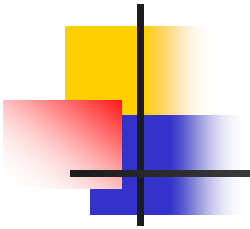
Ciência de Dados

Planejamento de experimentos

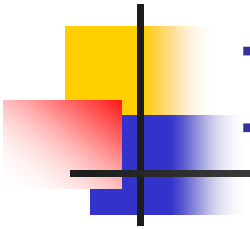
Revisado: Roseli Romero
SCC-ICMC-USP

Prof. Dr. André C. P. L. F. de Carvalho
Dr. Isvani Frias-Blanco
ICMC-USP



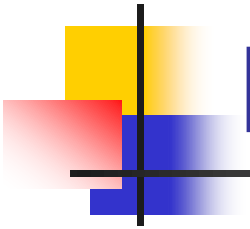
Principais tópicos

- Desempenho preditivo
- Partição dos dados
- Reamostragem
- Tipos de erro
- Avaliação do desempenho
- Curvas ROC



Introdução

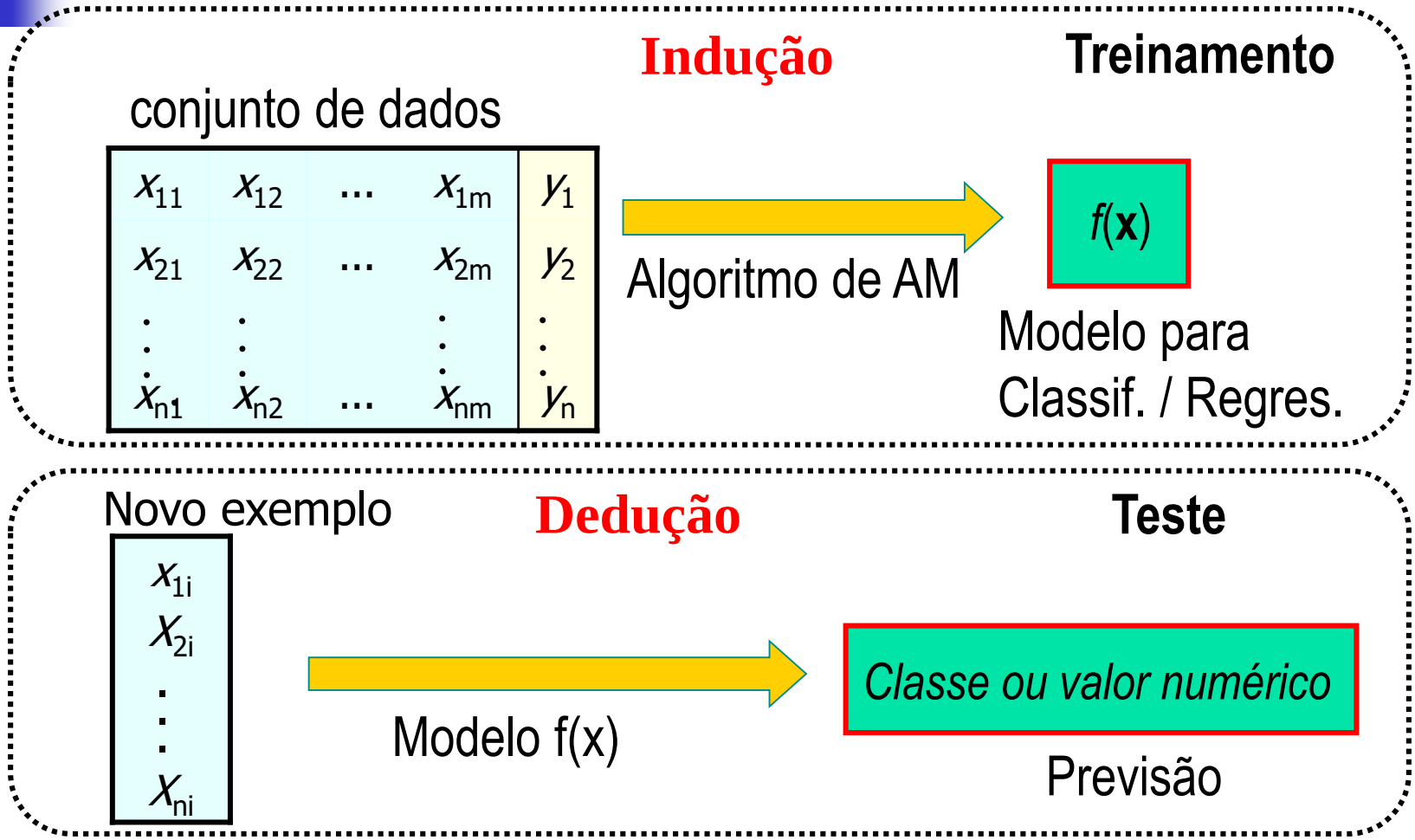
- Após a exploração e pré-processamento vem a modelagem
  - Permite avaliar benefícios do pré-processamento
    - E eventualmente retornar para as fases de exploração e pré-processamento
- Procedimentos experimentais e avaliação de desempenho
  - Diferente para tarefas descritivas e preditivas
  - Este módulo tratará de tarefas preditivas (classificação e regressão)

