

# Inteligência Artificial

## Tópico 04 - Parte 02

### Aprendizado de Máquina - Avaliação de Desempenho

Profa. Dra. Priscila Tiemi Maeda Saito  
✉ [priscilasaito@ufscar.br](mailto:priscilasaito@ufscar.br)

# Roteiro

## 1 Aprendizado de Máquina

# Avaliação de Desempenho Preditivo

- Algoritmos e modelos
- Algoritmos de AM induzem modelos
  - ▶ funções, hipóteses
- Desempenho a ser avaliado
  - ▶ saída de um algoritmo de AM:
    - ★ modelo induzido
  - ▶ saída de um modelo de classificação:
    - ★ classificação para um novo exemplo

# Avaliação de Desempenho

- Depende da tarefa
  - ▶ classificação: considera taxa de exemplos incorretamente classificados
    - ★ ex.: acurácia
  - ▶ regressão: considera diferença entre valor previsto e valor correto
    - ★ ex.: MSE
- Média dos erros obtidos em diferentes execuções de um experimento

# Desempenho de Classificação

- Principal **objetivo** de um modelo é a classificação **correta** de novos exemplos
  - ▶ desempenho preditivo
  - ▶ errar o mínimo possível
    - ★ minimizar taxa de erro de classificação
  - ▶ geralmente não é possível medir com exatidão essa taxa de erro
    - ★ deve ser estimada do erro de treinamento
    - ★ amostragem de dados

# Generalização

- Capacidade de **generalização** de uma hipótese
  - ▶ propriedade de continuar válida para outros objetos que não fazem parte de seu conjunto de treinamento

## Problemas

- **Overfitting**: especialização nos dados de treinamento, não generaliza
- **Underfitting**: baixo acerto mesmo nos dados de treinamento

# Overfitting

- Sobreajuste ou overtraining
- Fenômeno que ocorre quando o modelo estatístico se ajusta em demasiado ao conjunto de dados/amostras
- É comum que a amostra apresente dsvios causados por erros de medição ou fatores aleatórios, ocorre o sobreajuste quando o modelo se ajusta a estes

## Teorema do patinho feio (de Watanabe)

Caso haja um conjunto suficientemente grande de características em comum, sem uma outra referência previamente estabelecida, é possível fazer com que dois padrões arbitrários sejam considerados similares  
Um cisne e um pato e um par de cisnes podem ficar igualmente similares

# Overfitting





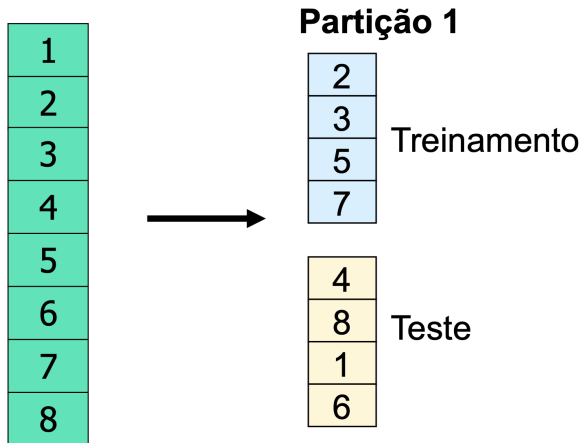
# Overfitting

- Como contornar esse problema?
- **Cross-validation**
  - ▶ consiste em separar os dados em treinamento e teste
    - ★ conjunto de **treinamento** usado para o aprendizado do conceito
    - ★ conjunto de **teste** usado para medir o grau de efetividade do conceito aprendido
  - ▶ essa divisão dos dados em subconjuntos ajuda a evitar que o modelo aprenda as particularidades dos dados

# Amostragem de Dados

- Permite melhor avaliação do desempenho preditivo
- Alternativas
  - ▶ amostragem única
    - ★ hold-out
  - ▶ re-amostragem

