

# Aprendizado de Máquina

Avaliação de Modelos Preditivos

Daniel Sabino A. de Araújo

# Avaliação Modelos Preditivos

- Não existe técnica de AM universal, que se saia melhor em qualquer tipo de problema
  - Implica na necessidade de experimentos
- Características do problema e das técnicas pode auxiliar em alguns casos
  - Ex. modelo deve ser interpretável  
⇒ técnicas simbólicas, dados possuem alta dimensão ⇒RNA, etc.
  - Mesmo assim diversos algoritmos podem ser candidatos

# Avaliação Modelos Preditivos

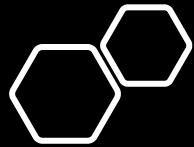
- Mesmo que um único algoritmo seja escolhido
  - Variações de parâmetros produzem diferentes modelos
- Domínio de AM: necessidade de experimentação
  - Experimentos controlados
  - Procedimentos que garantem a corretude e reproduzibilidade dos experimentos

# Avaliação Modelos Preditivos

- Mesmo que um único algoritmo seja escolhido
  - Variações de parâmetros produzem diferentes modelos
- Domínio de AM: necessidade de experimentação
  - Experimentos controlados
  - Procedimentos que garantem a corretude e reproduzibilidade dos experimentos

# Modelos preditivos

- Diferentes aspectos podem ser considerados:
  - Acurácia do modelo nas previsões
  - Compreensibilidade do conhecimento extraído
  - Tempo de aprendizado
  - Requisitos de armazenamento
  - Etc.
- Concentraremos discussões a medidas de desempenho preditivo



# Medidas de desempenho

- Desempenho na rotulação de objetos
  - Métricas para classificação:
    - Taxa de erro
    - Acurácia
    - Revocação
    - Precisão
  - Métricas para regressão:
    - Erro quadrático médio
    - Distância absoluta média



Medidas para  
Classificação

# Taxa de erro

- Taxa de erro de um classificador  $f$ 
  - classificações incorretas

$$err(f) = (1/n) \sum_{i=1 \dots n} I(y_i \neq f(\mathbf{x}_i))$$

- Proporção de exemplos classificados incorretamente em um conjunto com  $n$  objetos
  - Comparação da classe conhecida com a predita
  - $I$  é função identidade
    - = 1 se argumento é verdadeiro e 0 em caso contrário
- Varia entre 0 e 1 e valores próximos de 0 são melhores

# Taxa de acerto

- Taxa de acerto ou acurácia de um classificador  $f$ 
  - Complemento da taxa de erro

$$ac(f) = 1 - err(f) = (1/n) \sum_{i=1 \dots n} I(y_i = f(\mathbf{x}_i))$$

- Proporção de exemplos classificados corretamente em um conjunto com  $n$  objetos
  - Varia entre 0 e 1 e valores próximos de 1 são melhores

# Matriz de confusão

- Matriz de confusão
  - Alternativa para visualizar desempenho de classificador
  - Predições corretas e incorretas em cada classe

		Classe predita		
		c1	c2	c3
Classe verdadeira	c1	11	1	3
	c2	1	4	0
	c3	2	1	6

- Linhas representam classes verdadeiras
- Colunas representam classes preditas
- Elemento  $m_{ij}$ : número de exemplos da classe  $c_i$  classificados como pertencentes à classe  $c_j$
- Diagonal da matriz: acertos do classificador
- Outros elementos: erros cometidos

# Classificação binária

- Seja um problema com duas classes: + e -

Matriz de confusão:

		Classe predita	
		+	-
Classe verdadeira	+	VP	FN
	-	FP	VN
+   -	VP   FN	FP   VN	

**VP: verdadeiros positivos**

Número de exemplos da classe + classificados corretamente

**VN: verdadeiros negativos**

Número de exemplos da classe - classificados corretamente

**FP: falsos positivos**

Número de exemplos da classe - classificados incorretamente como +

**FN: falsos negativos**

Número de exemplos da classe + classificados incorretamente como -

Medidas obtidas  
a partir da matriz  
de confusão

# Erro e acurácia

Taxa de erro total:

Soma da diagonal secundária da matriz /  $n$

$$err(f) = \frac{FP + FN}{n}$$

Taxa de acerto ou acurácia total:

Soma da diagonal principal /  $n$

$$ac(f) = \frac{VP + VN}{n}$$

# Precisão e revocação

Precisão:

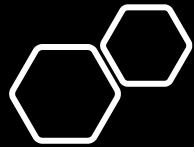
Proporção de exemplos + classificados corretamente entre os preditos  
como +

$$prec(f) = \frac{VP}{VP + FP}$$

Sensibilidade ou revocação:

Taxa de acerto na classe positiva (taxa de verdadeiros positivos)

$$sens(f) = \frac{VP}{VP + FN}$$



# Precisão vs revocação

Precisão: exatidão do modelo

- Ex. precisão 1,0 para uma classe C: itens rotulados como C realmente pertencem a C
  - Não fornece informação sobre exemplos de C que não foram corretamente classificados

Revocação: completude do modelo

- Ex. revocação 1,0 para uma classe C: itens da classe C foram rotulados como pertencendo a C
  - Não fornece informação sobre exemplos que foram classificados incorretamente como C

# F-measure

- Precisão e revocação costumam ser discutidas em conjunto, combinadas em uma medida F:

$$F(f) = \frac{(w + 1) \text{ rev}(f) \text{ prec}(f)}{\text{rev}(f) + w \text{ prec}(f)}$$

- Média harmônica da previsão e revocação
- Usando  $w = 1 \Rightarrow$  mesmo grau de importância para duas medidas  $\Rightarrow F_1$

$$F_1(f) = \frac{2 \text{ rev}(f) \text{ prec}(f)}{\text{rev}(f) + \text{prec}(f)}$$

# Especificidade

Especificidade:

Taxa de acerto na classe

Seu complemento é a taxa de falsos positivos.

$$esp(f) = \frac{VN}{VN + FP}$$

# Medidas de desempenho

- Ex.: avaliação de três classificadores

		Classe predita	
		p	n
Classe verdadeira	p	20	30
	n	15	35

Classificador 1  
TVP = 0.2  
TFP = 0.15

		Classe predita	
		p	n
Classe verdadeira	p	70	30
	n	50	50

Classificador 2  
TVP = 0.7  
TFP = 0.5

		Classe predita	
		p	n
Classe verdadeira	p	60	40
	n	20	80

Classificador 3  
TVP = 0.6  
TFP = 0.2

# Generalizando para mais classes

- Para mais que duas classes:
  - Considera cada uma + e as demais -
  - Ex. C1:

	C1	C2	C3
C1	TP	FN	FN
C2	FP	TN	TN
C3	FP	TN	TN

	C1	C2	C3
C1	49	1	0
C2	0	47	3
C3	0	2	48

C1		
	+	-
+	TP	FN
-	FP	TN

C1		
	+	-
+	49	1
-	0	100

$$\text{erro (+)} = \frac{\text{FN}}{\text{TP}+\text{FN}} = 0.02$$

$$\text{erro (-)} = \frac{\text{FP}}{\text{FP}+\text{TN}} = 0.00$$

# Generalizando para mais classes

- Para mais que duas classes:
  - Ex. C2:

	C1	C2	C3
C1	TN	FP	TN
C2	FN	TP	FN
C3	TN	FP	TN

	C1	C2	C3
C1	49	1	0
C2	0	47	3
C3	0	2	48

C2		
	+	-
+	TP	FN
-	FP	TN

C2		
	+	-
+	47	3
-	3	97

$$\text{erro (+)} = \frac{\text{FN}}{\text{TP} + \text{FN}} = 0.06$$

$$\text{erro (-)} = \frac{\text{FP}}{\text{FP} + \text{TN}} = 0.03$$

# Generalizando para mais classes

- Para mais que duas classes:

- Ex. C3:

	C1	C2	C3
C1	TN	TN	FP
C2	TN	TN	FP
C3	FN	FN	TP

	C1	C2	C3
C1	49	1	0
C2	0	47	3
C3	0	2	48

C3		
	+	-
+	TP	FN
-	FP	TN

C3		
	+	-
+	48	2
-	3	97

$$\text{erro (+)} = \frac{\text{FN}}{\text{TP} + \text{FN}} = 0.04$$

$$\text{erro (-)} = \frac{\text{FP}}{\text{FP} + \text{TN}} = 0.03$$

# Amostragem

- Tem-se usualmente um único conjunto de  $n$  objetos
  - Deve ser usado para induzir e avaliar o preditor
  - Desempenho no conjunto de treinamento é otimista
    - Todos algoritmos tentam de alguma forma melhorar seu desempenho no conjunto de treinamento na fase indutiva
    - Avaliar modelo no conjunto de treinamento é conhecido como resubstituição
      - Produz taxa de erro/acerto aparente

# Amostragem

- Observações:
  - Para médias de desempenho, é importante reportar também os valores de desvio-padrão
    - Alto desvio padrão ⇒ alta variabilidade dos resultados
    - Indicativo de sensibilidade a variações nos dados de treinamento
  - Estimativas mais precisas também podem ser obtidas usando intervalos de confiança

# Amostragem

- Métodos de amostragem: obter estimativas de desempenho mais confiáveis
  - Definindo subconjuntos disjuntos de:

## Treinamento

Dados empregados na **indução** e no **ajuste** do modelo

Qualquer ajuste de parâmetros deve ser feito **nos dados de treinamento**

## Teste

Simulam a apresentação de **novos exemplos** ao preditor  
(não vistos em sua indução)

**Somente avaliar o modelo obtido**

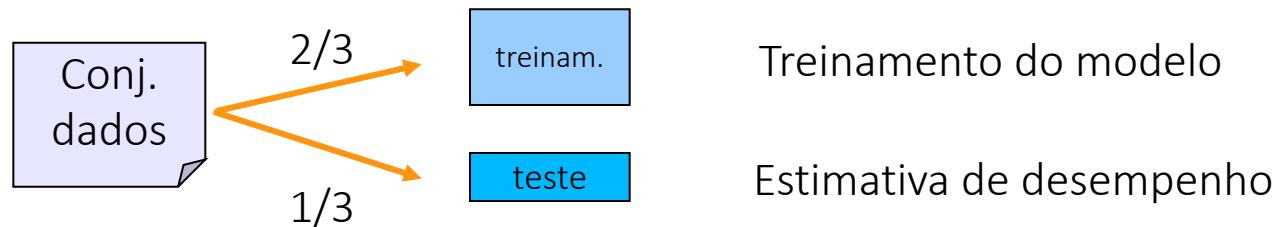
Em algumas situações, dados de treinamento são subdivididos, gerando conjunto de **validação** dedicado ao ajuste de parâmetros

# Amostragem

- Principais métodos de amostragem:
  - Holdout
  - Amostragem aleatória
  - Validação cruzada
  - Leave-one-out
  - Bootstrap

# Holdout

- Método mais simples:
  - Divide conjunto de dados em proporção  $p$  para treinamento e  $(1-p)$  para teste
    - Uma única partição
    - Valores típicos de  $p$ :  $\frac{1}{2}$ ,  $\frac{2}{3}$  ou  $\frac{3}{4}$



# Holdout

Objeto	Atributo 1	Atributo 2	Atributo 3	Classe
1	855	5142	2708	Safra 95
2	854	23155	2716	Safra 95
3	885	16586	2670	Safra 95
4	877	16685	2677	Safra 95
5	839	5142	2708	Safra 95
6	854	5005	2685	Safra 95
7	885	19455	2708	Safra 95
8	839	5027	2708	Safra 95
9	877	16823	2677	Safra 95
10	892	19180	2716	Safra 95
11	24628	39437	381	Safra 96
12	43183	39277	328	Safra 96
13	27871	39712	389	Safra 96
14	42329	40307	328	Safra 96
15	41627	40032	335	Safra 96
16	39399	40322	335	Safra 96
17	33677	40375	328	Safra 96
18	33539	40078	335	Safra 96
19	34150	40353	358	Safra 96
20	34485	40742	358	Safra 96