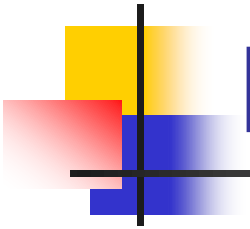


Desempenho

- Preditivo
  - Tarefa de classificação
  - Tarefa de regressão
- Custo
  - Tempo de processamento
  - Memória necessária
- Algoritmo e/ou modelo

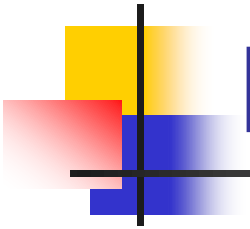
Tarefa preditiva





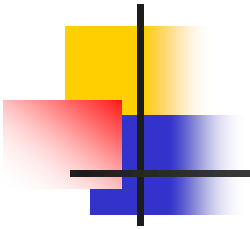
Desempenho preditivo

- Depende da tarefa a ser resolvida:
  - Classificação: considera taxa de exemplos incorretamente classificados
    - Acurácia
  - Regressão: considera diferença entre valor previsto e valor correto
  - Agrupamento: diferentes critérios
- Média dos erros obtidos em diferentes execuções de um experimento



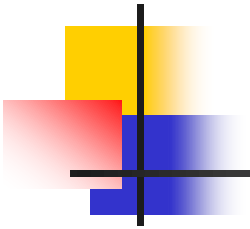
Desempenho preditivo

- Pode ser avaliado para:
  - Buscar o melhor modelo(s) de classificação
    - Gerados pelo mesmo algoritmo, variando
      - Valores de hiperparâmetros
      - Partições/atributos nos dado de treinamento
    - Para escolher melhor modelo preditivo
  - Buscar melhor algoritmo(s) de classificação
    - Avalia modelos gerados (funções, hipóteses)
      - Hiper-parâmtros de cada algoritmo com valores default ou otimizados
      - Conjuntos de dados com mesmas partições e atributos preditivos
    - Para escolher melhor algoritmo preditivo