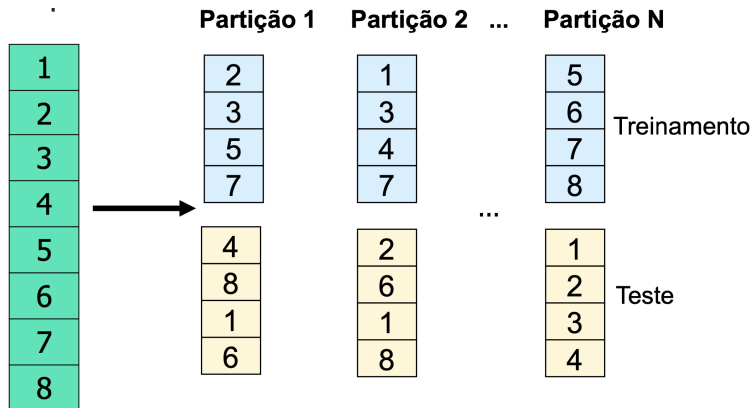
# Hold-out



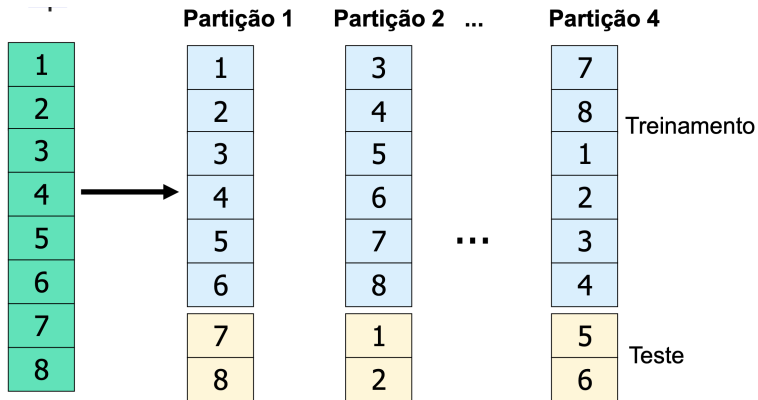
# Métodos de Reamostragem

- Utilizam várias partições para os conjuntos de treinamento e de teste
  - ▶ random subsampling
  - ▶ k-fold cross-validation
    - ★ leave-one-out
  - ▶ bootstrap

# Random Subsampling



# k-fold Cross-Validation



# Leave-one-out

- Cross-validation, em que  $k = n$ , sendo que  $n$  representa o número de amostras disponíveis
- Estimativa de erro é praticamente não tendenciosa
  - ▶ média das estimativas tende a taxa de erro verdadeiro
- Computacionalmente caro
  - ▶ geralmente utilizado para pequenos conjuntos de exemplos
  - ▶ 10-fold cross validation aproxima leave-one-out
- Variância tende a ser elevada

# Bootstrap

- Funciona melhor que cross-validation para conjuntos muito pequenos
- Existem diversas variações
- Forma mais simples de bootstrap
  - ▶ amostragem com reposição
    - ★ cada partição é uma amostra aleatória com reposição do conjunto total de exemplos
    - ★ conjunto de treinamento têm o mesmo número de exemplos do conjunto total
    - ★ exemplos que restarem são utilizados para teste



# Acurácia

- Quantos exemplos foram corretamente classificados?
  - ▶ avalia erro nas classes igualmente
- Pode não ser adequada para dados desbalanceados
  - ▶ pode prejudicar desempenho para classe minoritária
    - ★ geralmente mais interessante que a classe majoritária
  - ▶ acurácia balanceada

# Classificação Binária

- Classe de interesse é a classe positiva
- Dois tipos de erro:
  - ▶ classificação de um exemplo N como P
    - ★ falso positivo (alarme falso)
    - ★ ex.: diagnosticado como doente, mas está saudável
  - ▶ classificação de um exemplo P como N
    - ★ falso negativo
    - ★ ex.: diagnosticado como saudável, mas está doente


# Desempenho Preditivo

- Matriz de confusão (tabela de contingência) pode ser utilizada para distinguir os erros
  - ▶ base de várias medidas
  - ▶ pode ser utilizada com 2 ou mais classes

Classe verdadeira	Classe predita		
	1	2	3
1	25	0	5
2	10	40	0
3	0	0	20

# Exemplo

- Matriz de confusão para 200 exemplos divididos em 2 classes



Classe verdadeira	Classe predita	
	p	n
P	70	30
N	40	60

Classe verdadeira	Classe predita	
	p	n
P	VP	FN
N	FP	VN

# Medidas de Avaliação

$$\text{Taxa de FP (TFP)} = \frac{FP}{FP + VN}$$

(Alarmes falsos)

Erro do tipo I

		Classe predita	
		p	n
Classe verdadeira	P	VP	FN
	N	FP	VN

$$\text{Taxa de FN (TFN)} = \frac{FN}{VP + FN}$$

Erro do tipo II

		Classe predita	
		p	n
Classe verdadeira	P	VP	FN
	N	FP	VN

# Medidas de Avaliação

$$\text{Taxa de FP (TFP)} = \frac{FP}{FP + VN}$$

(Alarmes falsos)