# Holdout

## Conjunto de treinamento

Objeto	Atributo 1	Atributo 2	Atributo 3	Classe
4	877	16685	2677	Safra 95
6	854	5005	2685	Safra 95
8	839	5027	2708	Safra 95
2	854	23155	2716	Safra 95
10	892	19180	2716	Safra 95
1	855	5142	2708	Safra 95
6	854	5005	2685	Safra 95
18	33539	40078	335	Safra 96
15	41627	40032	335	Safra 96
19	34150	40353	358	Safra 96
12	43183	39277	328	Safra 96
17	33677	40375	328	Safra 96
20	34485	40742	358	Safra 96
11	24628	39437	381	Safra 96

## Conjunto de teste

Objeto	Atributo 1	Atributo 2	Atributo 3	Classe
3	885	16586	2670	Safra 95
5	839	5142	2708	Safra 95
9	877	16823	2677	Safra 95
13	27871	39712	389	Safra 96
14	42329	40307	328	Safra 96
16	39399	40322	335	Safra 96

# Holdout

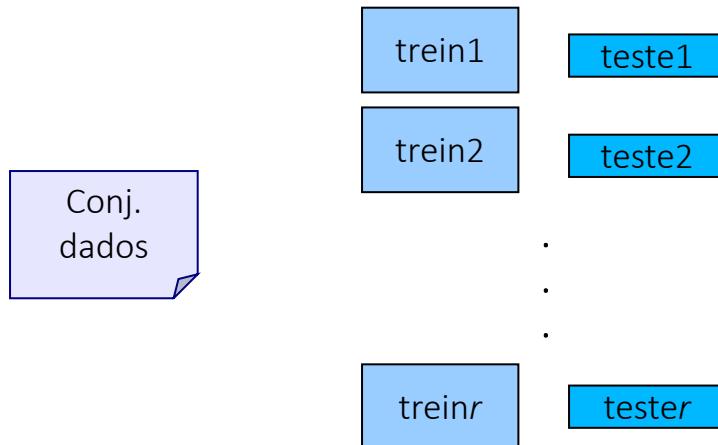
- Indicado para grande quantidade de dados
  - Se pequena quantidade de dados
    - Poucos exemplos são usados no treinamento
    - Modelo pode depender da composição dos conjuntos de treinamento e teste
      - Quanto menor conjunto de treinamento, maior a variância do modelo
      - Quanto menor conjunto de teste, menos confiável a acurácia estimada para ele
- Muito usado para definir subconjuntos de validação

# Holdout

- Não avalia o quanto o desempenho de uma técnica varia
  - Quanto a diferentes combinações de exemplos de treinamento
  - É possível que uma divisão deixe no subconjunto de teste exemplos “mais fáceis”
  - Para tornar os resultados menos dependentes da partição feita:
    - vários holdout
    - Random subsampling (amostragem aleatória)

# Amostragem aleatória

- Repetições de holdout
  - Há sobreposição entre os conjuntos de teste gerados
  - Fornece uma média de desempenho



Média e  
desvio-padrão  
de desempenho

# Validação cruzada

- Método mais usado: r-fold cross validation
  - Conjunto é dividido em r partes de tamanho aproximadamente igual
  - Objetos de r-1 partes são usados no treinamento e a parte restante é usada para teste
  - Procedimento é repetido r vezes usando cada partição para teste
    - subconjuntos de teste são independentes entre si
  - Desempenho é dado por média
  - Valor típico de r: 10

# Validação cruzada

- Variação: r-fold cross validation estratificado
  - Manter a distribuição de classes em cada partição
    - Ex: se conjunto de dados original tem 20% na classe c1 e 80% na classe c2, cada partição também deve manter essa proporção
  - Distribuição de classes: proporção de exemplos em cada classe
    - Para cada classe  $c_j$ ,  $dist(c_j) = \text{número de exemplos que possuem a classe } c_j / \text{número total de exemplos}$

$$dist(c_j) = \frac{1}{n} \sum_{i=1}^n |y_i = c_j|$$

# Distribuição de classes

- Ex.: conjunto de dados com 100 exemplos
  - 60 são da classe c1
  - 15 são da classe c2
  - 25 são da classe c3
  - A distribuição de classe é  $\text{dist}(c1,c2,c3) = (0,60, 0,15, 0,25) = (60\%, 15\%, 25\%)$
  - A classe c1 é a classe majoritária ou prevalente
  - A classe c2 é a classe minoritária

