

The background of the slide is a complex, abstract composition. It features a dark blue and black base with glowing binary code (0s and 1s) in light blue and white. Overlaid on this are several financial-style charts: a line graph with a red line, a bar chart with red and blue bars, and a candlestick chart. The overall effect is a blend of digital data and financial analysis.

Avaliação de Modelos / Amostragem de Dados

Huei Diana Lee

Inteligência Artificial
CECE/UNIOESTE-FOZ

Avaliação de Modelos

Estimativa de
performance

a
priori



a
posteriori

Avaliação de Sistemas de Aprendizado

Teórica – *a priori*

Analisar algoritmos matematicamente:

- Complexidade computacional
- Habilidade de se adaptar aos dados de treinamento
- Número de exemplos de treinamento necessários para se aprender uma função correta

Experimental – *a posteriori*

- Conduzir experimentos
- Coletar dados sobre o seu desempenho, por exemplo, acurácia, tempo de treinamento, tempo de teste
- Analisar a significância estatística

Avaliação Experimental de Hipóteses

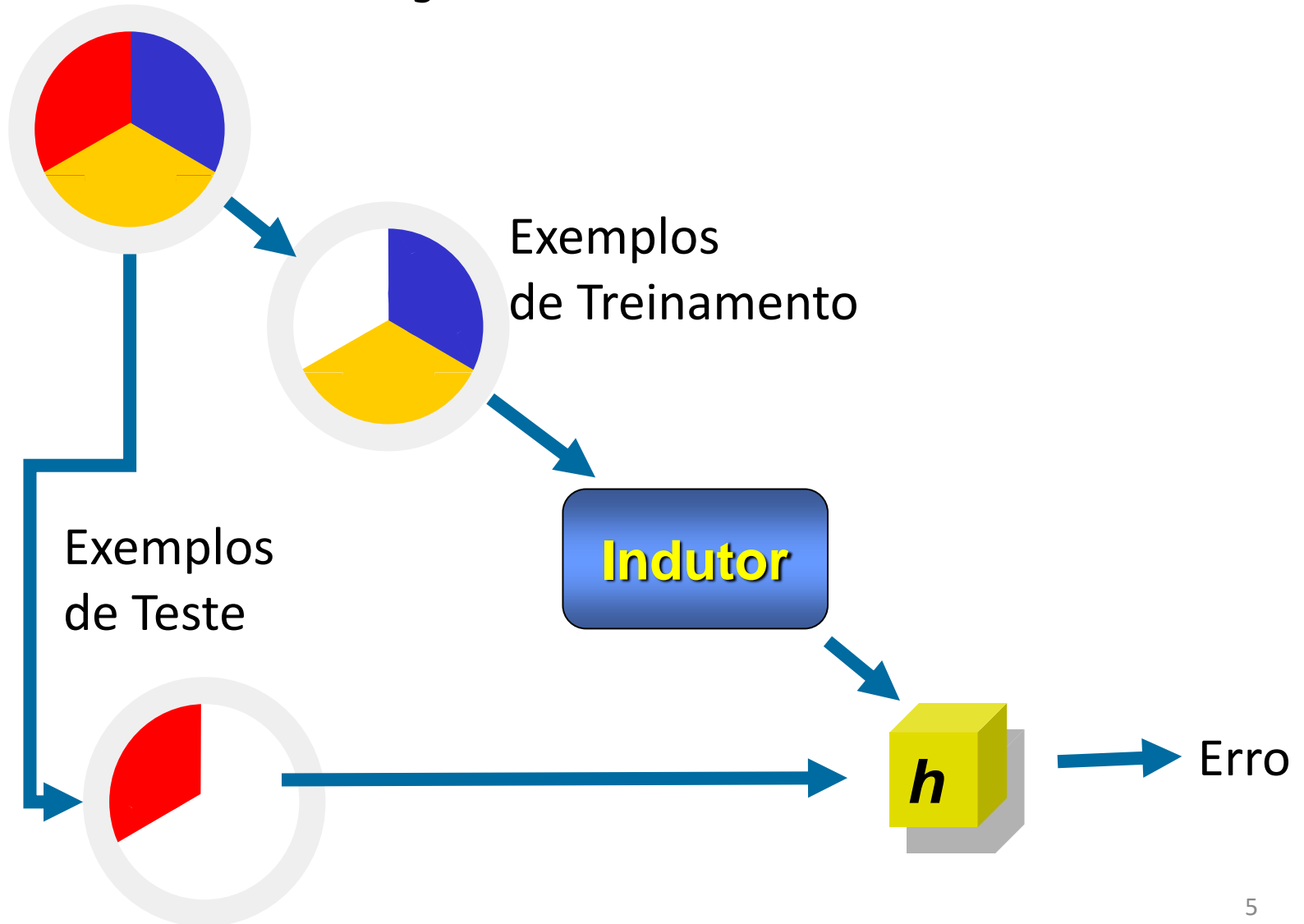
Classificador:

- Por si só não fornece uma boa estimativa de sua capacidade de previsão
- Possui boa capacidade de **descrever** os dados, não de **predizer**

Conjunto de dados:

- Treinamento
- Validação
- Teste

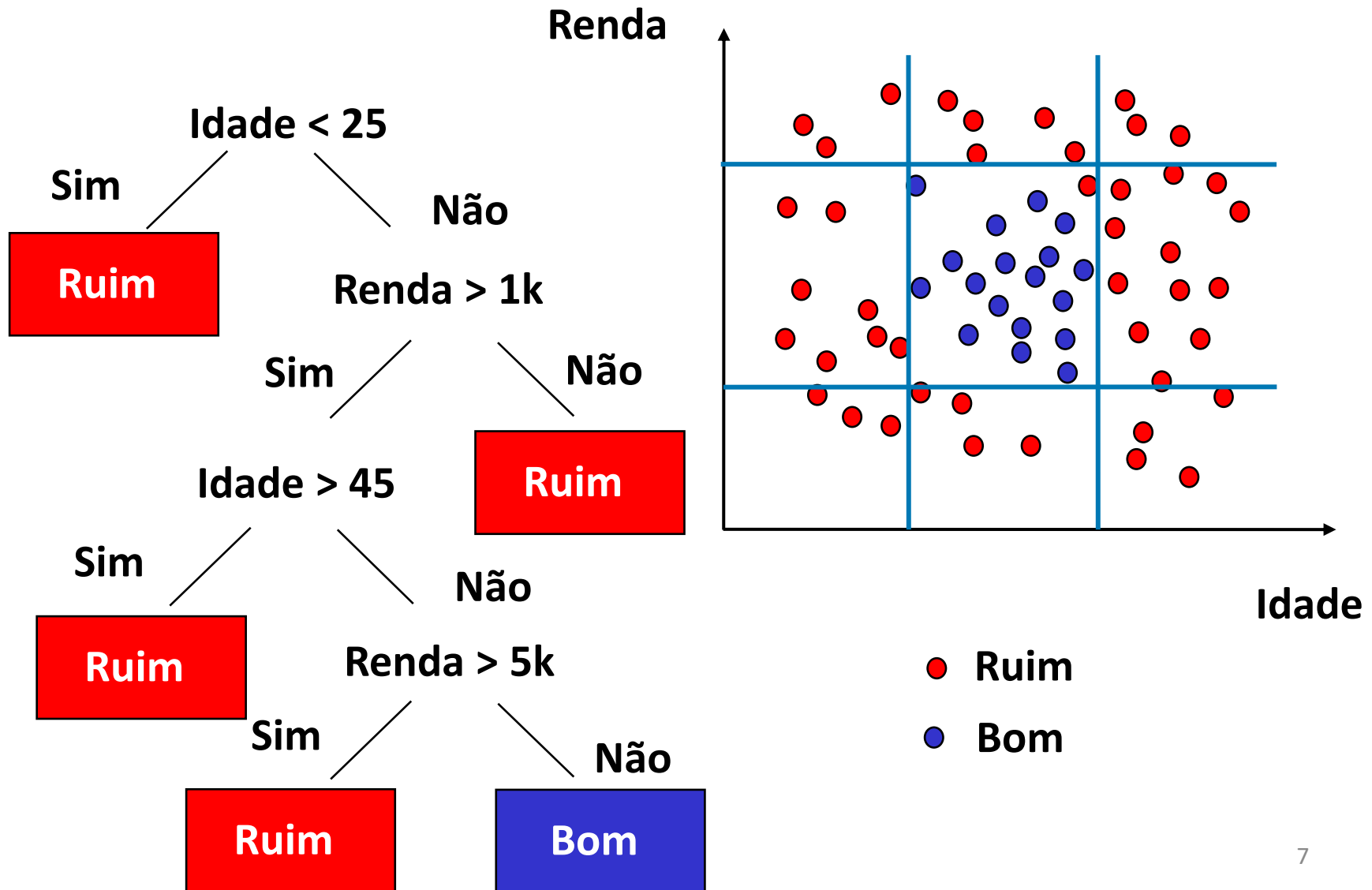
Avaliação de Modelos



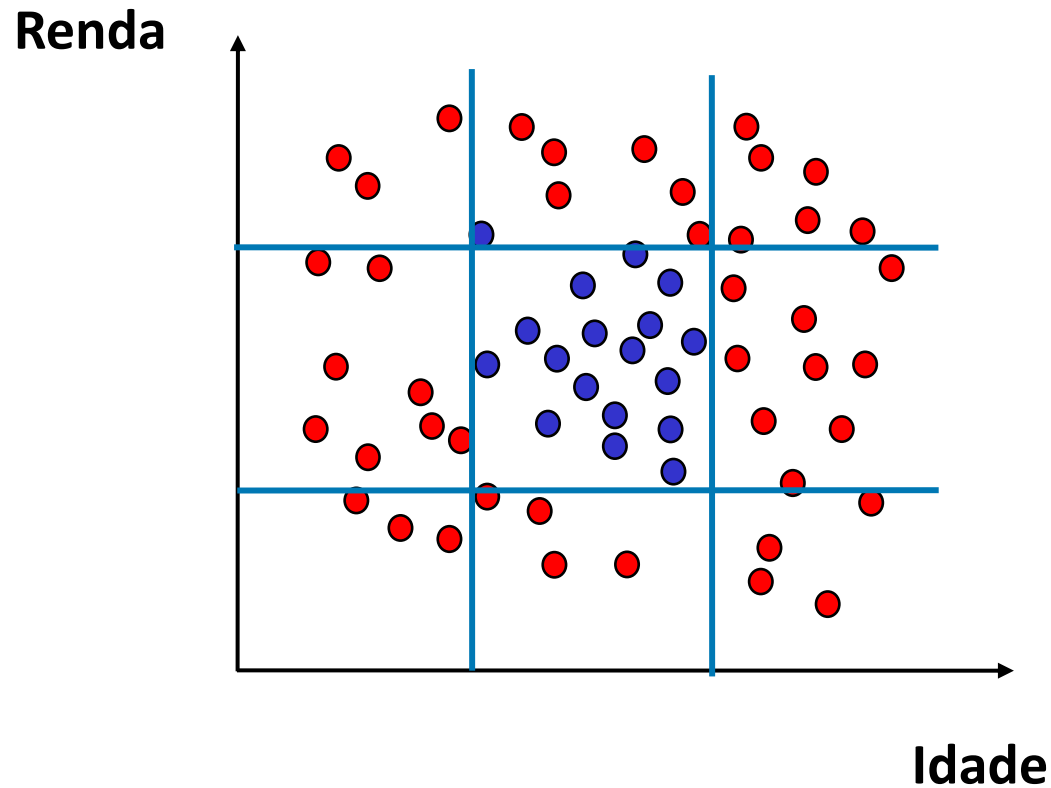
Avaliação Experimental de Hipóteses

- Classe Majoritária (CM):
 - Classe de maior ocorrência
- Erro da Classe Majoritária (ECM):
 - Erro cometido ao se atribuir um novo exemplo a ser classificado à CM
- Erro Aparente (EA):
 - Erro cometido ao se testar o classificador usando o conjunto de treinamento
- Erro Verdadeiro (EV):
 - Estimativa do desempenho futuro do classificador induzido utilizando o conjunto de treinamento com amostra **aleatória**, isto é, os exemplos não devem ser pré-selecionados

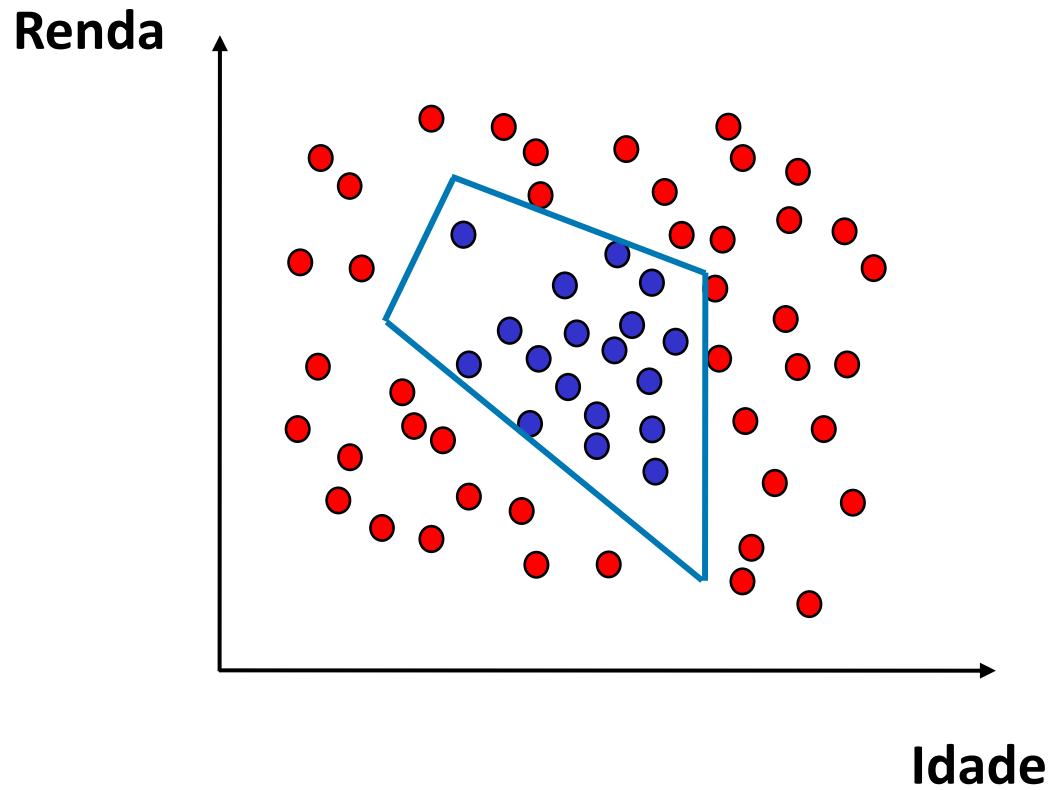
Árvore de Decisão



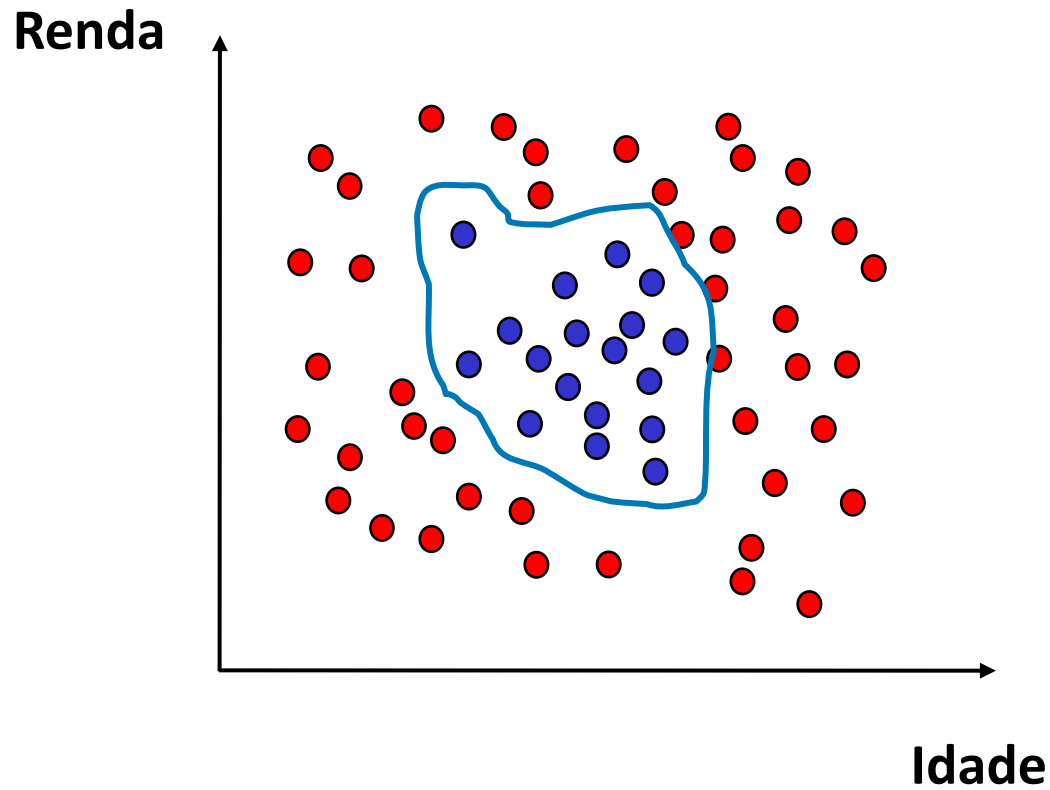
Árvore de Decisão (H1)