Custo

		Classe predita	
		p	n
Classe verdadeira	P	VP	FN
	N	FP	VN

$$\text{Taxa de VP (TVP)} = \frac{VP}{VP + FN}$$

Benefício

		Classe predita	
		p	n
Classe verdadeira	P	VP	FN
	N	FP	VN

# Exemplo

- Avaliação de 3 classificadores

$$\frac{VP}{VP+FN}$$

$$\frac{FP}{FP+VN}$$

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

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

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

## Classificador 1

TVP =

TFP =

## Classificador 2

TVP =

TFP =

## Classificador 3

TVP =

TFP =

# Exemplo

- Avaliação de 3 classificadores

$$\frac{VP}{VP+FN}$$

$$\frac{FP}{FP+VN}$$

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

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

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

## Classificador 1

$$TVP = 0.4$$

$$TFP = 0.3$$

## Classificador 2

$$TVP = 0.7$$

$$TFP = 0.5$$

## Classificador 3

$$TVP = 0.6$$

$$TFP = 0.2$$



# Medidas de Avaliação

- Medidas frequentemente utilizadas

TFP (Erro tipo I)

$$\frac{FP}{FP+VN}$$

Precisão

$$\frac{VP}{VP+FP}$$

TVP

$$\frac{VP}{VP+FN}$$

Sensibilidade  
Revocação (Recall)

TFN (Erro tipo II)

$$\frac{FN}{VP+FN}$$

Especificidade

$$\frac{VN}{VN+FP} = 1 - \text{TFP}$$

Acurácia

$$\frac{VP+VN}{VP+VN+FP+FN}$$

Medida F1

$$\frac{2}{1/prec+1/rev}$$

# Revocação x Precisão

- Revocação (recall)
  - ▶ porcentagem de exemplos positivos classificados como positivos
    - ★ nenhum exemplo positivo é deixado de fora
- Precisão
  - ▶ porcentagem de exemplos classificados como positivos que são realmente positivos
    - ★ nenhum exemplo negativo é incluído

Revocação

$$\frac{VP}{VP+FN}$$

Precisão

$$\frac{VP}{VP+FP}$$

# Sensibilidade x Especificidade

- Sensibilidade

- ▶ porcentagem de exemplos positivos classificados como positivos
  - ★ igual a revocação

- Especificidade

- ▶ porcentagem de exemplos negativos classificados como negativos
  - ★ nenhum exemplo negativo é deixado de fora

sensibilidade

$$\frac{VP}{VP+FN}$$

Especificidade

$$\frac{VN}{VN+FP}$$

# Avaliação

- Medida-F
  - ▶ média harmônica ponderada da precisão e da revocação
- Medida-F1
  - ▶ precisão e revocação têm o mesmo peso

## Medida-F

$$\frac{(1+\alpha) \times (prec \times rev)}{\alpha \times prec + rev}$$

## Medida-F1

$$\frac{2 \times (prec \times rev)}{prec + rev} = \frac{2}{1 / prec + 1 / rev}$$

# Exemplo

- Seja um classificador com a seguinte matriz de confusão, definir:
  - ▶ acurácia
  - ▶ precisão
  - ▶ revocação
  - ▶ especificidade

		Classe predita	
		p	n
Classe verdadeira	P	70	30
	N	40	60

# Exemplo

## Acurácia

$$\frac{VP+VN}{VP+VN+FP+FN}$$

## Precisão

$$\frac{VP}{VP+FP}$$

## Revocação

$$\frac{VP}{VP+FN}$$

## Especificidade

$$\frac{VN}{VN+FP}$$

		Predito	
		p	n
Verdadeiro	P	VP	FN
	N	FP	VN
	P	70	30
	N	40	60

# Exemplo

## Acurácia

$$\frac{VP+VN}{VP+VN+FP+FN} = (70+60) / (70+30+40+60) = 0.65$$

## Precisão

$$\frac{VP}{VP+FP} = 70 / (70+40) = 0.64$$

## Revocação

$$\frac{VP}{VP+FN} = 70 / (70+30) = 0.70$$

## Especificidade

$$\frac{VN}{VN+FP} = 60 / (40+60) = 0.60$$

		Predito	
		p	n
Verdadeiro	P	VP	FN
	N	FP	VN
		p	n
	P	70	30
	N	40	60

# Gráficos ROC

- Do inglês, Receiver operating characteristics
- Medida de desempenho originária da área de processamento de sinais
  - ▶ muito utilizada nas áreas médica e biológica
  - ▶ mostra relação entre custo (TFP) e benefício (TVP)

