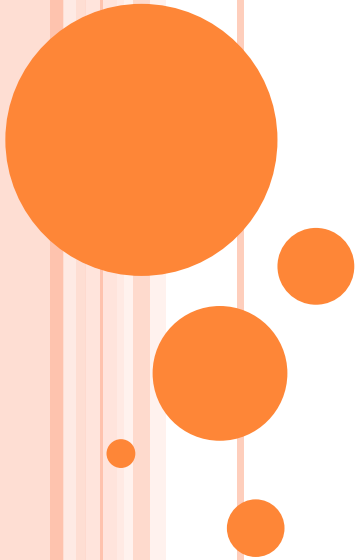


AVALIAÇÃO DA APRENDIZAGEM

DIN4034 – Aprendizagem de Máquina

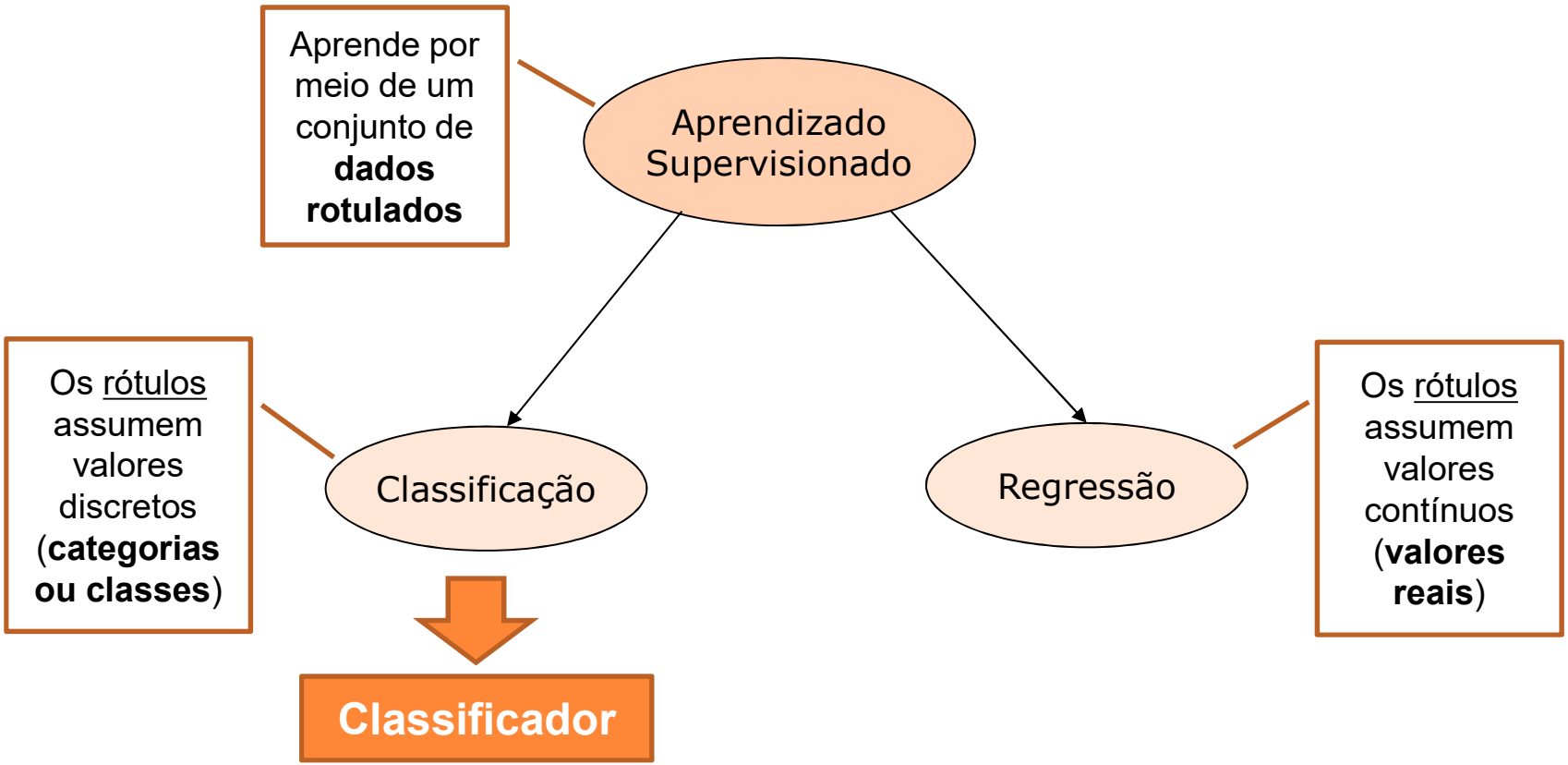
Profa. Dra. Valéria Delisandra Feltrim

PCC – DIN – UEM

1º sem/2016

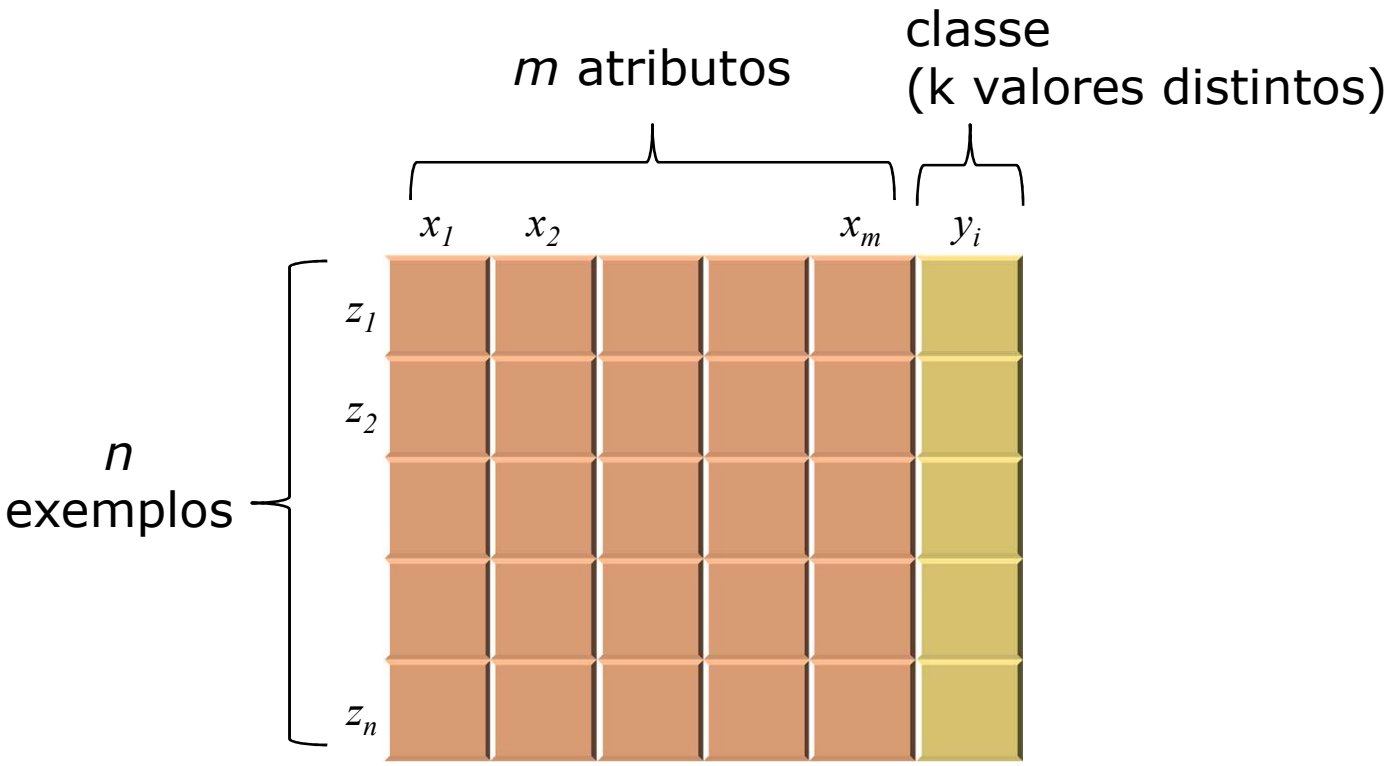
Slides preparados com base no material do Prof. José A. Baranauskas (DCM-FFCLRP-USP) e dos Profs. Maria Carolina Monard e Gustavo Batista (ICMC-USP)