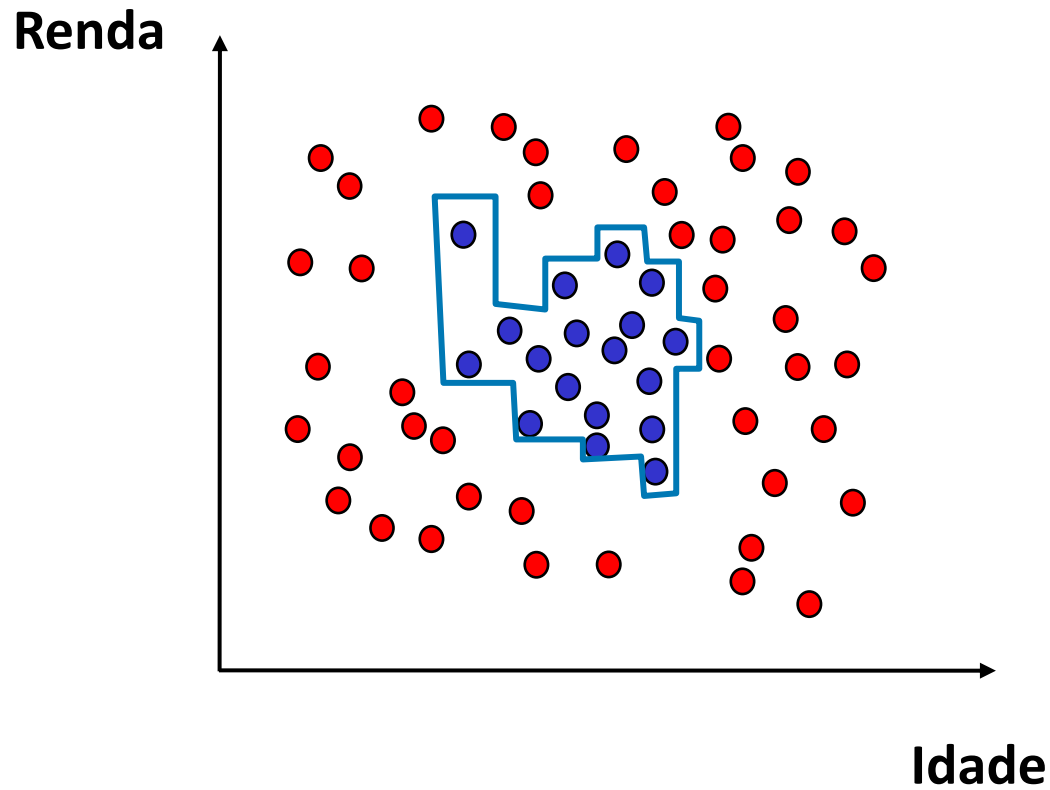
...Outra Possível H2



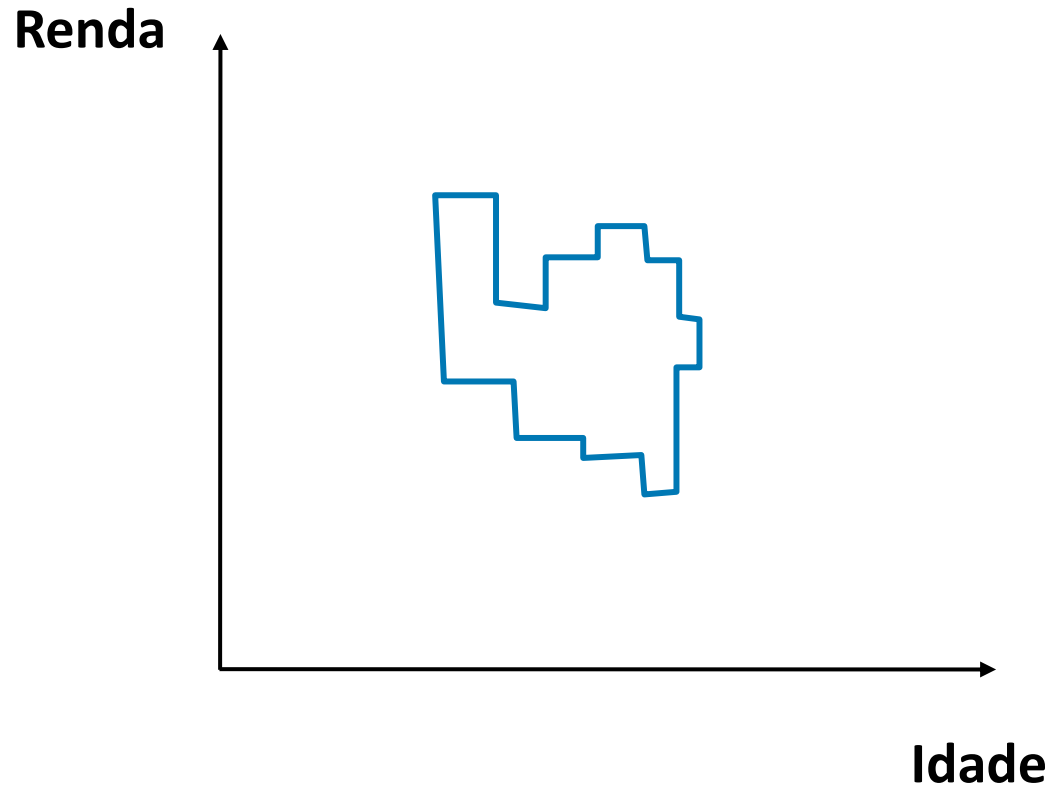
...Outra Possível H3



...Outra Possível H4



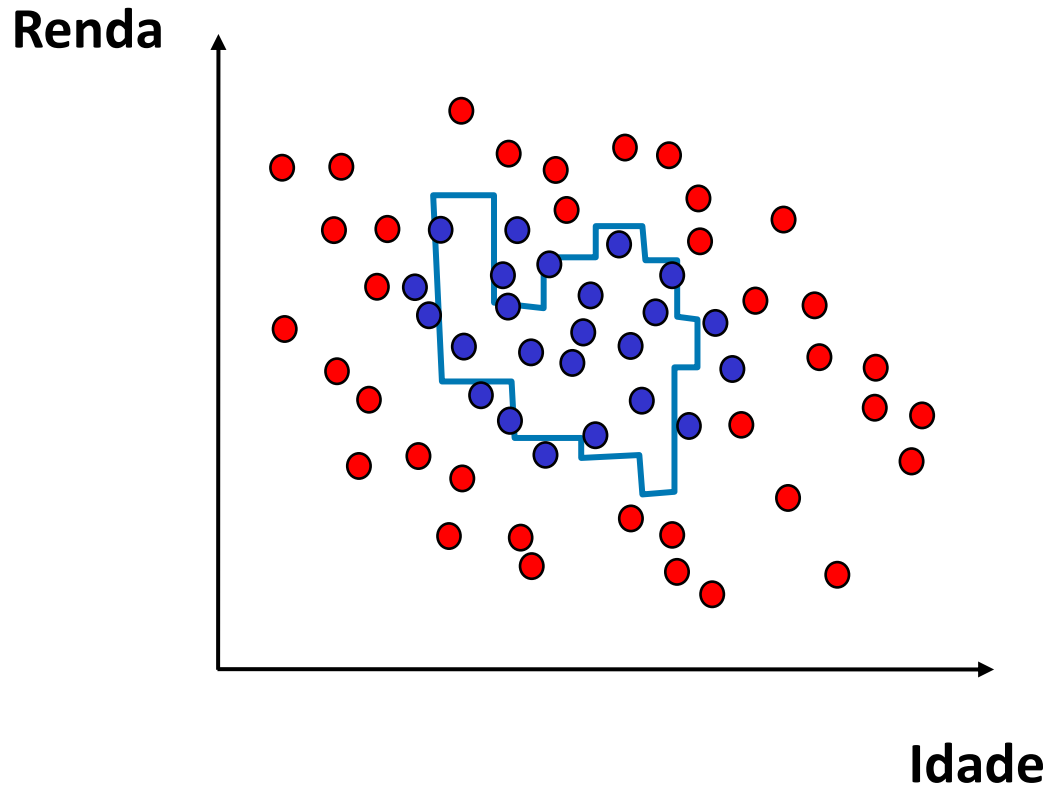
H4...