$$\frac{FP}{FP+VN} \times \frac{VP}{VP+FN}$$



# Exemplo

- Colocar no gráfico ROC os 3 classificadores do exemplo anterior

Classificador 1

TFP = 0.3

TVP = 0.4



Classificador 2

TFP = 0.5

TVP = 0.7



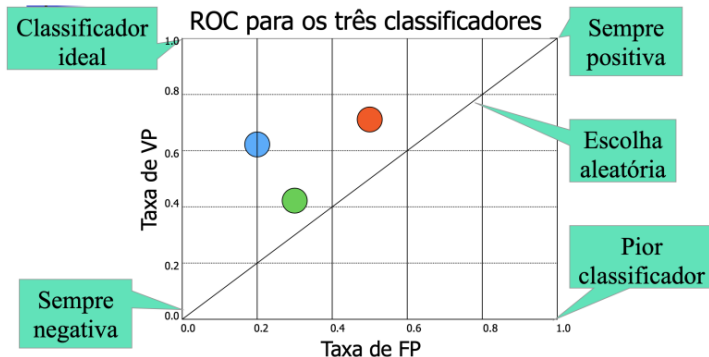
Classificador 3

TFP = 0.2

TVP = 0.6



# Gráficos ROC



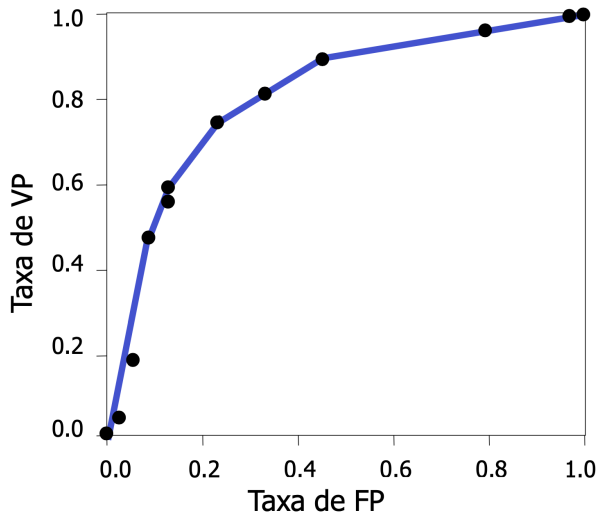
# Gráficos ROC

- Classificadores discretos produzem um simples ponto no gráfico ROC
  - ▶ ADs e conjuntos de regras
- Outros classificadores produzem uma probabilidade ou score
  - ▶ RNAs e NB
- Curvas ROC permitem uma melhor comparação de classificadores
  - ▶ são insensíveis a mudanças na distribuição das classes

# Curvas ROC

- Mostram ROC para diferentes variações
- Classificadores que geram valores contínuos (threshold, probabilidade)
  - ▶ diferentes valores de threshold podem ser utilizados para gerar vários pontos
    - ★ ligação dos pontos gera uma curva ROC
- Classificadores discretos
  - ▶ convertidos internamente ou comitês

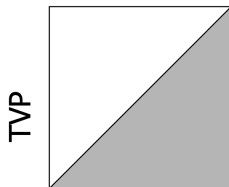
# Curvas ROC



# Área sob a curva ROC (AUC)

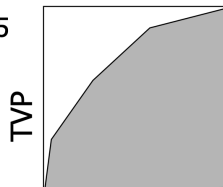
- Fornece uma estimativa do desempenho de classificadores
- Gera um valor contínuo no intervalo  $[0,1]$ 
  - ▶ quanto maior melhor
  - ▶ adição de áreas de sucessivos trapezóides
- Um classificador com maior AUC pode apresentar AUC pior em trechos da curva
- Mais confiável utilizar médias de AUCs

# Área Sob Curvas ROC



Área = 0,5  
Nenhuma

TFP



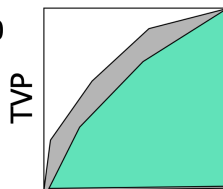
Área = 0,74

TFP



Área = 1,0  
Perfeita

TFP



Área = 0,74  
Área = 0,67

TFP

- Teste de Hipóteses

- ▶ permite afirmar que uma técnica é melhor que outra com X% de confiança
- ▶ podem assumir que os dados seguem uma dada distribuição de probabilidade
  - ★ paramétricos
  - ★ não paramétricos
- ▶ número de técnicas comparadas
  - ★ duas
  - ★ mais que duas



# Referências e Leituras Complementares

- Cap. 18 → livro Russel e Norvig
- Cap. 10 → livro Ben Coppin