RELEMBRANDO 1



RELEMBRANDO 2

- Conjunto de exemplos para aprendizado supervisionado



Cada exemplo é um vetor $\overrightarrow{z_i} = (\overrightarrow{x_i}, y_i)$

MEDIDAS DE AVALIAÇÃO

- Algumas medidas são específicas de um conjunto de exemplos particular → independentes do classificador induzido
  - Distribuição de classes
  - Prevalência de classe
  - Erro majoritário
- Outras medidas dependem tanto do conjunto de exemplos como do classificador induzido
  - Taxa de erro/acerto
  - Precisão
  - Cobertura
  - ...

MEDIDAS DE AVALIAÇÃO

- **Distribuição de classes** → dá a proporção de cada classe no conjunto de exemplos
- Para cada classe C_i no conjunto T sua distribuição é dada por

$$distr(C_i) = \frac{1}{n}\sum_{i=1}^{n} \| y_i = C_i \|$$

- Exemplo
  - Um conjunto com 100 exemplos que possui 60 exemplos da classe *A*, 15 exemplos da classe *B* e 25 exemplos da classe *C*, tem a seguinte distribuição de classes:
    - **distr(*A*, *B*, *C*)** = (0,60, 0,15, 0,25) = (60%, 15%, 25%)
  - A classe *A* é a classe **majoritária** ou **prevalente**
  - A classe *B* é a classe **minoritária**

MEDIDAS DE AVALIAÇÃO

- O **balanceamento das classes** no conjunto de exemplos é um aspecto muito importante
- Suponha um conjunto de exemplos *T* com a seguinte distribuição de classes
  - distr(C_1, C_2, C_3) = (99%, 0,25%, 0,75%)
  - **Prevalência da classe C_1**
- Um classificador simples que sempre classifique novos exemplos como pertencentes à classe majoritária C_1 acertaria 99% das vezes
  - err(*h*) = 0,01 e acc(*h*) = 0,99
  - maj-err(T) = 0,01

MEDIDAS DE AVALIAÇÃO

- **Erro majoritário** → denotado por **maj-err(*T*)**, é calculado com base na distribuição de classes em um conjunto de exemplos *T*
- Sabendo-se a distribuição de classes do conjunto de exemplos *T*, pode-se calcular seu erro majoritário:

$$maj - err(T) = 1 - \max_{i=1,\ldots,k} distr(C_i)$$

  - Para o exemplo do slide anterior, o erro majoritário é maj-err(*T*) = 1 – 0,99 = 0,01 = 1%

- O erro majoritário é **independente do algoritmo de aprendizado** utilizado
  - Ele fornece um limiar abaixo do qual o erro de um classificador deve ficar

MEDIDAS DE AVALIAÇÃO

- Uma medida de desempenho comumente usada é a **taxa de erro** de um classificador ***h***, denotada por **err(*h*)**
- Usualmente, a taxa de erro é obtida comparando-se a classe verdadeira de cada exemplo com a classe atribuída pelo classificador induzido

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

$$err(h) = \frac{exemplos\_\boldsymbol{incorretamente}\_classificados}{total\_exemplos\_classificados}$$

MEDIDAS DE AVALIAÇÃO

- Considere um classificador h_1. Suponha que dentre 10 exemplos submetidos à h_1, 7 foram classificados corretamente e 3 foram receberam classes erradas
  - Então err(h_1) = $\frac{3}{10} = 0{,}3$ ou 30%
- O complemento da taxa de erro, chamada de **taxa de acerto** ou **acurácia** ou **precisão**, denotada por **acc(*h*)**, é o número de acertos do classificador dividido pelo número total de exemplos classificados

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

$$acc(h) = \frac{exemplos\_\boldsymbol{corretamente}\_classificados}{total\_exemplos\_classificados}$$

  - Se err(h_1) = 0,3, então acc(h_1) = 0,7 ou 70%

EXEMPLO

- Número de exemplos?
- Número de classes?
- Distribuição de classes?
- Classe prevalente ou majoritária?
- Classe minoritária?
- Erro majoritário?

Cabeça X_1	Peso X_2	Sorri X_3	Classe Y=f(x)
redonda	10.0	não	amigo
triangular	12.0	sim	amigo
redonda	5.6	sim	amigo
quadrada	11.0	não	chato
quadrada	10.0	sim	amigo
triangular	5.5	não	inimigo
redonda	5.7	sim	chato
quadrada	15.3	sim	chato
quadrada	10.2	sim	amigo
redonda	5.0	não	inimigo

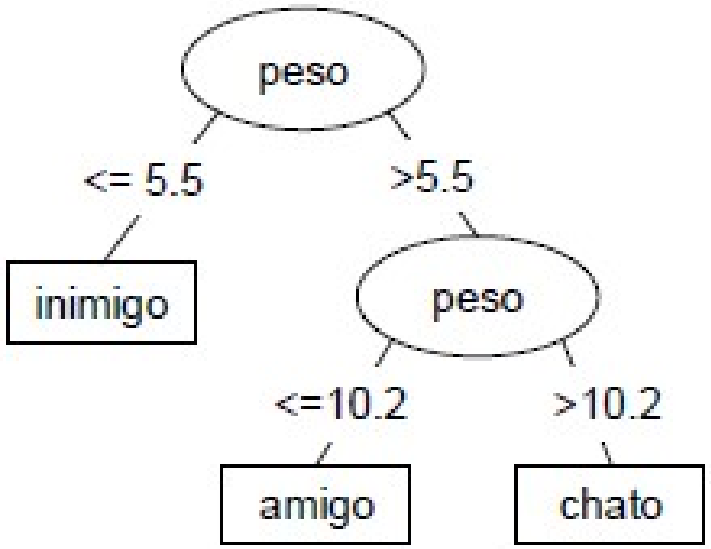
EXEMPLO

- Número de exemplos? N = 10
- Número de classes? k = 3
  - C1=amigo; C2=chato; C3=inimigo
- Distribuição de classes?
  - distr(amigo) = 5/10 = 50%
  - distr(chato) = 3/10 = 30%
  - distr(inimigo) = 2/10 = 20%
- Classe *amigo* é a classe majoritária
- Classe *inimigo* é a classe minoritária
- Erro majoritário?
  - *maj-err(T)* = 1 – 5/10 = 50%

Cabeça X_1	Peso X_2	Sorri X_3	Classe Y=f(x)
redonda	10.0	não	amigo
triangular	12.0	sim	amigo
redonda	5.6	sim	amigo
quadrada	11.0	não	chato
quadrada	10.0	sim	amigo
triangular	5.5	não	inimigo
redonda	5.7	sim	chato
quadrada	15.3	sim	chato
quadrada	10.2	sim	amigo
redonda	5.0	não	inimigo

EXEMPLO

- Seja h(x)



Cabeça X_1	Peso X_2	Sorri X_3	Classe Y=f(x)	Predita $\hat{Y}$=h(x)
redonda	10.0	não	amigo	amigo
triangular	12.0	sim	amigo	chato
redonda	5.6	sim	amigo	amigo
quadrada	11.0	não	chato	chato
quadrada	10.0	sim	amigo	amigo
triangular	5.5	não	inimigo	inimigo
redonda	5.7	sim	chato	amigo
quadrada	15.3	sim	chato	chato
quadrada	10.2	sim	amigo	amigo
redonda	5.0	não	inimigo	inimigo