Avaliação em um conjunto de teste

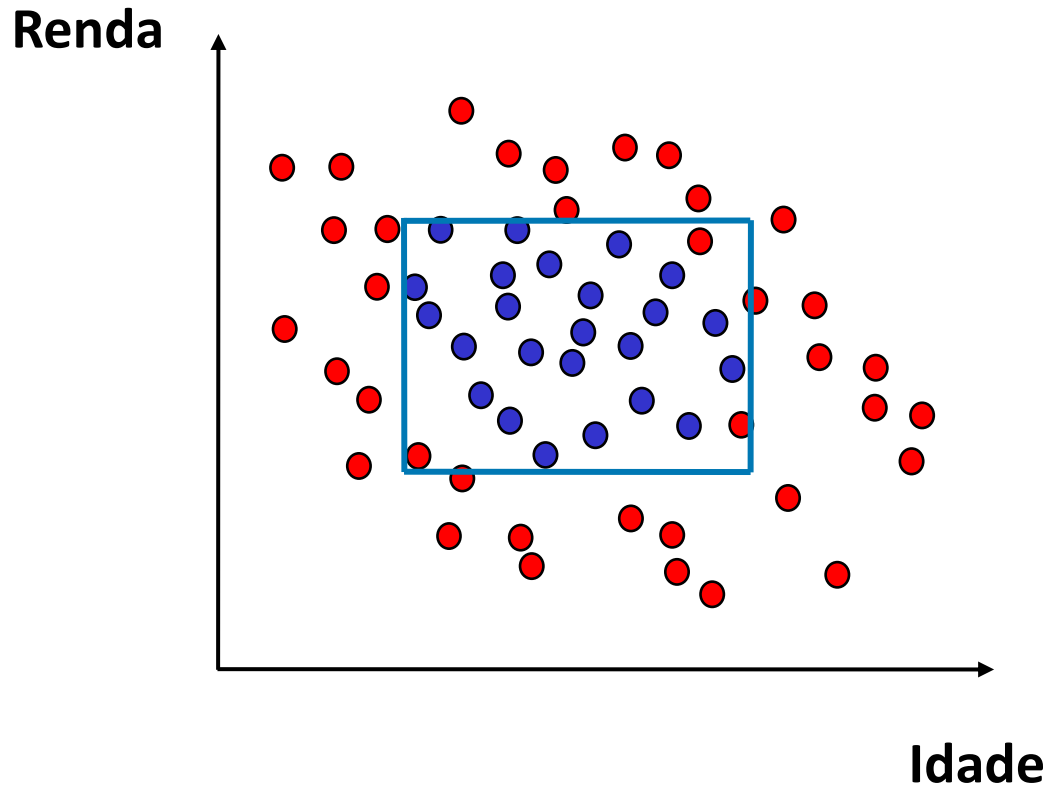
Erro de H4

Conjunto de Teste

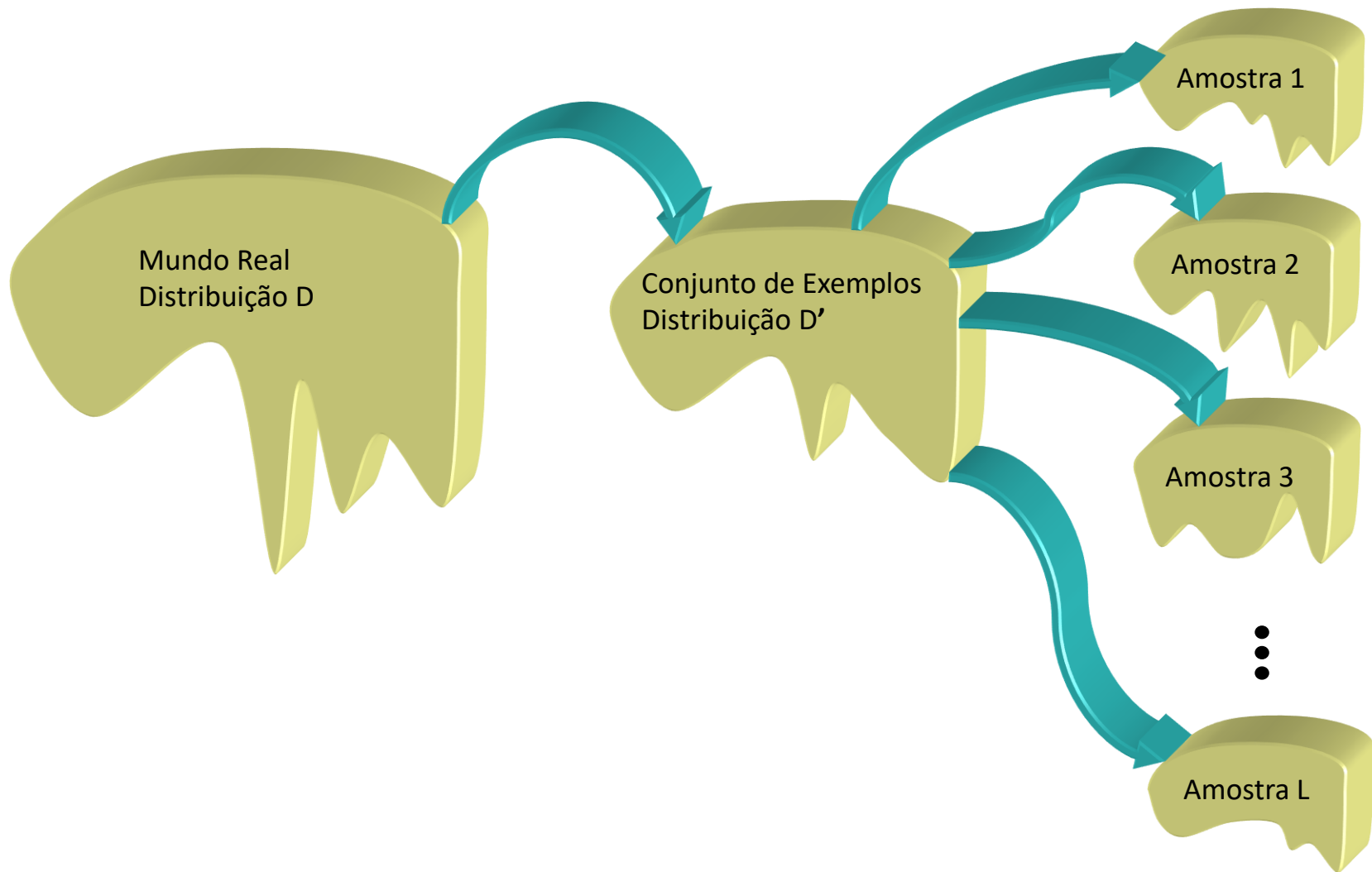


Erro de H1

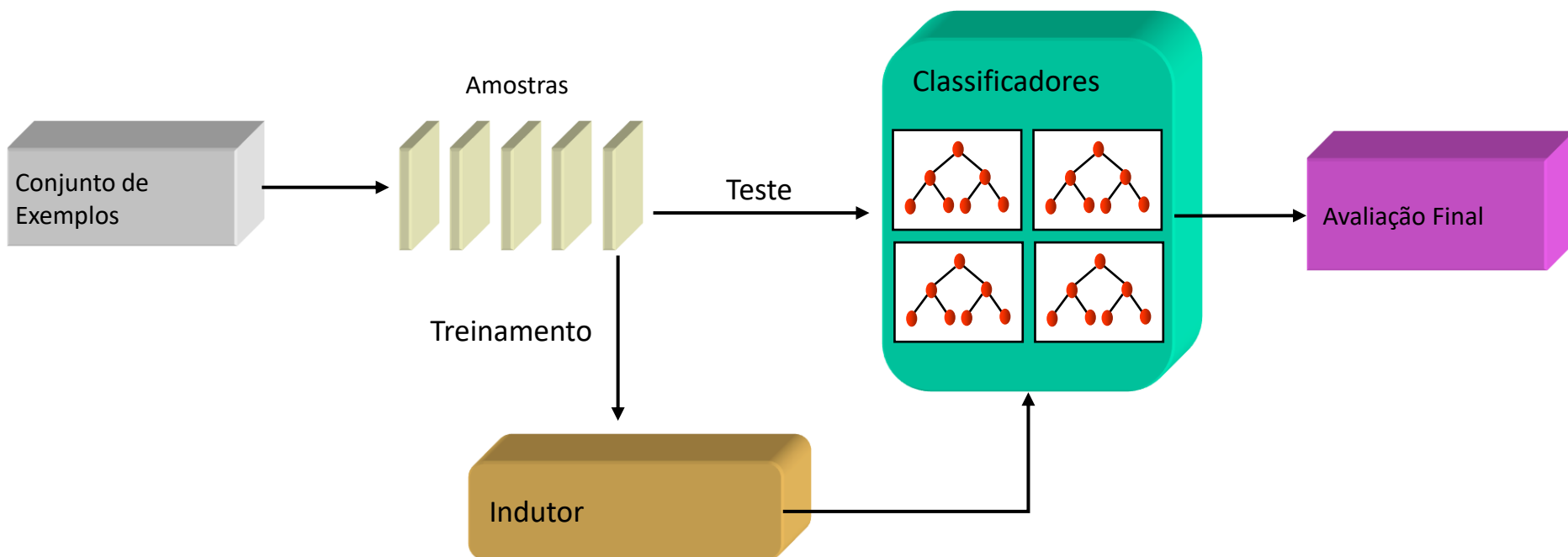
Conjunto de Teste



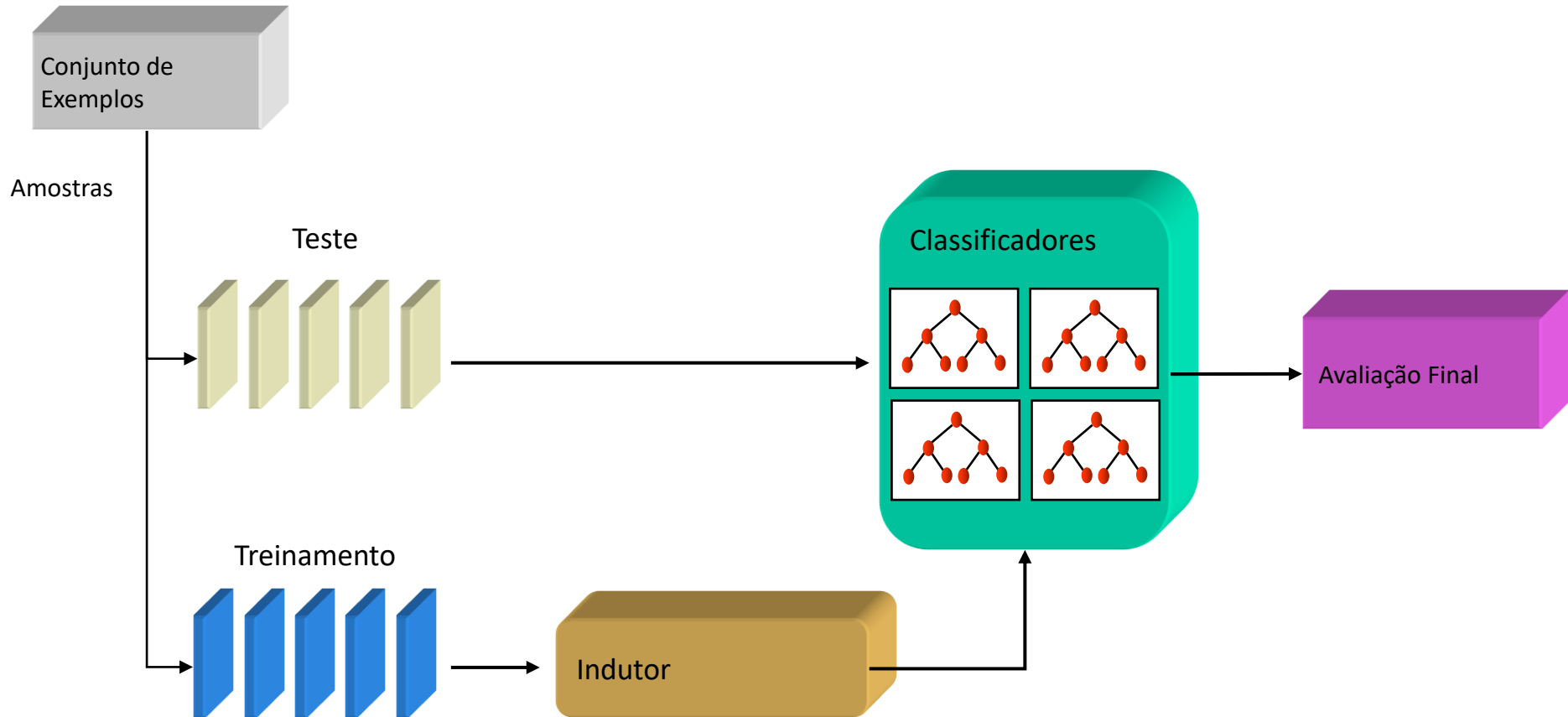
Métodos de Amostragem



Métodos de Amostragem (Resubstituição)

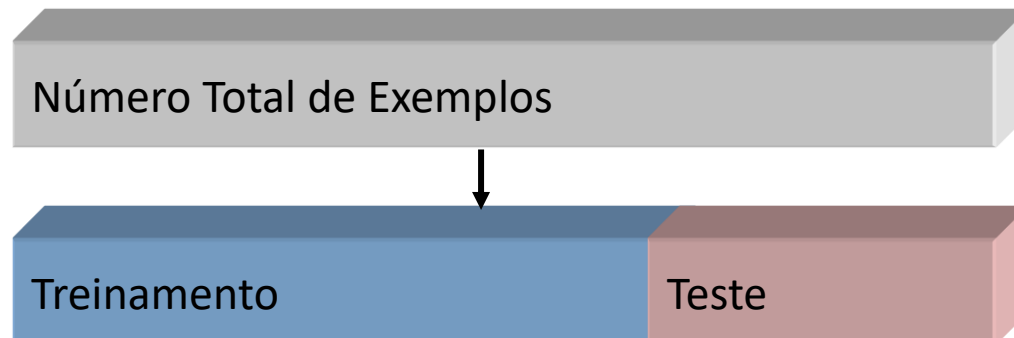


Métodos de Amostragem (Exceto Resubstituição)



Holdout

- Exemplos são divididos em uma porcentagem fixa de exemplos p para treinamento e $(1-p)$ para teste, considerando normalmente $p > 1/2$
- Valores típicos são $p = 2/3$ e $(1-p) = 1/3$, embora não existam fundamentos teóricos sobre estes valores
- n é o número total de exemplos



Holdout

- Este método tem a tendência de super estimar o erro verdadeiro
- Para pequenos conjuntos, nem sempre é possível separar uma parte dos exemplos