Desempenho preditivo

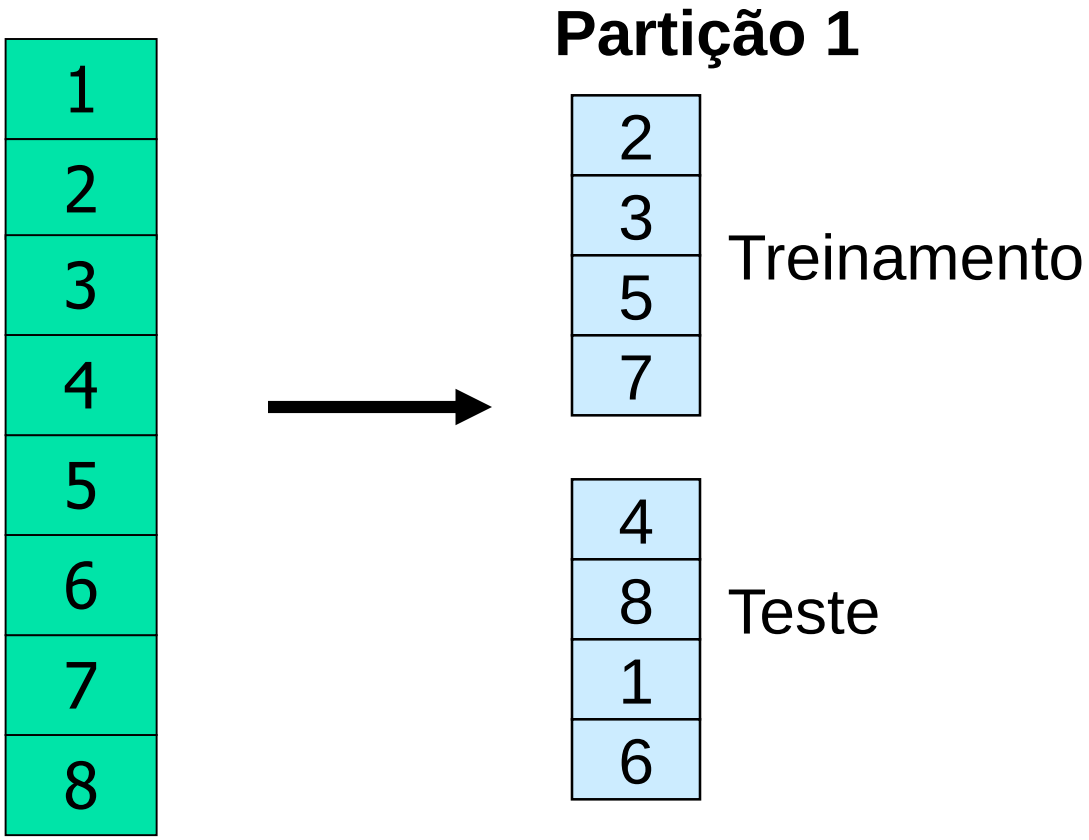
- Principal objetivo:
  - Classificação correta de novos exemplos
    - Errar o mínimo possível
      - Minimizar taxa de erro para novos exemplos
- Geralmente não é possível medir com exatidão essa taxa de erro
  - Deve ser estimada
    - Para uma amostra de teste (simula novos exemplos) do conjunto de dados disponível
    - Utilizando modelo induzido com uma amostra de treinamento do conjunto de dados disponível

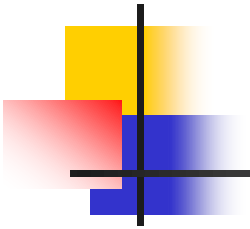


Amostragem de dados

- Permite melhor estimativa do desempenho de um modelo ou algoritmo
  - Treinamento (validação) e teste
- Procedimentos
  - Amostragem única
    - *Hold-out*
  - Re-amostragem