- err(h) = 2/10 = 20%
- acc(h) = 1 – 2/10 = 80%

Medidas de Avaliação

- **Matriz de confusão** de um classificador *h*
  - Mostra o número de **classificações "verdadeiras"** ***versus*** **as classificações preditas** para cada classe C_i, sobre um conjunto de exemplos *T*
  - Os resultados são apresentados em duas dimensões: classes "verdadeiras" e classes preditas, para *k* classes diferentes
  - Cada elemento da matriz fora da diagonal principal representa o número de exemplos de *T* que pertencem à classe C_i, mas foram classificados como sendo da classe C_j

MATRIZ DE CONFUSÃO: EXEMPLO

		Respostas do classificador							
		C_1	C_2	C_3	C_4	C_5	C_6	C_7	Total
Respostas verdadeiras	C_1	**57**	10	2	1	7	0	0	77
	C_2	11	**23**	0	0	2	0	0	36
	C_3	6	1	**49**	0	8	1	0	65
	C_4	5	0	0	**26**	14	0	0	45
	C_5	2	2	0	9	**101**	3	0	117
	C_6	0	0	0	0	9	**10**	1	20
	C_7	0	0	0	0	5	1	**0**	6
	Total	81	36	51	36	146	15	1	**366**

MEDIDAS DE AVALIAÇÃO

- O **número de acertos**, para cada classe, se localiza na ***diagonal principal*** da matriz
  - Os demais elementos representam **erros** de classificação
- A matriz de confusão de um *classificador ideal* possui todos os elementos fora da diagonal principal iguais a zero
- A partir da matriz de confusão, outras medidas podem ser obtidas, em especial:
  - ***Precision*** (Precisão) e ***Recall*** (Cobertura ou Abrangência)

MATRIZ DE CONFUSÃO

- *h*: if X1= a and X2 = s then classe = + else classe = –

	Atributos				
Exemplo	X_1	X_2	X_3	Classe (Y)	h
z_1	a	s	2	+	+
z_2	a	s	1	-	+
z_3	b	n	1	+	-
z_4	b	s	2	-	-
z_5	c	n	2	+	-

		Predita		
	Classe	+	–	Total
verdadeira	+	1	2	3
	–	1	1	2
	Total	2	3	5

MATRIZ DE CONFUSÃO

		Classe predita por h				
		C_1	C_2	...	C_k	$M(C_i,*)=\sum_{j=1}^{k} M(C_i,C_j)$
Classe Verdadeira	C_1	$M(C_1,C_1)$	$M(C_1,C_2)$	...	$M(C_1,C_k)$	$M(C_1,*)$
	C_2	$M(C_2,C_1)$	$M(C_2,C_2)$	...	$M(C_2,C_k)$	$M(C_2,*)$
	...	...	...	...	...	...
	C_k	$M(C_k,C_1)$	$M(C_k,C_2)$	...	$M(C_k,C_k)$	$M(C_k,*)$
		$M(*,C_1)$	$M(*,C_2)$	...	$M(*,C_k)$	n

$$M(*,C_j)=\sum_{i=1}^{k} M(C_i,C_j) \qquad n=\sum_{i=1}^{k} M(C_i,*)=\sum_{i=1}^{k} M(*,C_i)$$

MEDIDAS DE AVALIAÇÃO

- ***Precision*** e ***Recall*** medem o desempenho do classificador para cada classe, separadamente
- Dada uma classe ***C*:**
  - ***Precision*** (Precisão) é o total de exemplos corretamente classificados como ***C*** sobre o total de exemplos classificados como ***C***
  - ***Recall*** (Cobertura) é o total de exemplos corretamente classificados como ***C*** sobre o total de exemplos pertencentes à classe ***C*** presentes no conjunto *T*

PRECISION E RECALL

- *Precision*: dos exemplos classificados como *C*, quantos foram classificados corretamente?

$$\Pr ecision(C_i) = \frac{M(C_i, C_i)}{M(*, C_i)} = \frac{corretamente\_reconhecidos\_C}{total\_reconhecidos\_C}$$

- *Recall*: dos exemplos que deveriam ser classificados como *C*, quantos foram classificados corretamente?

$$\mathrm{Re}\, call(C_i) = \frac{M(C_i, C_i)}{M(C_i, *)} = \frac{corretamente\_reconhecidos\_C}{total\_exemplos\_C}$$

MEDIDAS DE AVALIAÇÃO

- ***F-measure***: combinação das medidas *Precision* e *Recall*, sendo forma conveniente de expressá-las como um único valor
  - Média harmônica das medidas *precision* e *recall*

$$F - measure(C) = \frac{2 \times recall(C) \times precision(C)}{recall(C) + precision(C)}$$

- ***Macro-F***: média aritmética das *F-measures* de todas as categorias

$$Macro - F(h) = \frac{1}{k} \sum_{i=1}^{k} F - measure(C_i)$$

MATRIZ DE CONFUSÃO

- *h*: if X1= a and X2 = s then classe = + else classe = –

	Atributos				
Exemplo	X_1	X_2	X_3	Classe (Y)	h
z_1	a	s	2	+	+
z_2	a	s	1	-	+
z_3	b	n	1	+	-
z_4	b	s	2	-	-
z_5	c	n	2	+	-

Classe	Predita +	Predita -	Total
Verdadeira +	1	2	3
Verdadeira -	1	1	2
Total	2	3	5

$$\text{Precision}(+) = \frac{M(+,+)}{M(*,+)} = \frac{1}{2} = 0{,}5$$

$$\text{Precision}(-) = \frac{M(-,-)}{M(*,-)} = \frac{1}{3} = 0{,}33$$

$$\text{Recall}(+) = \frac{M(+,+)}{M(+,*)} = \frac{1}{3} = 0{,}33$$

$$\text{Recall}(-) = \frac{M(-,-)}{M(-,*)} = \frac{1}{2} = 0{,}5$$

$$F-measure(+) = \frac{2 \times 0{,}33 \times 0{,}5}{0{,}33 + 0{,}5} = 0{,}398$$

$$F-measure(-) = \frac{2 \times 0{,}5 \times 0{,}33}{0{,}5 + 0{,}33} = 0{,}398$$

$$Macro-F(h) = \frac{0{,}398 + 0{,}398}{2} = 0{,}398$$

MATRIZ DE CONFUSÃO

- Matriz de confusão para problemas binários (apenas duas classes)

Classe	Predita +	Predita -
Verdadeira +	**TP**	**FN**
Verdadeira -	**FP**	**TN**

$$Precision(+) = \frac{TP}{TP + FP}$$

$$Precision(-) = \frac{TN}{TN + FN}$$

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

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

TP (*True Positive*) = Exemplos positivos classificados como positivos
FP (*False Positive*) = Exemplos negativos classificados como positivos
FN (*False Negative*) = Exemplos positivos classificados como negativos
TN (*True Negative*) = Exemplos negativos classificados como negativos

MEDIDAS DE AVALIAÇÃO

- ***Kappa***: mede a concordância entre *k* juízes sobre *n* exemplos

$$kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

$$P(A) = \frac{1}{n} \sum_{i=1}^{k} M(C_i, C_i)$$

$$P(E) = \frac{1}{n^2} \sum_{i=1}^{k} M(C_i, *) \times M(*, C_i)$$

- *P(A)* estima a concordância observada
- *P(E)* estima a concordância esperada ao acaso
- O valor *Kappa* varia entre -1 e 1
  - -1 indica discordância sistemática entre os juízes
  - 0 indica concordância ao acaso
  - 1 indica concordância perfeita entre os juízes
- Geralmente, considera-se que há um alto índice de concordância quando *Kappa* é superior a 0,8, mas esse valor varia de uma tarefa para outra