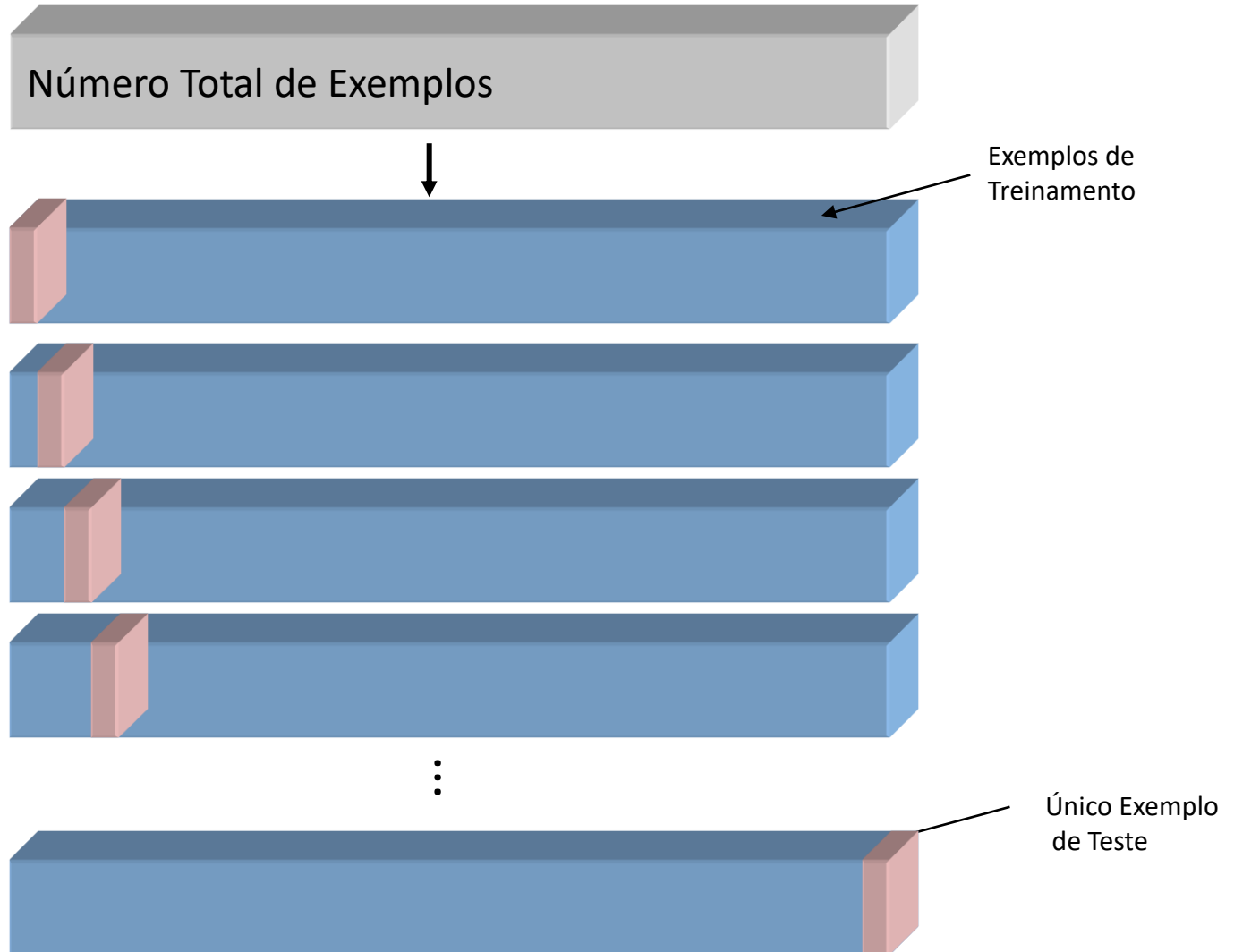
Método *Repeated Holdout*

- Estimativa do Holdout pode ser tornada mais confiável repetindo o processo com diferentes subamostras
 - A cada iteração, uma certa porção é randomicamente selecionada para treinamento (possivelmente com estratificação)
 - O erro é a média dos erros nas diferentes iterações

Leave-one-out

- Computacionalmente dispendioso e frequentemente é usado em amostras pequenas
- Para uma amostra de tamanho n :
 - uma hipótese é induzida utilizando $(n-1)$ exemplos
 - a hipótese é então testada no único exemplo remanescente
- Processo é repetido n vezes, cada vez induzindo uma hipótese deixando de considerar um único exemplo
- O erro é a soma dos erros em cada teste dividido por n

Leave-one-out



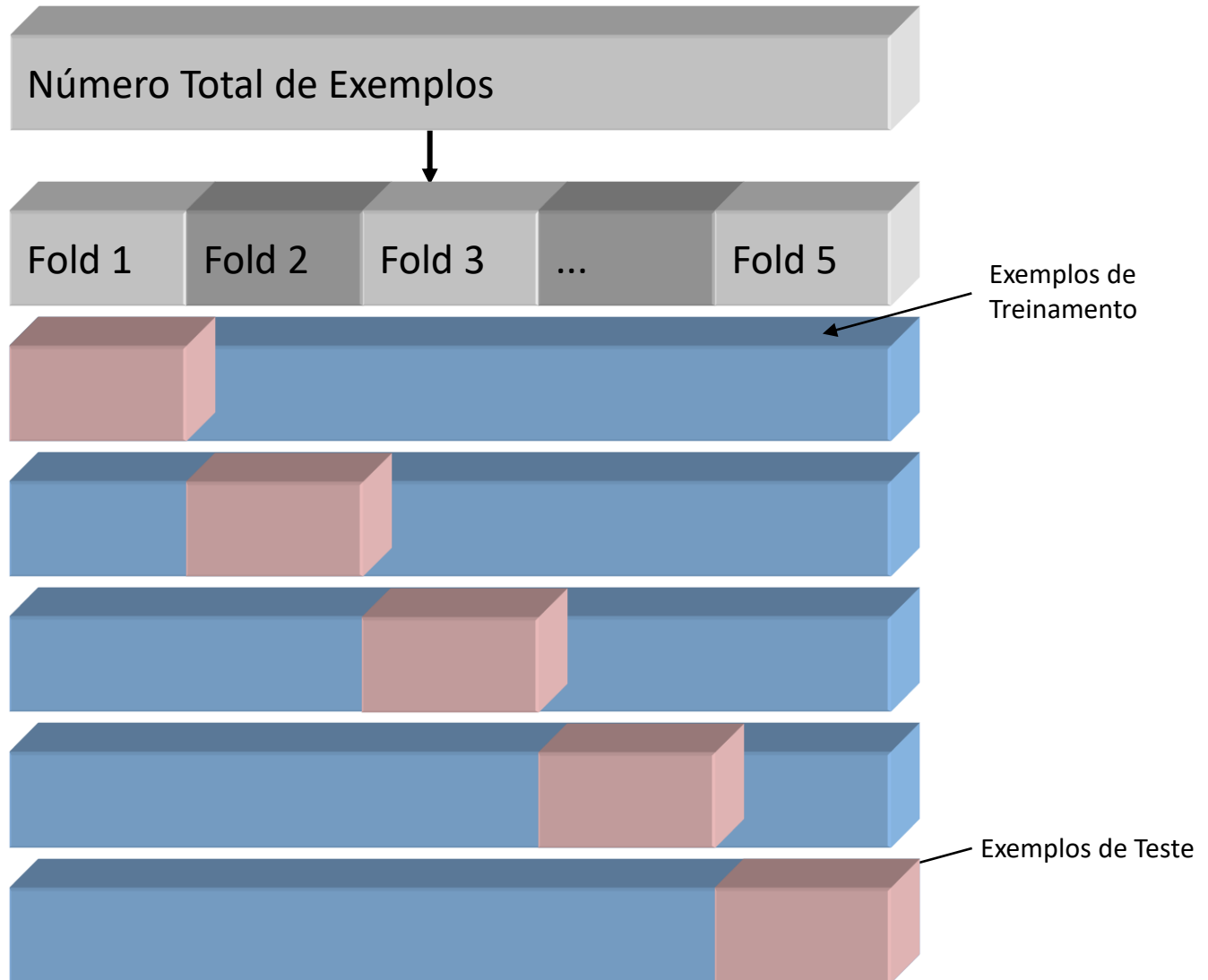
Cross-Validation

- Meio termo entre os estimadores Holdout e Leave-one-out
- **r**-fold cross-validation (CV):
 - exemplos são aleatoriamente divididos em **r** partições mutuamente exclusivas (folds) de tamanho aproximadamente igual a **n/r** exemplos
 - exemplos nos (**r**-1) folds são usados para treinamento e a hipótese induzida é testada no fold remanescente
- Processo é repetido **r** vezes, cada vez considerando um fold diferente para teste

