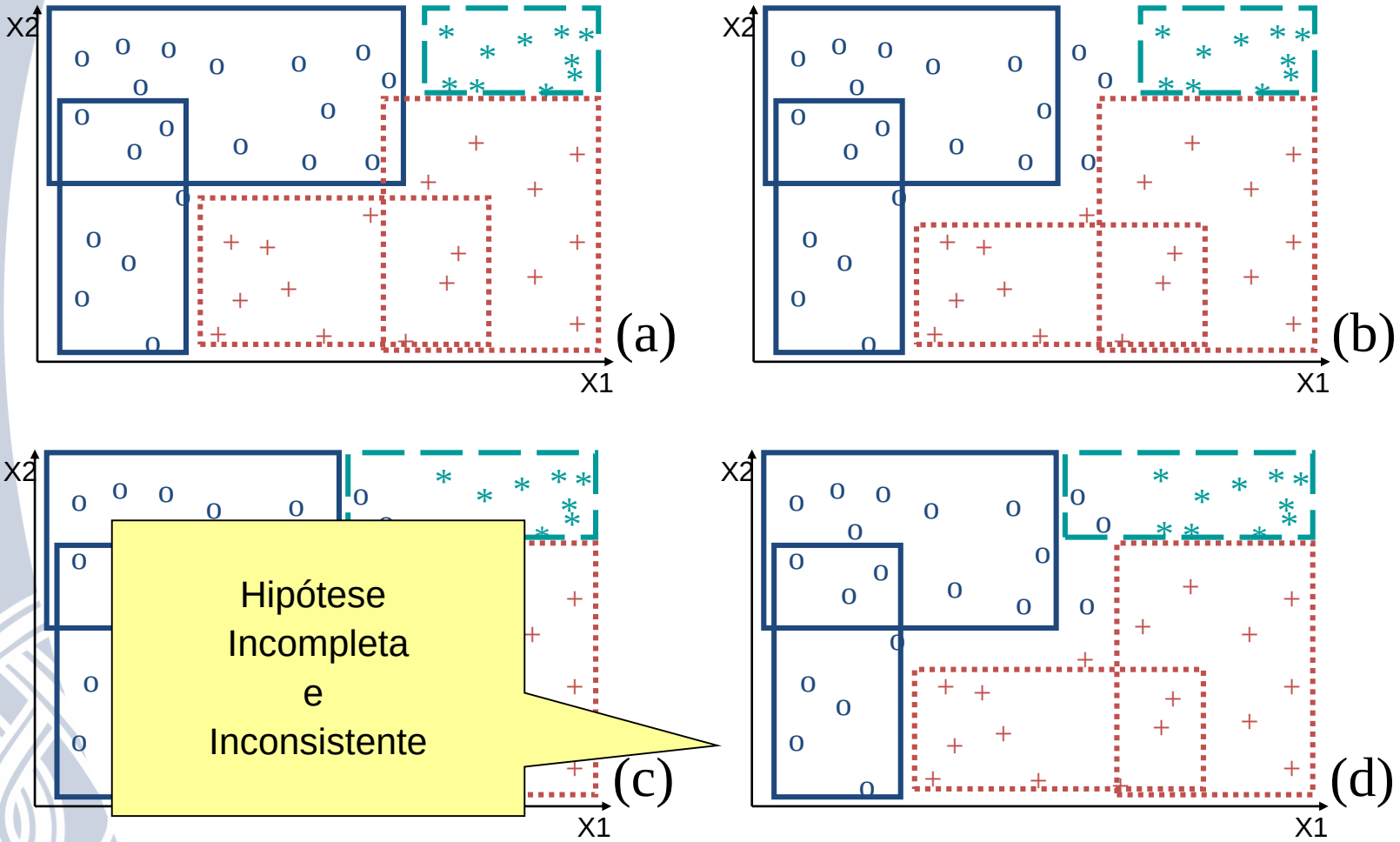
Relação entre Completude e Consistência



Relação entre Completude e Consistência



Relação entre Completude e Consistência



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

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

Class Label	predicted C_1	predicted C_2	$\cdots$	predicted C_1
true C_1	$M(C_1, C_1)$	$M(C_1, C_2)$	$\cdots$	$M(C_1, C_k)$
true C_2	$M(C_2, C_1)$	$M(C_2, C_2)$	$\cdots$	$M(C_2, C_k)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
true C_k	$M(C_k, C_1)$	$M(C_k, C_2)$	$\cdots$	$M(C_k, C_k)$