MEDIDAS DE AVALIAÇÃO

- As medidas vistas até agora servem para problemas de classificação → acerto ou erro
- Para problemas de regressão, além de verificar acerto ou erro, queremos saber o tamanho do erro
- Várias medidas disponíveis:
  - **Erro médio quadrático** (*Mean-squared error*) é a mais comum
    - **Raiz do erro médio quadrático** (*Root mean-squared error*)
    - Tende a maximizar o efeito de valores "muito errados"

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

$$rmse(h) = \sqrt{mse(h)}$$

  - **Erro médio absoluto** *(Mean absolute erros)*
    - Trata todos os erros (pequenos ou grandes) igualmente de acordo com a sua magnitude

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

MEDIDAS DE AVALIAÇÃO

- **Erro relativo quadrático** (*Relative squared error*)
  - **Raiz do erro relativo quadrático** (*Root Relative squared error*)
  - Pondera o erro de acordo com a sua previsibilidade: considera a distribuição dos valores em torno da média

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

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$rrse(h) = \sqrt{rse(h)}$$

- **Erro relativo absoluto** (*Relative absolute error*)
  - Faz a mesma ponderação de valores de acordo com distribuição em torno da média

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

- **Coeficiente de correlação** (*Correlation coefficient*)
  - Mede a correlação estatística entre os valores reais e os valores previstos
  - Varia de 1 a -1:
    1 (correlação perfeita),
    -1 (correlação inversa perfeita), 0 (nenhuma correlação)

$$corr(h) = \frac{\frac{\sum_{i=1}^{n}(y_i - \bar{y})(h(x_i) - \bar{h})}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} \times \frac{\sum_{i=1}^{n}(h(x_i) - \bar{h})^2}{n-1}}}$$

MEDIDAS DE AVALIAÇÃO

- Qual dessas medidas usar?
  - Depende do problema que está sendo tratado
- Na prática, geralmente o melhor modelo de regressão será o melhor independente da medida usada na avaliação
  - Quanto menor a medida do *erro*, melhor
  - Quanto maior o valor de *correlação*, melhor

COMO AVALIAR UM CLASSIFICADOR?

- Dados um conjunto de exemplos e um algoritmo indutor, deseja-se estimar o desempenho do modelo induzido

- Já vimos quais medidas podem ser utilizadas, mas a forma como serão obtidos os resultados também é importante para que a avaliação seja válida e realista
  - Avaliação (teste) realizada com os mesmos exemplos que foram utilizados para induzir (treinar) o modelo produz resultados tendenciosos!

Medidas de avaliação

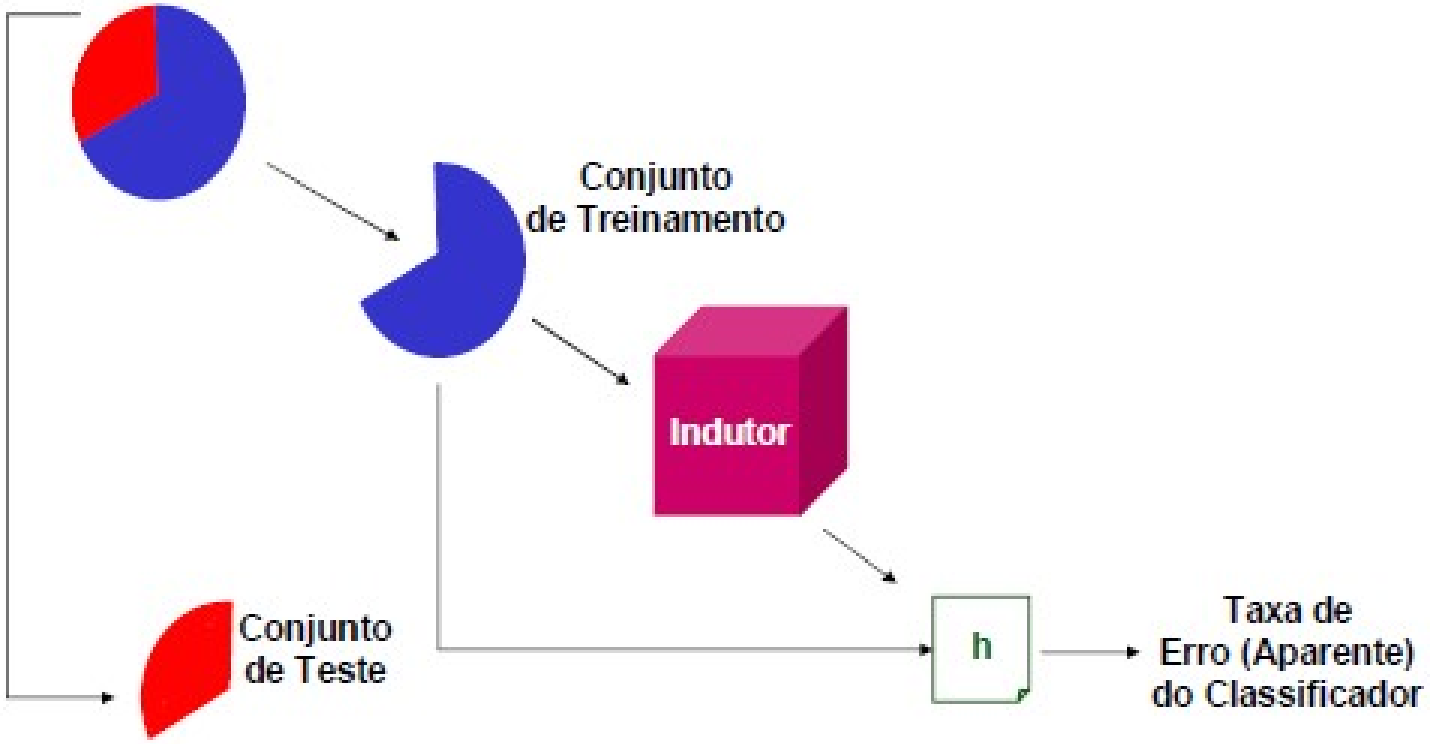
- Usualmente, o conjunto de exemplos é dividido em dois subconjuntos disjuntos:
  - **Conjunto de treinamento** que é usado para a indução (aprendizado) da hipótese
  - **Conjunto de teste** usado para medir o desempenho da hipótese induzida

- Os subconjuntos são disjuntos para assegurar que as medidas obtidas utilizando o conjunto de teste sejam de um conjunto diferente do usado para realizar o aprendizado