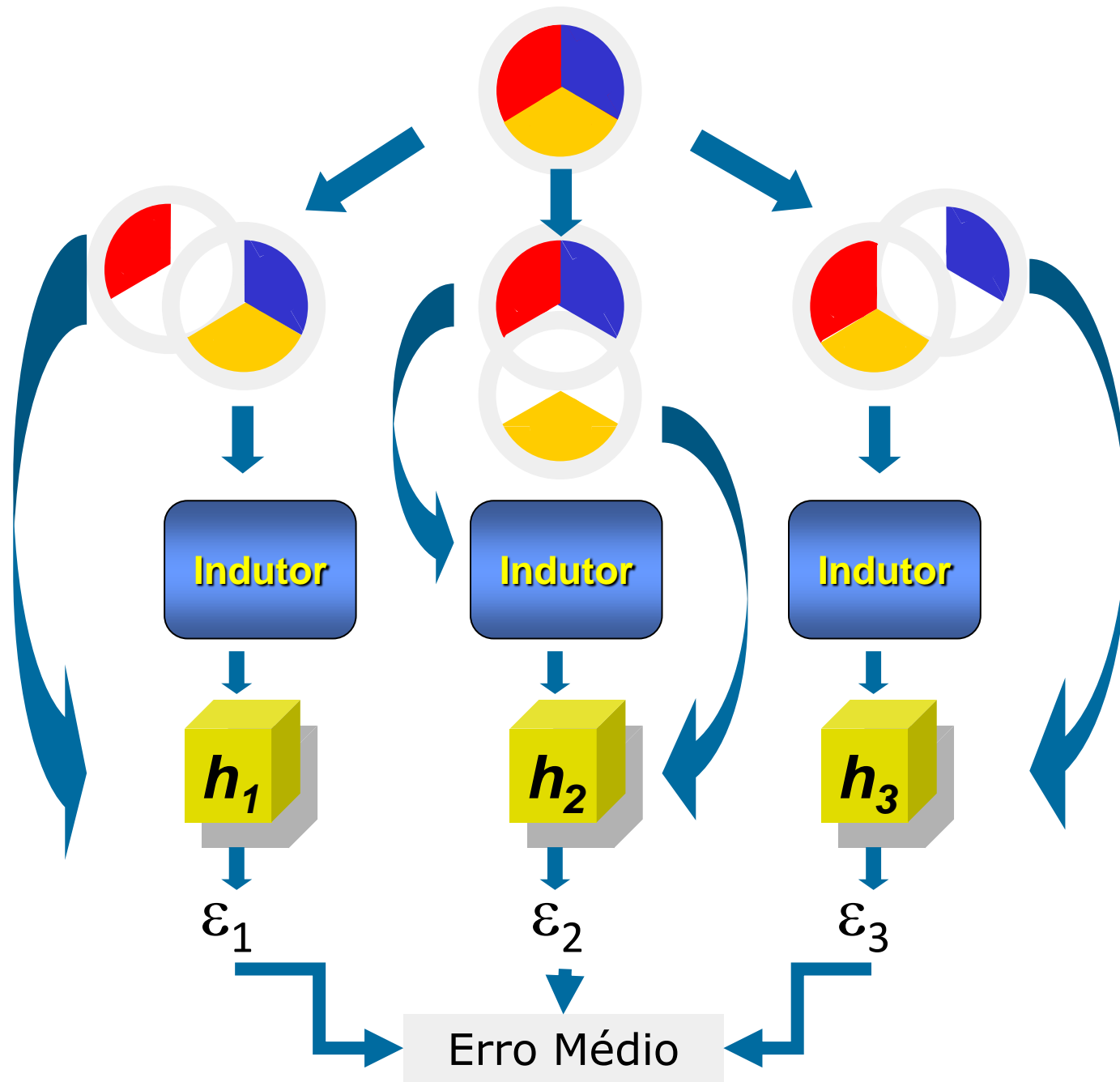
Cross-Validation

- Erro no cross-validation é a média dos erros calculados em cada um dos **r** folds
- Procedimento de rotação reduz tanto o *bias* inerente ao método de *Holdout* quanto o custo computacional do método *Leave-one-out*

Cross-Validation



Validação cruzada com 3 folds



Stratified Cross-Validation

- Similar ao *cross-validation*, mas ao gerar os *folds* mutuamente exclusivos, a distribuição de classe é considerada durante a amostragem
- Por exemplo:
 - conjunto original de exemplos possui duas classes com distribuição de 20% e 80%
 - cada *fold* também terá esta proporção de classes

Cross-Validation

- Número usual de folds: 10
- Porque 10?
 - Experimentos extensivos mostraram que, em geral, essa é uma boa escolha
- Estratificação reduz a variância na estimativa
- Melhor ainda: repetir stratified cross-validation
 - 10-fold cross-validation repetido 10 vezes e calculada a média

Avaliação de Modelos

Matriz de Confusão para um conjunto de dados com 2 classes e com n exemplos:

Rótulo do exemplo	Predito como +	Predito como -
+	5 (<i>TP</i>)	1 (<i>FN</i>)
-	2 (<i>FP</i>)	4 (<i>TN</i>)

Avaliação de Modelos

$$n = TP + FP + TN + FN$$

Erro do modelo:

$$Erro = \frac{FP + FN}{n}$$

Erro na classe:

$$Erro(+) = \frac{FN}{TP + FN}$$

$$Erro(-) = \frac{FP}{TN + FP}$$

Avaliação de Modelos

OBS: Bases com mais de duas classes

Rótulo do exemplo	Predito como C_1	Predito como C_2	Predito como C_3
C_1	5	0	1
C_2	0	4	0
C_3	1	1	4

Avaliação de Modelos

OBS: Bases com mais de duas classes

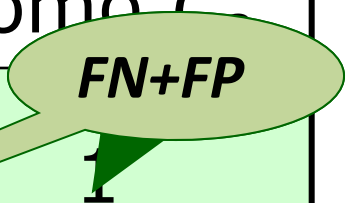
Rótulo do exemplo	Predito como C_1	Predito como C_2	Predito como C_3
C_1	5	0	1
C_2	0	4	0
	1	1	4

Como calcular o erro do modelo?

Avaliação de Modelos

OBS: Bases com mais de duas classes

Rótulo do exemplo	Predito como C_1	Predito como C_2	Predito como C_3
C_1	5	0	1
C_2	0	4	0
C_3	1	1	4



Ou seja, $FN+FP$ = soma do número de exemplos classificados erroneamente!

Avaliação de Modelos

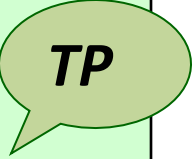
OBS: Bases com mais de duas classes

Rótulo do exemplo	Predito como C_1	Predito como C_2	Predito como C_3
C_1	5	0	1
C_2	0	4	0
	1	1	4

Como calcular o erro nas classes (Classe C_1 , por exemplo)?

Avaliação de Modelos

C_1 : Classe +

Rótulo do exemplo	Predito como C_1	Predito como C_2	Predito como C_3
C_1	5 	0	1
C_2	0	4	0
C_3	1	1	4

Avaliação de Modelos

C_1 : Classe +

Rótulo do exemplo	Predito como C_1	Predito como C_2	Predito como C_3
C_1	5	0	1
C_2	0	4	0
C_3	1	1	4

TN

Avaliação de Modelos

C_1 : Classe +

Rótulo do exemplo	Predito como C_1	Predito como C_2	Predito como C_3
C_1	5	0	1
C_2	0	4	0
C_3	1	1	4

FN

Avaliação de Modelos

C_1 : Classe +

Rótulo do exemplo	Predito como C_1	Predito como C_2	Predito como C_3
C_1	5	0	1
C_2	0	4	0
C_3	1	1	4

FP

Avaliação de Modelos

$$n = TP + FP + TN + FN$$

Erro do modelo:

$$Erro = \frac{FP + FN}{n}$$

Erro na classe:

$$Erro(+) = \frac{FN}{TP + FN}$$

$$Erro(-) = \frac{FP}{TN + FP}$$

Avaliação de Modelos

- Acurácia:

$$Acc = \frac{TP+TN}{n}$$

- Precisão:

$$Prec = \frac{TP}{TP+FP}$$

Avaliação de Modelos (Precisão)

$$Prec = \frac{TP}{TP + FP}$$

Rótulo do exemplo	Predito como +	Predito como -
+	5 (<i>TP</i>)	1 (<i>FN</i>)
-	2 (<i>FP</i>)	4 (<i>TN</i>)