Hold-out

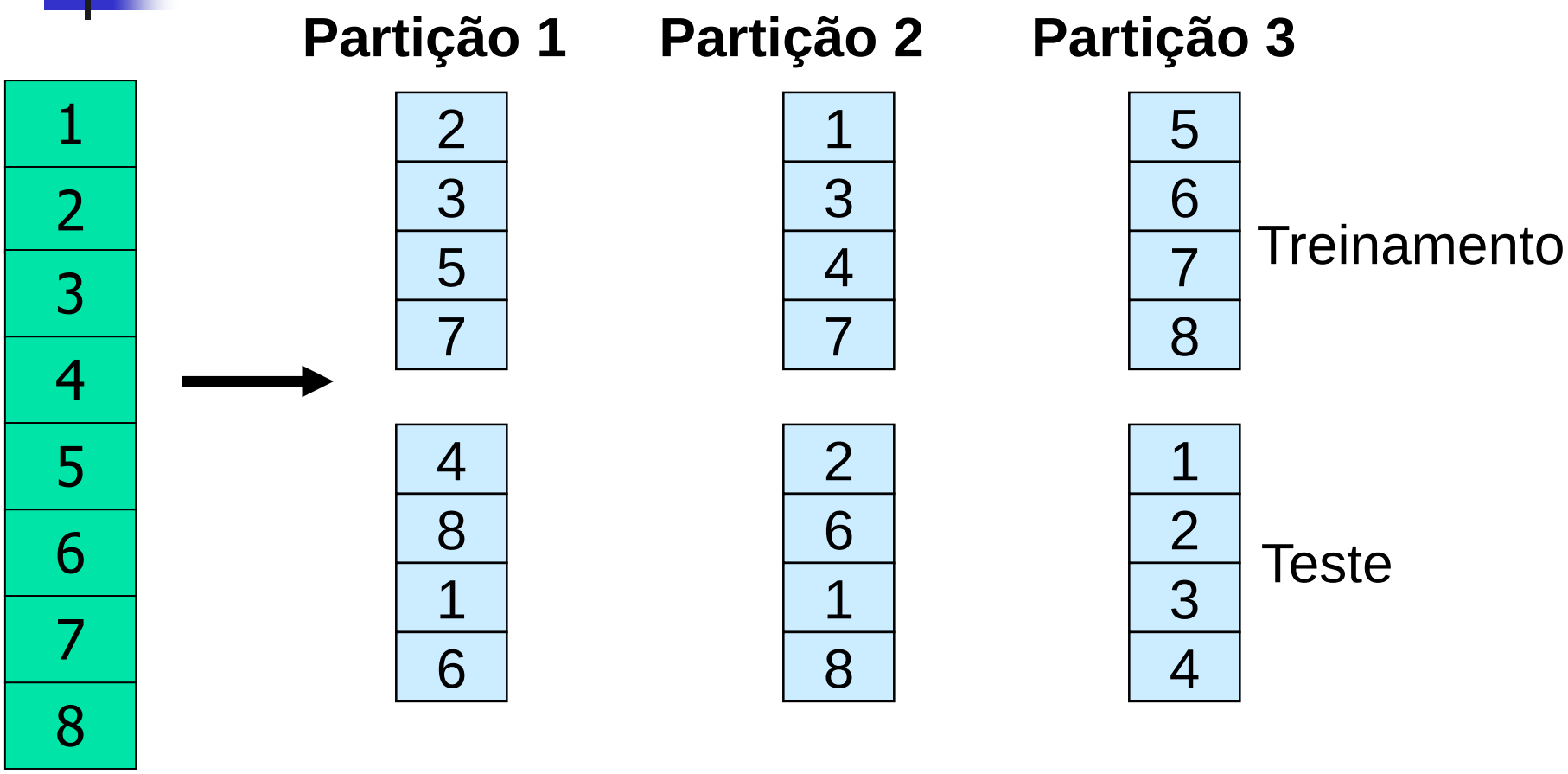




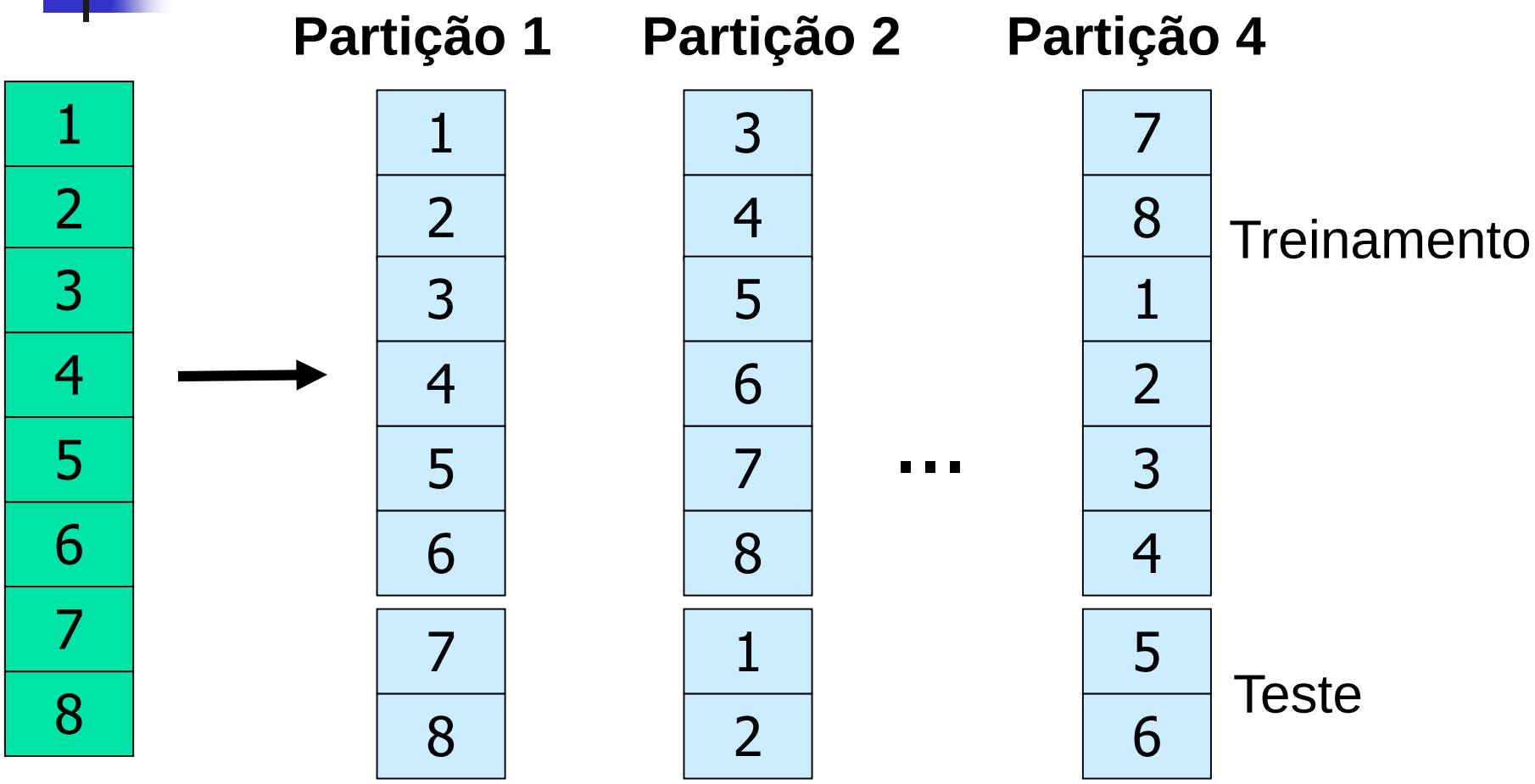
Métodos de reamostragem

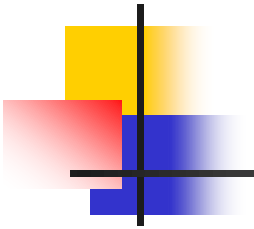
- Amostragem única é pouco confiável
- Geram várias partições para conjuntos de treinamento e teste (validação)
  - *Random subsampling*
  - *K-fold Cross-validation*
    - *Leave-one-out*
  - *Bootstrap (ou Bootstrapping)*