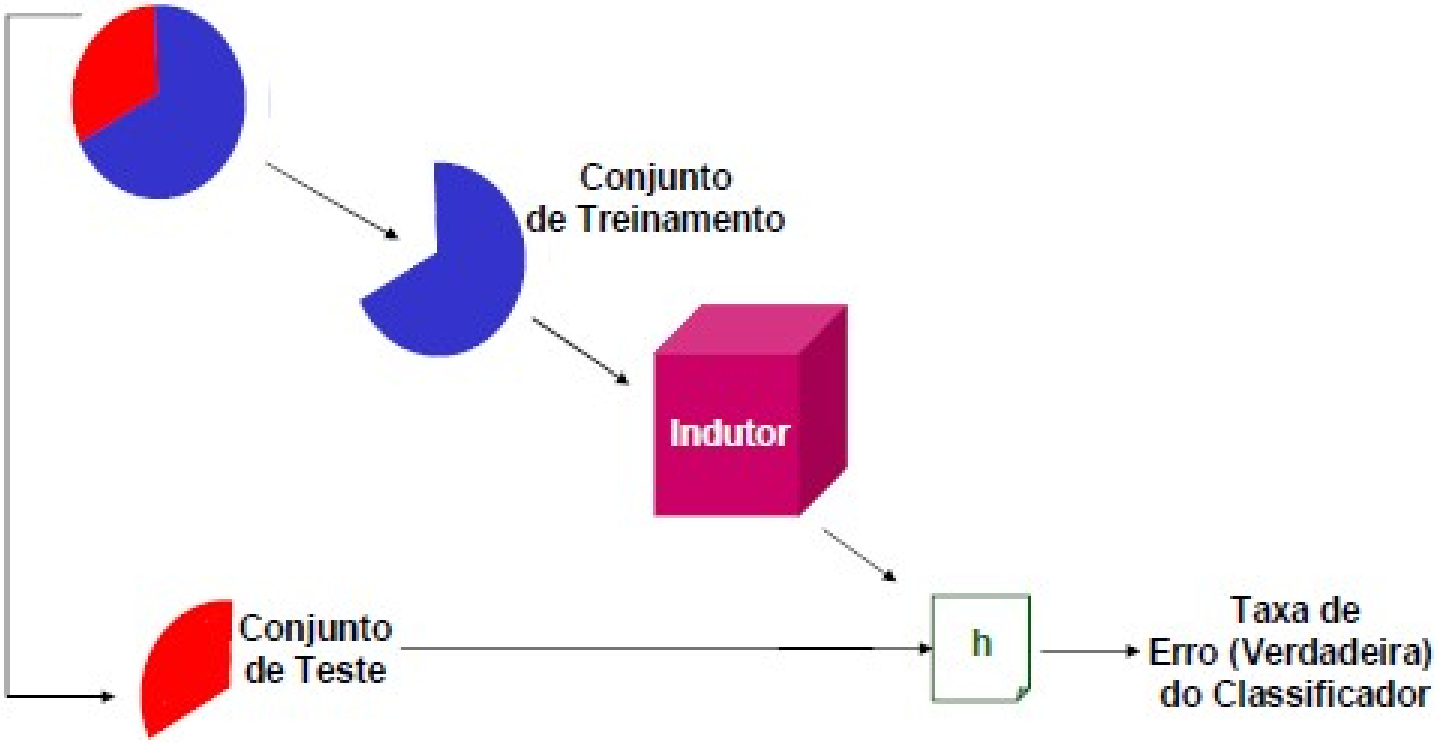
MEDIDAS DE AVALIAÇÃO

- Medidas de desempenho efetuadas sobre o conjunto de treinamento são chamadas **aparentes**
- Medidas efetuadas sobre o conjunto de teste são chamadas medidas **reais**
  - Por exemplo, caso a medida seja o **erro**, teremos o **erro aparente** e o **erro real**
- Para a maioria das hipóteses, a medida aparente é um estimador ruim do seu desempenho futuro
  - Em geral, o erro calculado sobre o conjunto de exemplos de treinamento (erro aparente) é menor que o erro calculado sobre o conjunto de exemplos de teste (erro verdadeiro)

MEDIDAS DE AVALIAÇÃO



MEDIDAS DE AVALIAÇÃO

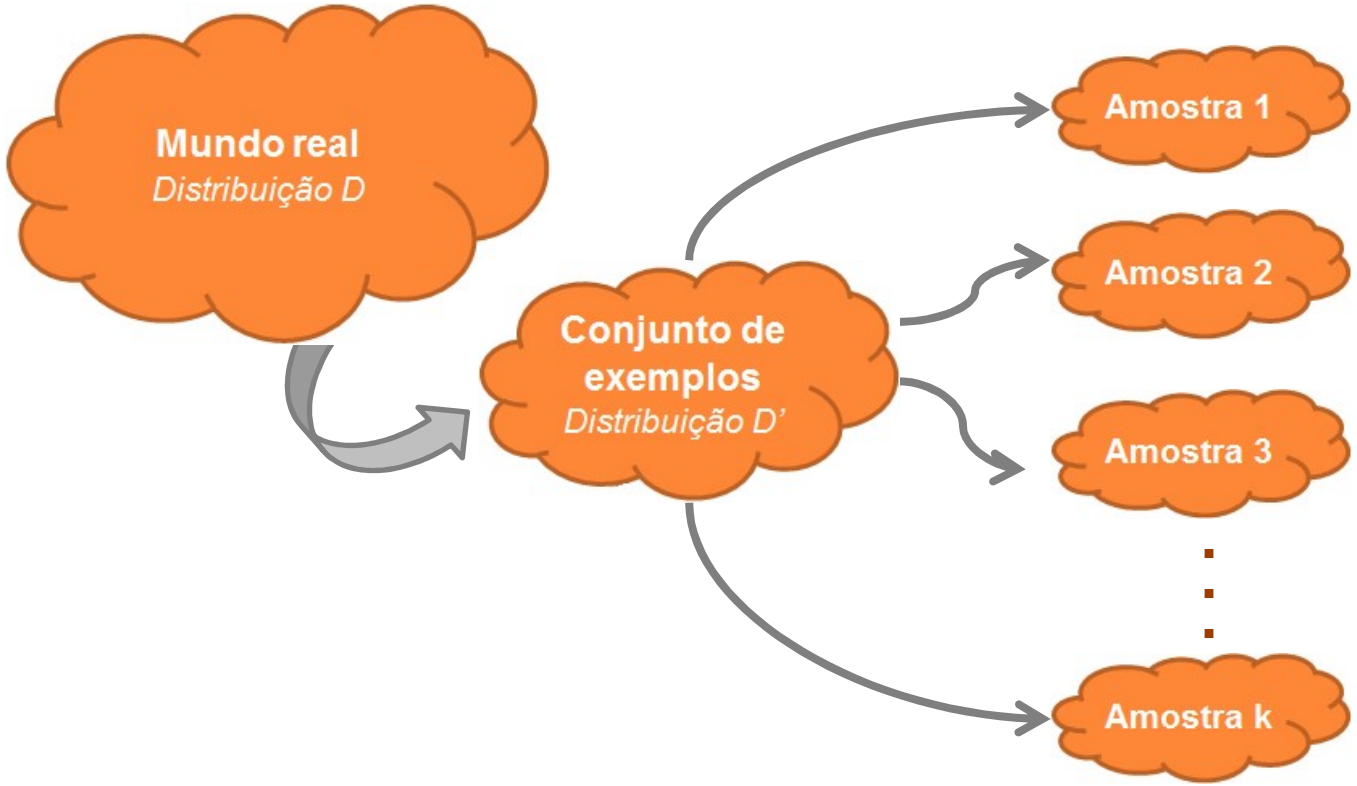


COMO AVALIAR UM CLASSIFICADOR?

- São duas as formas de avaliação mais utilizadas para algoritmos de aprendizado supervisionado
  - ***Holdout*** e **validação cruzada** (*cross-validation*)

- Ambas são técnicas de **amostragem** usadas para dividir o conjunto de exemplos em conjunto de treinamento e conjunto de teste