# Cross-validation estratificado

- Exemplo:
  - $r = 5$

Objeto	Atributo 1	Atributo 2	Atributo	Classe
4	877	16685	2677	Safra 95
9	877	16823	2677	Safra 95
18	33539	40078	335	Safra 96
11	24628	39437	381	Safra 96
1	855	5142	2708	Safra 95
3	885	16586	2670	Safra 95
14	42329	40307	328	Safra 96
20	34485	40742	358	Safra 96
7	885	19455	2708	Safra 95
10	892	19180	2716	Safra 95
15	41627	40032	335	Safra 96
19	34150	40353	358	Safra 96
6	854	5005	2685	Safra 95
2	854	23155	2716	Safra 95
17	33677	40375	328	Safra 96
12	43183	39277	328	Safra 96
8	839	5027	2708	Safra 95
5	839	5142	2708	Safra 95
15	41627	40032	335	Safra 96
16	39399	40322	335	Safra 96

# Cross-validation estratificado

Conjunto de treinamento

Objeto	Atributo 1	Atributo 2	Atributo	Classe
1	855	5142	2708	Safra 95
3	885	16586	2670	Safra 95
14	42329	40307	328	Safra 96
20	34485	40742	358	Safra 96
7	885	19455	2708	Safra 95
10	892	19180	2716	Safra 95
15	41627	40032	335	Safra 96
19	34150	40353	358	Safra 96
6	854	5005	2685	Safra 95
2	854	23155	2716	Safra 95
17	33677	40375	328	Safra 96
12	43183	39277	328	Safra 96
8	839	5027	2708	Safra 95
5	839	5142	2708	Safra 95
15	41627	40032	335	Safra 96
16	39399	40322	335	Safra 96

Conjunto de teste

Objeto	Atributo 1	Atributo 2	Atributo	Classe
4	877	16685	2677	Safra 95
9	877	16823	2677	Safra 95
18	33539	40078	335	Safra 96
11	24628	39437	381	Safra 96

# Cross-validation estratificado

Conjunto de treinamento

Objeto	Atributo 1	Atributo 2	Atributo	Classe
4	877	16685	2677	Safra 95
9	877	16823	2677	Safra 95
18	33539	40078	335	Safra 96
11	24628	39437	381	Safra 96
7	885	19455	2708	Safra 95
10	892	19180	2716	Safra 95
15	41627	40032	335	Safra 96
19	34150	40353	358	Safra 96
6	854	5005	2685	Safra 95
2	854	23155	2716	Safra 95
17	33677	40375	328	Safra 96
12	43183	39277	328	Safra 96
8	839	5027	2708	Safra 95
5	839	5142	2708	Safra 95
15	41627	40032	335	Safra 96
16	39399	40322	335	Safra 96

Conjunto de teste

Objeto	Atributo 1	Atributo 2	Atributo	Classe
1	855	5142	2708	Safra 95
3	885	16586	2670	Safra 95
14	42329	40307	328	Safra 96
20	34485	40742	358	Safra 96

# Cross-validation estratificado

Conjunto de treinamento

Objeto	Atributo 1	Atributo 2	Atributo	Classe
4	877	16685	2677	Safra 95
9	877	16823	2677	Safra 95
18	33539	40078	335	Safra 96
11	24628	39437	381	Safra 96
1	855	5142	2708	Safra 95
3	885	16586	2670	Safra 95
14	42329	40307	328	Safra 96
20	34485	40742	358	Safra 96
6	854	5005	2685	Safra 95
2	854	23155	2716	Safra 95
17	33677	40375	328	Safra 96
12	43183	39277	328	Safra 96
8	839	5027	2708	Safra 95
5	839	5142	2708	Safra 95
15	41627	40032	335	Safra 96
16	39399	40322	335	Safra 96

Conjunto de teste

Objeto	Atributo 1	Atributo 2	Atributo	Classe
7	885	19455	2708	Safra 95
10	892	19180	2716	Safra 95
15	41627	40032	335	Safra 96
19	34150	40353	358	Safra 96