Random subsampling



K-fold cross-validation





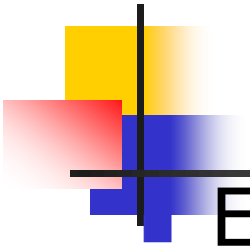
Leave-one-out

- Estimativa de erro praticamente não tendenciosa
  - Tende a taxa de erro verdadeiro
- Computacionalmente caro para conjuntos grandes
  - Geralmente utilizado para pequenos conjuntos de dados
  - 10-fold cross validation aproxima leave-one-out
- Variância tende a ser elevada

5 x 2 Cross-validation

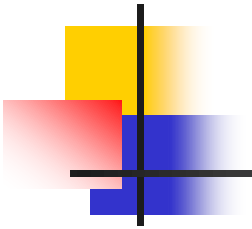
- Conjuntos de treinamento e teste com mesmo tamanho

Seja um conjunto de N exemplos
Para i = 1 até 5
Dividir N aleatoriamente em duas metades
Usar metade 1 para treinamento e metade 2 para teste
Usar metade 2 para treinamento e metade 1 para teste



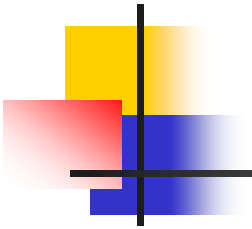
Bootstrap

- Estocástico, com diversas variações
  - Alguns exemplos podem não participar do treinamento
- Variação mais simples:
  - 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
  - Esta reamostragem é feita muitas vezes(de 1000 a 10000 vezes) para criar uma estimativa da função de distribuição acumulada.



Bootstrap

- Se conjunto original tem N exemplos
  - A probabilidade de um exemplo não ser amostrado é de: (1-1/n)^n ~e^-1~0.368.
  - Amostra de tamanho N tem ≈ 63,2% dos exemplos originais
- Processo é repetido k vezes
  - Resultado final é a média dos k experimentos