AMOSTRAGEM



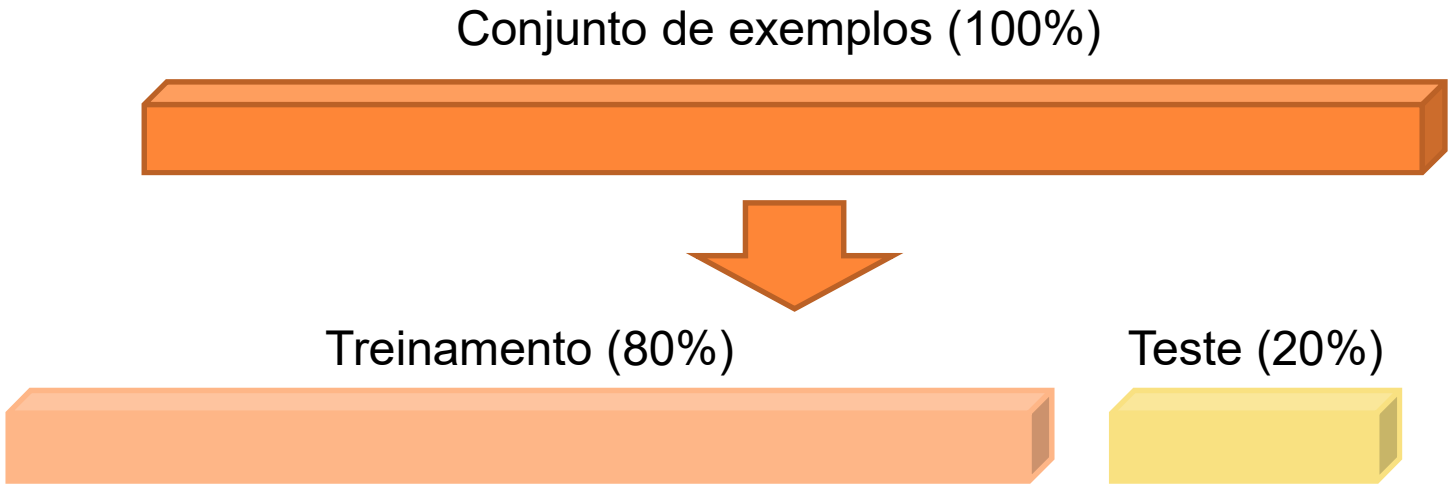
- É importante que as amostras sejam ***aleatórias***, i.e., os exemplos não devem ser pré-selecionados
  - A pré-seleção de exemplos tornaria a estimativa tendenciosa

COMO AVALIAR UM CLASSIFICADOR?

- ***Holdout:*** consiste em dividir os exemplos em uma porcentagem fixa de exemplos *p* para treinamento e *(1-p)* para teste, considerando sempre $p > 1/2$
- Valores típicos de *p* são
  - 2/3 para o treinamento e 1/3 *(1-p)* para o teste
- Outros valores comumente utilizados são:
  - 75% para treinamento e 25% para teste
  - 80% para treinamento e 20% para teste

- Não existem fundamentos teóricos sobre esses valores, apenas empíricos

HOLDOUT



- Lembrando que os exemplos que compõem o conjunto de teste devem ser selecionados aleatoriamente
- Por conta disso, partições diferentes podem gerar resultados diferentes

HOLDOUT

- Para tornar o resultado menos dependente do modo como foi feita a divisão dos exemplos, pode-se calcular a média de vários resultados de *holdout* por meio da construção de várias partições obtendo-se, assim, uma estimativa média do *holdout*
- ***holdout*** tende a superestimar o erro verdadeiro
  - Uma vez que uma hipótese construída utilizando todos os exemplos, em média, apresenta desempenho melhor que uma hipótese construída utilizando apenas uma parte dos exemplos
- Para pequenos conjuntos, nem sempre é possível separar uma parte dos exemplos para teste

COMO AVALIAR UM CLASSIFICADOR?

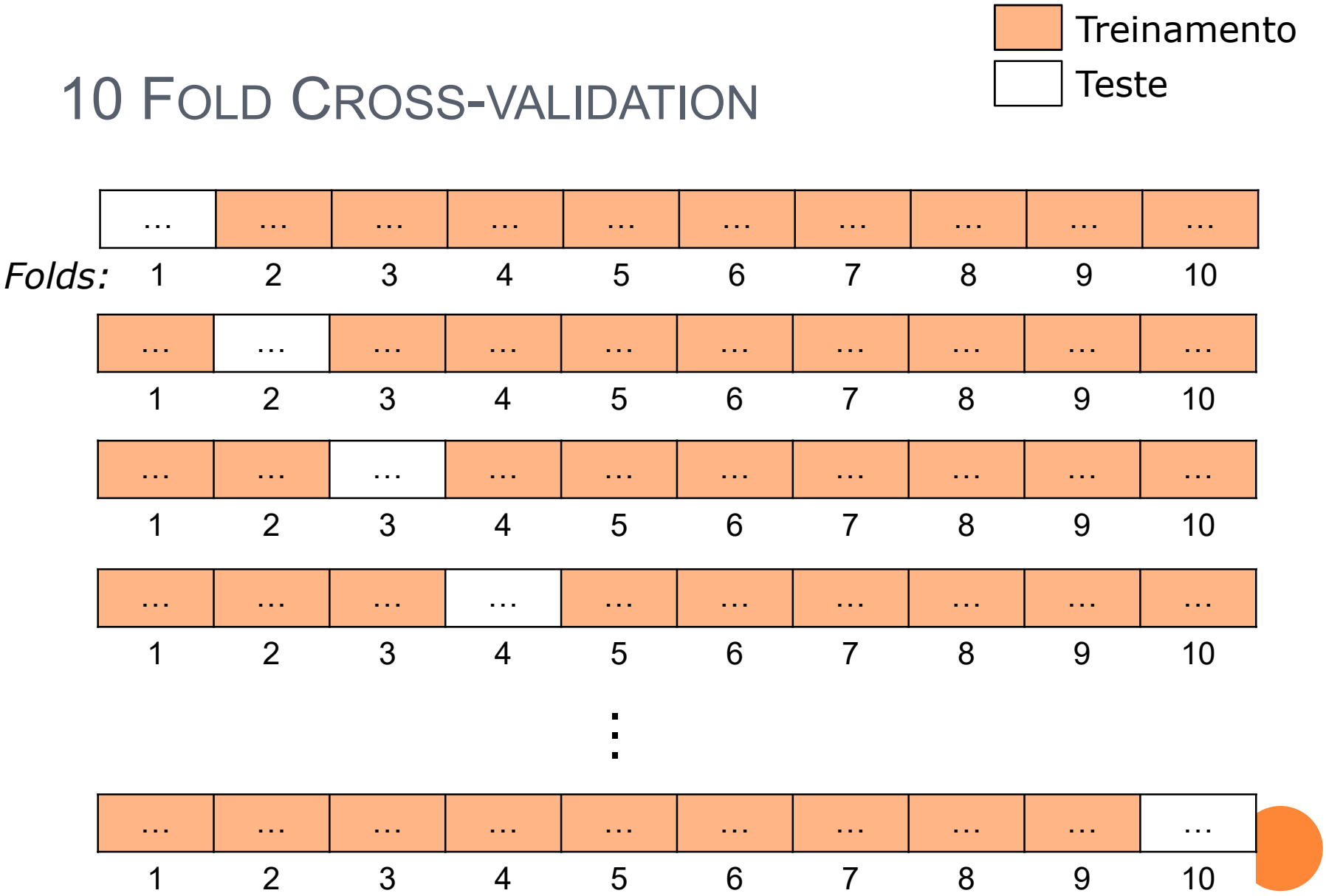
- **Validação cruzada (*cross-validation*)**: os exemplos são divididos aleatoriamente em *k* partições mutuamente exclusivas (chamadas de *folds*) de tamanho aproximadamente igual a *n/k*
- Os exemplos de (*k*-1) *folds* são usados para o treinamento e o classificador induzido é testado com os exemplos do *fold* remanescente
- O processo é repetido *k* vezes, usando um *fold* diferente para o teste em cada vez
- O **erro** é a média dos erros calculados nos *k folds*

$$err(h) = \frac{1}{k}\sum_{i=1}^{k} err(fold_i)$$

Validação Cruzada

- Um valor típico para *k* é 10 (*10-fold cross-validation*), mas outros valores podem ser adotados, dependendo do conjunto de treinamento
- A vantagem da validação cruzada é que se usa todos os exemplos disponíveis para o treinamento e para o teste, sem que o mesmo exemplo seja utilizado em ambos os processos
  - Ideal para quando se tem conjuntos de exemplos pequenos

10 Fold Cross-validation



STRATIFIED CROSS-VALIDATION

- ***Stratified cross-validation***: similar à *cross-validation*, mas ao gerar os *folds* mutuamente exclusivos, a distribuição de classes (proporção de exemplos em cada uma das classes) é considerada durante a amostragem
- Isso significa que todos os *folds* terão aproximadamente a mesma distribuição de classes
- Por exemplo, se o conjunto de exemplos original possui duas classes com distribuição de 20% e 80%, então cada *fold* também terá essa proporção de classes