Modelo de classificação de duas classes

- Por exemplo, dada uma regra, um exemplo e uma classe, podemos ter 4 casos:
  - **true positive**: o exemplo satisfaz todas as condições da regra e a classe do exemplo é a mesma prevista na regra
  - **false positive**: o exemplo satisfaz todas as condições da regra e a classe do exemplo não é mesma prevista na regra
  - **false negative**: o exemplo não satisfaz todas as condições da regra e a classe do exemplo é a mesma prevista na regra
  - **true negative**: o exemplo não satisfaz todas as condições da regra e a classe do exemplo não é a mesma prevista na regra

Matriz de Confusão para 2 Classes

Class label	predicted C_+	predicted C_-	Class error rate	Total error rate
true C_+	T_P	F_N	$\frac{F_N}{T_P + F_N}$	$\frac{F_P + F_N}{n}$
true C_-	F_P	T_N	$\frac{F_P}{F_P + T_N}$	

T_P = True Positive
F_N = False Negative
F_P = False Positive
T_N = True Negative
$n = (T_P+F_N+F_P+T_N)$

Matriz de Confusão para 2 Classes

Outras métricas derivadas da tabela anterior:

$$C_+ \text{PredictiveValue} = \frac{T_P}{T_P + F_P}$$

$$C\_\text{PredictiveValue} = \frac{T_N}{T_N + F_N}$$

$$\text{True } C_+ \text{ Rate ou Sensitivity ou Recall} = \frac{T_P}{T_P + F_N}$$

$$\text{True } C\_ \text{ Rate ou Specifity} = \frac{T_N}{F_P + T_N}$$

$$\text{Precisão} = \frac{T_P + T_N}{n}$$

Resampling

- 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

Métodos para estimar o erro verdadeiro de um classificador

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

Resubustituition

- Gera o classificador e testa a sua performance com o mesmo conjunto de dados
  - os desempenhos computados com este método são otimistas e tendenciosos

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 divididas 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 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 a 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*
- Caso especial de cross-validation
- Computacionalmente caro e usado apenas quando o conjunto de exemplos é pequeno

Bootstrap

- Repetir diversas vezes o processo inteiro de classificação e cada experimento é baseado em um conjunto de treinamento novo, obtido por resampling com reposição do conjunto de dados original
  - Existem muitos estimadores bootstrap

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 acurácia e escolher entre os algoritmos aquele que apresentar maior acurácia, por exemplo, para um domínio específico

Metodologia de Avaliação (Russel e Norvig, 2003)

1. Coletar um conjunto de exemplos, de preferencia sem "ruido"
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

Calculando Média e Desvio Padrão usando Resampling

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

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

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

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

Calculando Média e Desvio Padrão usando Resampling

- 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:

$$média(A) = \frac{90.0}{10} = 9.0$$

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

Comparando dois Algoritmos

$A_S \Rightarrow$ algoritmo standart

$Ap \Rightarrow$ algoritmo proposto

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

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

$$ad(A_S - A_P) = \frac{média(A_S - A_P)}{sd(A_S - A_P)}$$

Ad - Diferença absoluta em Desvios Padrões

Comparando dois Algoritmos

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

Comparando dois Algoritmos

- Ao comparara dois indutores no mesmo domínio Exemplo: considerando que A_S = 9.00±1.00 (alg. padrão) e A_P = 7.50±0.80 (alg. proposto)

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

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

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

Como $ad(A_S\text{-}A_P) > 0$, A_P supera A_S

Sugestão de leitura

- Mitchell, T. M., ***Machine Learning***, McGraw-Hill, 1997.
- P.-N. Tan, Steinbach, M., and Kumar, V., ***Introduction to Data Mining***, Addison-Wesley, 2006.
- Witten, I. H. & Frank, E., ***Data Mining – Pratical Machine Learning Tools and Techniques***, Elsevier, 2005.
- Rezende, S. O; *Sistemas Inteligentes: Fundamentos e Aplicações;* Ed Manole 2003 (cap 4 e 5).

- **Agradecimentos:**
  - Material desenvolvido com ajuda de Rafael Geraldeli Rossi, Ronaldo Prati