# Cross-validation estratificado

Conjunto de treinamento

Objeto	Atributo 1	Atributo 2	Atributo	Classe
4	877	16685	2677	Safra 95
9	877	16823	2677	Safra 95
18	33539	40078	335	Safra 96
11	24628	39437	381	Safra 96
1	855	5142	2708	Safra 95
3	885	16586	2670	Safra 95
14	42329	40307	328	Safra 96
20	34485	40742	358	Safra 96
7	885	19455	2708	Safra 95
10	892	19180	2716	Safra 95
15	41627	40032	335	Safra 96
19	34150	40353	358	Safra 96
8	839	5027	2708	Safra 95
5	839	5142	2708	Safra 95
15	41627	40032	335	Safra 96
16	39399	40322	335	Safra 96

Conjunto de teste

Objeto	Atributo 1	Atributo 2	Atributo	Classe
6	854	5005	2685	Safra 95
2	854	23155	2716	Safra 95
17	33677	40375	328	Safra 96
12	43183	39277	328	Safra 96

# Cross-validation estratificado

Conjunto de treinamento

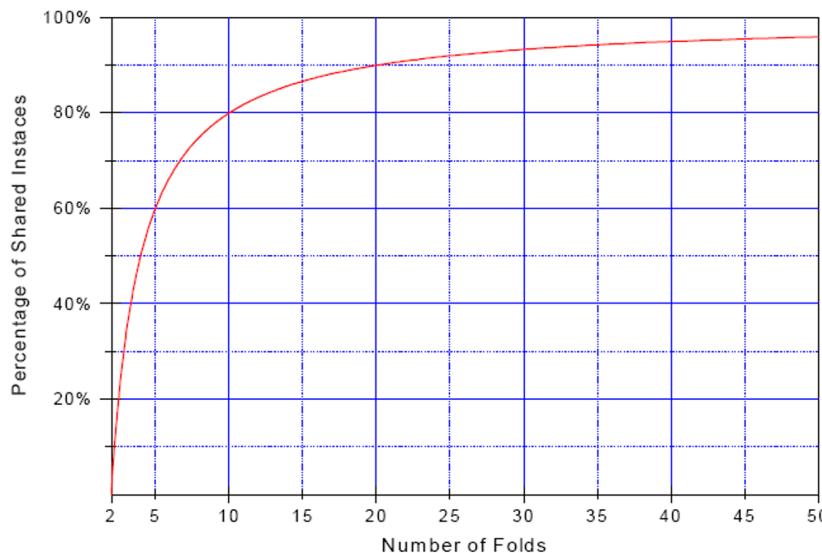
Objeto	Atributo 1	Atributo 2	Atributo	Classe
4	877	16685	2677	Safra 95
9	877	16823	2677	Safra 95
18	33539	40078	335	Safra 96
11	24628	39437	381	Safra 96
1	855	5142	2708	Safra 95
3	885	16586	2670	Safra 95
14	42329	40307	328	Safra 96
20	34485	40742	358	Safra 96
7	885	19455	2708	Safra 95
10	892	19180	2716	Safra 95
15	41627	40032	335	Safra 96
19	34150	40353	358	Safra 96
6	854	5005	2685	Safra 95
2	854	23155	2716	Safra 95
17	33677	40375	328	Safra 96
12	43183	39277	328	Safra 96

Conjunto de teste

Objeto	Atributo 1	Atributo 2	Atributo	Classe
8	839	5027	2708	Safra 95
5	839	5142	2708	Safra 95
15	41627	40032	335	Safra 96
16	39399	40322	335	Safra 96

# Validação cruzada

Crítica: uma parte dos dados é partilhada entre os subconjuntos de treinamento



Para  $r \geq 2$ , uma proporção de  $(1 - 2/r)$  dos objetos é compartilhada  
Ex.  $r = 10 \Rightarrow 80\%$  dos objetos são compartilhados

Slides construídos com base no material fornecido pela autora do livro  
‘inteligência artificial: uma abordagem de aprendizado de máquina’  
(Faceli, 2011).