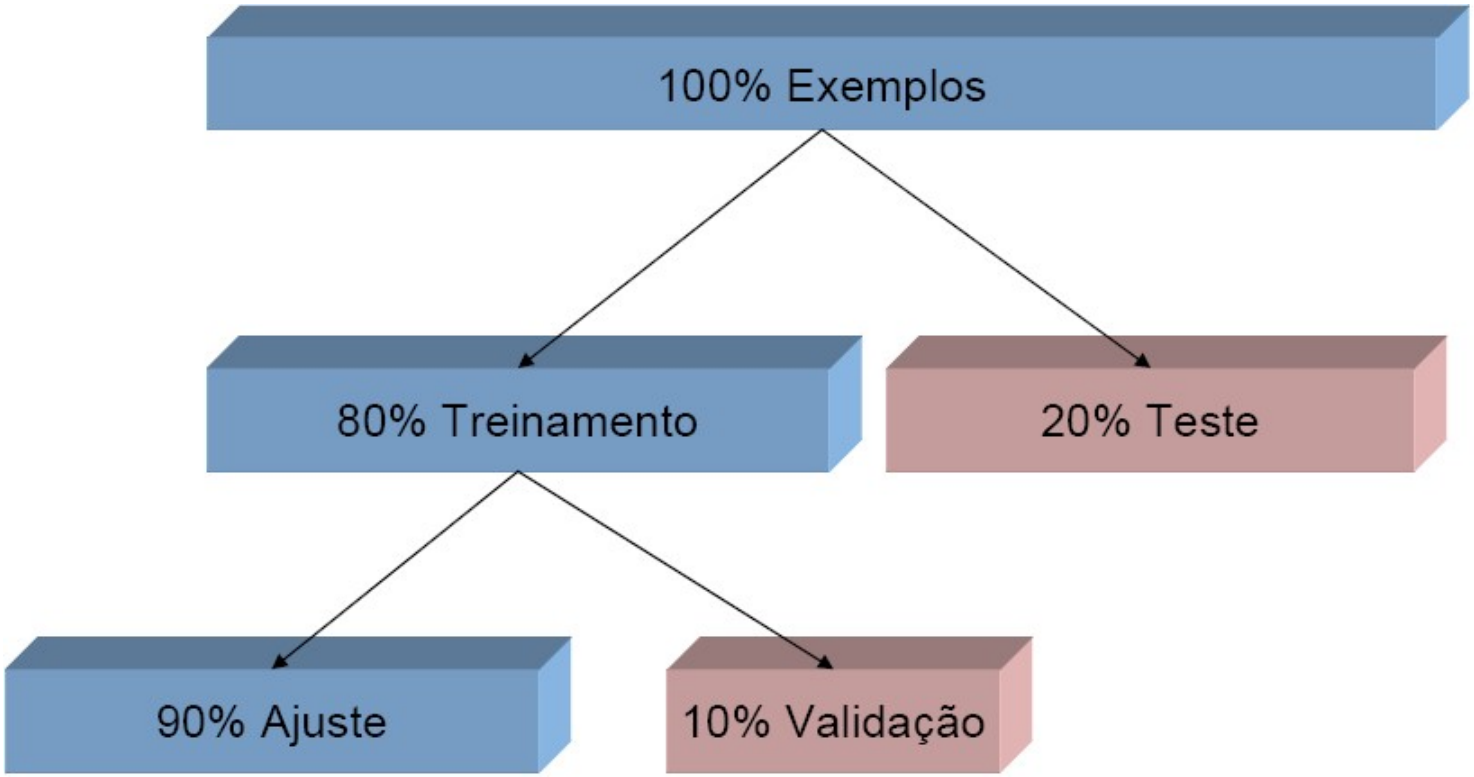
LEAVE-ONE-OUT

- ***Leave-one-out**:* é um caso especial de validação cruzada
- É computacionalmente caro e frequentemente é usado com 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
  - Esse processo é repetido **n** vezes, cada vez induzindo uma hipótese deixando de fora um único exemplo
  - Similar a fazer **n**-*folds cross-validation*
- O erro é a soma dos erros em cada teste individual dividido por **n**

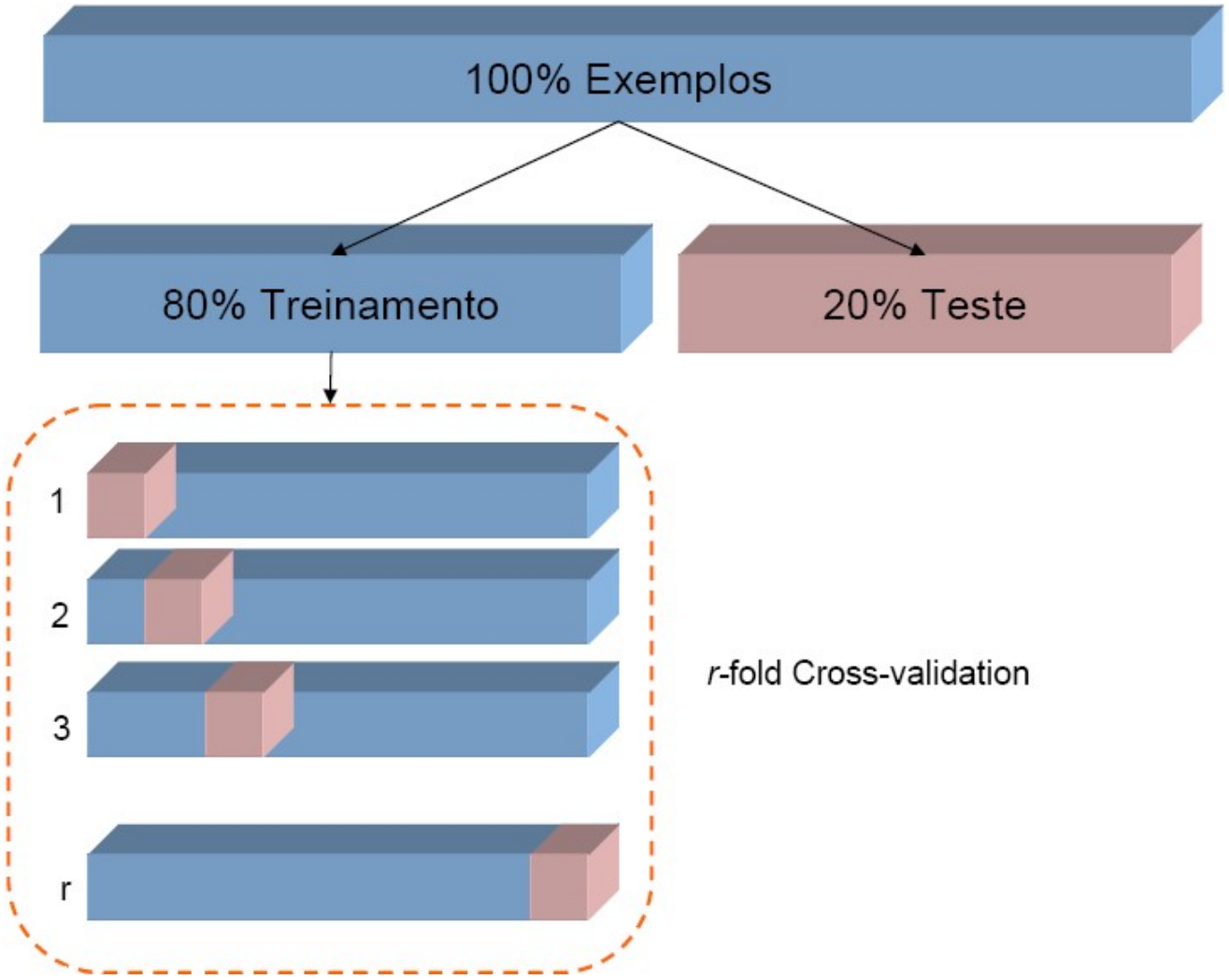
Conjunto de validação

- Em algumas situações torna-se necessário realizar ajustes de parâmetros no indutor
  - Fator de confiança (poda), número mínimo de exemplos em cada folha, etc. (Árvore de decisão)
  - Número de condições por regra, suporte, etc. (Indução de Regras)
  - Número de neurônios por camadas, tipo de função de ativação, número de camadas, etc. (RNA)
- Nesses casos, é necessário reservar uma parte dos exemplos para ajuste dos parâmetros e outra parte para teste
- O conjunto usado para ajustar parâmetros é chamado de **conjunto de validação** ou **de desenvolvimento**