Avaliação de Modelos

- Sensibilidade (Recall):

$$Sen = \frac{TP}{TP+FN}$$

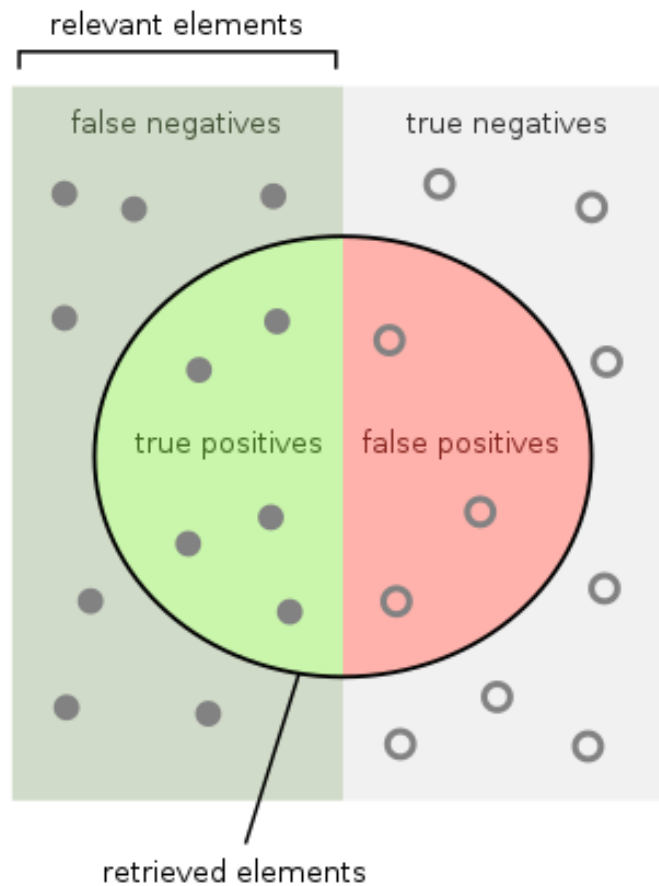
- Especificidade:

$$Esp = \frac{TN}{TN+FP}$$

Avaliação de Modelos (Sensibilidade/Recall)

$$Sen = \frac{TP}{TP + FN}$$

Rótulo do exemplo	Predito como +	Predito como -
+	5 (<i>TP</i>)	1 (<i>FN</i>)
-	2 (<i>FP</i>)	4 (<i>TN</i>)



How many retrieved items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are retrieved?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Avaliação de Modelos

- Escore F1:

$$F1 = \frac{Prec * Recall}{Prec + Recall}$$

- Equilíbrio entre Precisão e Recall
- Especialmente interessante quando há um desbalanço entre classes

Medidas de Qualidade de Regras

Tabela de Contingência:

usada para registrar observações independentes de duas ou mais variáveis aleatórias, normalmente qualitativas

	Fumante	Não fumante	Total
Desenvolveu	43	9	52
Não desenvolveu	44	4	48
Total	87	13	100

Avaliação de Regras

Classificadores Simbólicos podem ser avaliados como:

- Caixa-preta;
- Regras individuais:

If $\underbrace{cond_1 \text{ and } cond_2 \dots \text{ and } cond_m}_b$ **then** $\underbrace{\text{classe } C_k}_h$

Medidas de Qualidade de Regras

Tabela de Contingência:

	classe H	classe não H	
R cobre	$b \ h$	$b \ \sim h$	b
R não cobre	$\sim b \ h$	$\sim b \ \sim h$	$\sim b$
	h	$\sim h$	n

Medidas de Avaliação de Regras

- *Cobertura*
 - $Cov(R) = b/n$
- *Precisão*
 - $Prec(R) = bh/b$
- *Sensibilidade (ou Recall)*
 - $Sens(R) = bh/h$
- *Especificidade*
 - $Esp(R) = \sim b \sim h / \sim h$
- *Novidade*
 - $Nov(R) = 1/n (bh - bh/n)$

Medidas de Avaliação de Regras

- *Cobertura*
 - $Cov(R) = b/n$
- *Precisão*
 - $Prec(R) = bh/b$
- *Sensibilidade (ou Recall)*
 - $Sens(R) = bh/h$
- *Especificidade*
 - $Esp(R) = \sim b \sim h / \sim h$
- *Novidade*
 - $Nov(R) = 1/n (bh - bh/n)$

Número de exemplos cobertos pela regra

Medidas de Avaliação

Quanto uma regra é específica
para o problema
(dos exemplos cobertos pela regra, quantos são
cobertos corretamente)

- Cobertura
 - $Cov(R) = b/n$
- Precisão
 - $Prec(R) = bh/b$
- Sensibilidade (ou *Recall*)
 - $Sens(R) = bh/h$
- Especificidade
 - $Esp(R) = \sim b \sim h / \sim h$
- Novidade
 - $Nov(R) = 1/n (bh - bh/n)$

Medidas de Avaliação de Regras

- *Cobertura*
 - $Cov(R) = b/n$
- *Precisão*
 - $Prec(R) = bh/b$
- *Sensibilidade (ou Recall)*
 - $Sens(R) = bh/h$
- *Especificidade*
 - $Esp(R) = \sim b \sim h / \sim h$
- *Novidade*
 - $Nov(R) = 1/n (bh - bh/n)$

Número de exemplos da classe h que são cobertos por R

(mede a fração de Verdadeiros Positivos que são corretamente classificados)

Medidas de Avaliação de Regras

- *Cobertura*
 - $Cov(R) = b/n$
- *Precisão*
 - $Prec(R) = bh/b$
- *Sensibilidade (ou Recall)*
 - $Sens(R) = bh/h$
- *Especificidade*
 - $Esp(R) = \sim b \sim h / \sim h$
- *Novidade*
 - $Nov(R) = 1/n (bh - bh/n)$

Equivalente
ao Recall, mas para
exemplos
que NÃO são cobertos
pela regra

Medidas de Avaliação de Regras

- Cobertura
 - $Cov(R) = b/n$
- Precisão
 - $Prec(R) = bh/b$
- Sensibilidade (ou Recall)
 - $Sens(R) = bh/h$
- Especificidade
 - $Esp(R) = \sim b \sim h / \sim b$
- Novidade
 - $Nov(R) = 1/n (bh - bh/n)$

Mede a probabilidade de b e h ocorrerem juntos dado que b e h não são estatisticamente independentes

Aproveitando ao Máximo os Dados

- Uma vez que a avaliação esteja completa, todos os dados podem ser utilizados para construir o classificador final
- Em geral:
 - Quanto maior o conjunto de dados, melhor o classificador
 - Quanto maior o conjunto de teste, mais acurada a estimativa do erro

Comparando Diversos Modelos

- Questão frequente: Qual modelo possui melhor desempenho?
 - É dependente de domínio
 - Caminho óbvio: usar estimativas 10-fold CV
 - Problema: variância na estimativa
 - Variância pode ser reduzida com CV repetidos
- Porém, ainda não sabemos se os resultados são **confiáveis**

Testes de Significância (TS)

- Nos dizem quão confiantes podemos ser de que realmente há uma diferença entre os modelos
 - *Null hypothesis*: não há “real” diferença
 - *Alternative hypothesis*: há uma diferença
- TS medem quanto de evidência há a favor da rejeição da hipótese nula

Testes de Significância (TS)

Exemplo:

- 03-fold CV repetidos 02 vezes
- Queremos saber se as duas médias são significativamente diferentes
- Neste caso podemos usar:
 - Teste pareado t-student, se dados forem paramétricos e pareados
 - Teste Mann-Whitney, se dados forem não paramétricos e não pareados
 - Teste Willcoxon, se dados forem não paramétricos e pareados

Custos Diferentes

- Na prática, verdadeiros positivos (TP) e falso negativos (FN) em geral representam custos diferentes
- Exemplos:
 - Testes de diagnóstico médico: o paciente possui leucemia?
 - Decisão de empréstimos: deve-se aprovar o empréstimo para o cliente?
 - Mineração na Web: o usuário irá clicar neste link?
 - Propaganda direcionada: o cliente irá comprar esse produto?
 - ...

Critério para Seleção de Modelos

- Encontrar boa proporção entre:
 - Complexidade do modelo
 - Acurácia
- Raciocínio: um bom modelo é um modelo simples que permite alto desempenho
- Occam's Razor :
a melhor teoria é a “menor” que descreve todos os fatos

William of Ockham, born in the village of Ockham in Surrey (England) about 1285, was the most influential philosopher of the 14th century and a controversial theologian.



- slides baseados em apresentações de:
 - Profa. Maria Carolina Monard
 - Prof. José Augusto Baranauskas
 - Profa. Huei Diana Lee
 - Prof. Ronaldo Cristiano Prati.
 - Prof. Gustavo E.A.P.A. Batista
 - Profa. Bianca Zadrozny
 - Prof. G. Piatetsky-Shapiro