VALIDAÇÃO COM *HOLDOUT*



VALIDAÇÃO COM *CROSS-VALIDATION*



AVALIANDO O DESEMPENHO DE UMA HIPÓTESE

- Existem testes estatísticos que podem ser aplicados para estimar a precisão de hipóteses
- Quando se tem um conjunto de dados grande, estimar a precisão é fácil
- O problema está em estimar a precisão de uma hipótese quando o conjunto de exemplos é pequeno
  - Suponha dois classificadores h_1 e h_2, sendo $acc(h_1) = acc(h_2) = 0,8 = 80\%$
  - h_1 foi avaliado com 1.000 exemplos
  - h_2 foi avaliado com 100 exemplos
  - **Qual resultado é mais confiável?**
  - **Qual é a taxa de sucesso verdadeira?**

AVALIANDO O DESEMPENHO DE UMA HIPÓTESE

- Existem testes estatísticos que podem ser aplicados para estimar
- Quando se tem um cor
  estimar a precisão é fá
- O problema está em es
  hipótese quando o conjunto dados
  - Suponha dois classificadores, h_1 e h_2, se
    $acc(h_1) = acc(h_2) = 0,8 = 80\%$
  - h_1 foi avaliado com 1.000 exemplos
  - h_2 foi avaliado com 100 exemplos
  - **Qual resultado é mais confiável?**
  - **Qual é a taxa de sucesso verdadeira?**

Quanto maior for o conjunto usado na estimativa, maior a probabilidade da taxa verdadeira estar perto da taxa estimada

Para h_1 (1.000), podemos afirmar com **80% de confiança** que a taxa de acerto verdadeira está entre **0,78** e **0,82**.
Para h_2 (100), mantendo 80% de confiança, a taxa de acerto verdadeira está entre **0,74** e **0,85**.

AVALIANDO O DESEMPENHO DE UMA HIPÓTESE

- Quando calculamos a taxa de erro/acerto de uma hipótese *h* estamos fazendo uma estimativa do desempenho de *h* em casos futuros (população) com base no desempenho medido sobre o conjunto de dados (amostra) → inferência
- Queremos saber qual é o $erro_X(h)$ sobre a população *X* com distribuição desconhecida *D*, ou seja, o erro esperado quando se aplica *h* a novos exemplos
- Só podemos calcular o $erro_S(h)$ sobre uma amostra *S* contendo exemplos extraídos aleatoriamente de *X*
- A questão que se coloca é:

  Quão boa uma estimativa do $erro_X(h)$ *é fornecida pelo* $erro_S(h)$ *dada uma amostra de tamanho n?*

- Essa questão é respondida calculando-se um **intervalo de confiança** para $erro_S(h)$

AVALIANDO O DESEMPENHO DE UMA HIPÓTESE

- O **intervalo de confiança** é uma amplitude de valores que tem certa probabilidade de conter o valor verdadeiro da população
- A probabilidade associada ao intervalo é chamada de **nível de confiança**
- Os valores referentes a intervalos de confiança para uma medida amostral ***p*** são calculados considerando-se a *tabela de probabilidades da distribuição normal*
  - Para o nível de confiança *N%*, usa-se o respectivo valor da constante Z_N no cálculo do intervalo

Nível de confiança $N\%$:	50%	68%	80%	90%	95%	98%	99%
Constante Z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

$$p \pm Z_N \sqrt{\frac{p \times (1-p)}{n}}$$

AVALIANDO O DESEMPENHO DE UMA HIPÓTESE

- Para um caso de classificação, isto é, a hipótese *h* erra ou acerta, e considerando que:
  - A amostra *S* contém *n* exemplos retirados um a um de acordo com a distribuição *D*, independente de *h*
  - $n \geq 30$ (restrição vinda do Teorema do Limite Central)
  - A hipótese *h* classifica erroneamente *r* desses exemplos, ou seja, $erro_S(h) = r/n$
- Com base na teoria da inferência estatística é possível afirmar que $erro_S(h)$ é o valor provável de $erro_X(h)$ e, com probabilidade de 95%, o valor de $erro_X(h)$ está no intervalo dado por:

$$erro_S(h) \pm 1{,}96\sqrt{\frac{erro_S(h) \times (1 - erro_S(h))}{n}}$$

AVALIANDO O DESEMPENHO DE UMA HIPÓTESE

- Por exemplo, se a amostra S contém n = 50 exemplos e a hipótese *h* classifica erroneamente r = 10 desses exemplos, então $erro_S(h)$ = 10/50 = 0,2
- Se o experimento fosse repetido várias vezes retirando-se outras amostras S1, S2, etc., de 50 exemplos, espera-se que os erros dessas amostras $erro_{S1}$, $erro_{S2}$, etc., apresentem valores ligeiramente diferentes que $erro_S(h)$
- Mas será encontrado que para 95% desses experimentos, o intervalo calculado contém o erro verdadeiro
- Por isso, esse intervalo é denominado intervalo de confiança de 95% da estimativa para $erro_X(h)$

$$0,2 \pm 1,96\sqrt{\frac{0,2 \times (1-0,2)}{50}} = 0,2 \pm 0,110 = \begin{matrix} 0,09 \\ 0,31 \end{matrix}$$

AVALIANDO O DESEMPENHO DE UMA HIPÓTESE

- Intervalos de Confiança para n = 50 e r = 10

$N\%$	$Z_N\sqrt{\frac{erro_S(h)(1-erro_S(h))}{n}}$	Intervalo de Confiança
50%	0.037	[0.163 , 0.237]
68%	0.056	[0.144 , 0.256]
80%	0.072	[0.128 , 0.272]
90%	0.092	[0.108 , 0.292]
95%	0.110	[0.090 , 0.310]
98%	0.131	[0.069 , 0.331]
99%	0.145	[0.055 , 0.345]

- Observe que os intervalos com maior nível de confiança são maiores, pois está aumentando a probabilidade de $erro_X(h)$ estar no intervalo

AVALIANDO O DESEMPENHO DE UMA HIPÓTESE

- Quando se mantém o mesmo nível de confiança e se aumenta o tamanho da amostra, o intervalo diminui
  - Quanto maior for a amostra, maior é a probabilidade de $erro_X(h)$ estar no intervalo

$$0{,}2 \pm 1{,}96\sqrt{\frac{0{,}2 \times (1-0{,}2)}{50}} = 0{,}2 \pm 0{,}110 = \begin{matrix} 0{,}09 \\ 0{,}31 \end{matrix}$$

$$0{,}2 \pm 1{,}96\sqrt{\frac{0{,}2 \times (1-0{,}2)}{100}} = 0{,}2 \pm 0{,}078 = \begin{matrix} 0{,}12 \\ 0{,}28 \end{matrix}$$

$$0{,}2 \pm 1{,}96\sqrt{\frac{0{,}2 \times (1-0{,}2)}{1000}} = 0{,}2 \pm 0{,}025 = \begin{matrix} 0{,}18 \\ 0{,}23 \end{matrix}$$