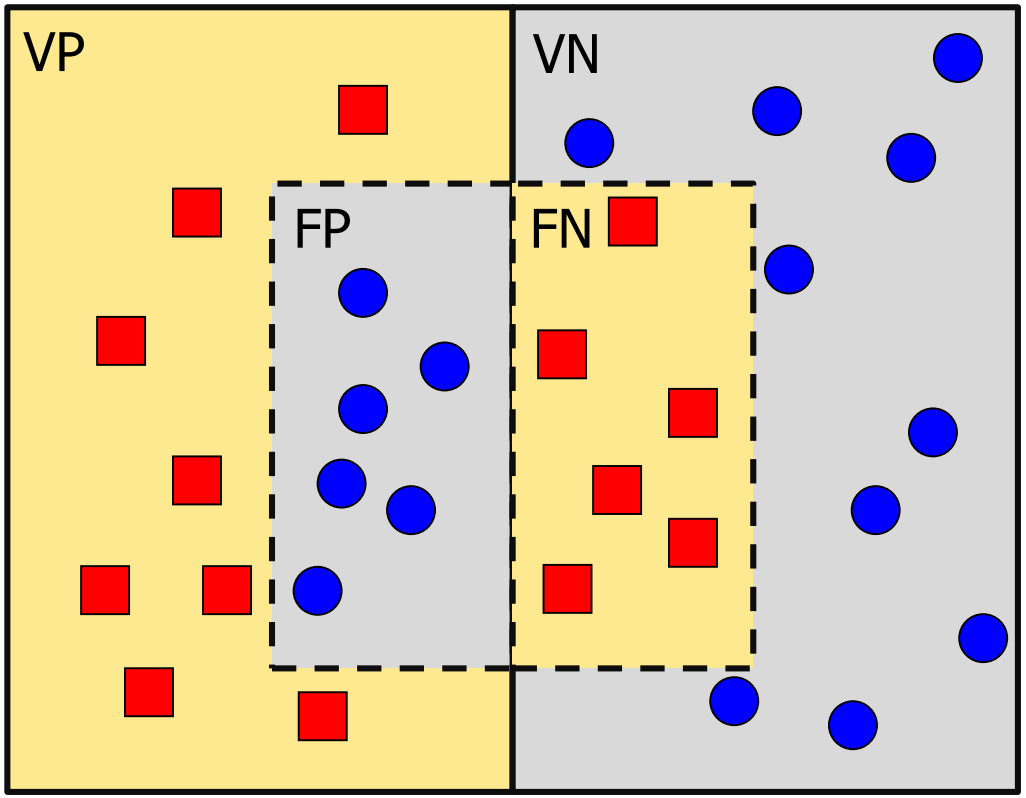
Bootstrap

- Estima incerteza de um algoritmo
  - *K-fold cross-validation* é mais usado para estimar acurácia preditiva
    - Seleção de algoritmos/modelos
- Tende a ter menor variância e ser mais pessimista que *k-fold cross-validation*

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

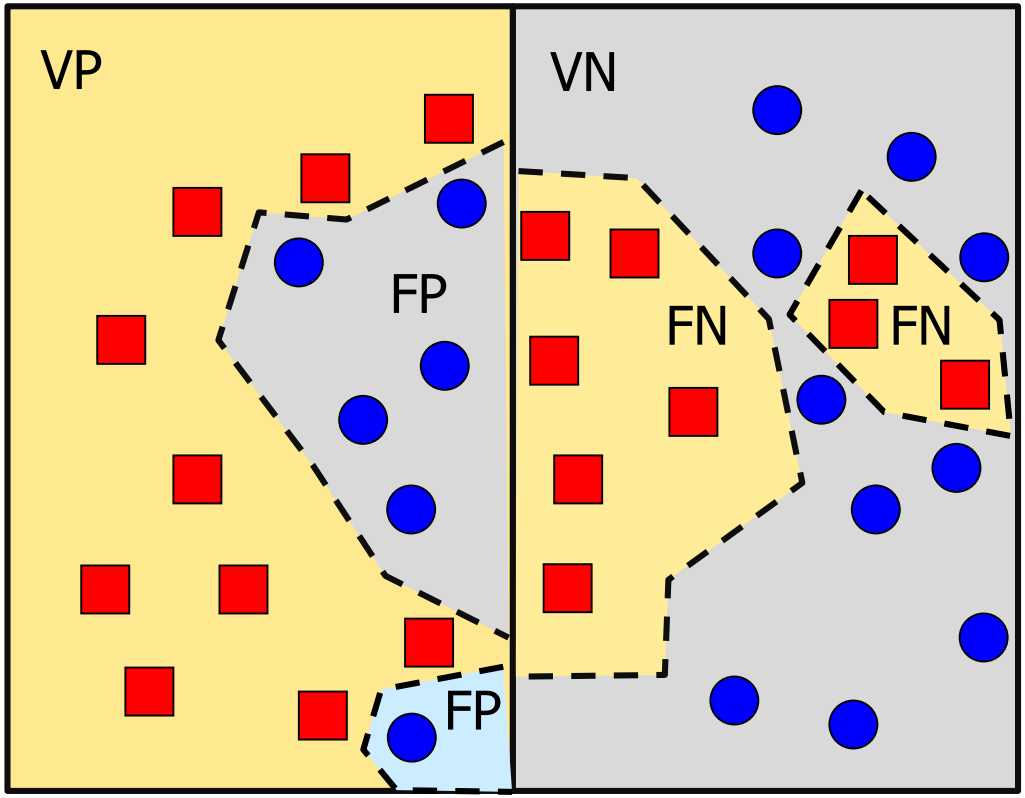
Classificação binária



Predito como positivo

Predito como negativo

Classificação binária



Predito como positivo

Predito como negativo

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

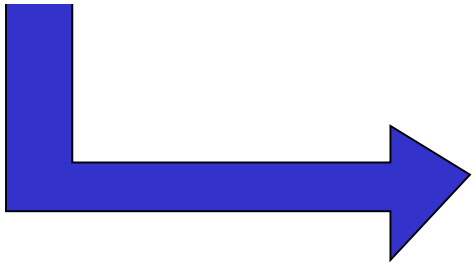
Exemplo

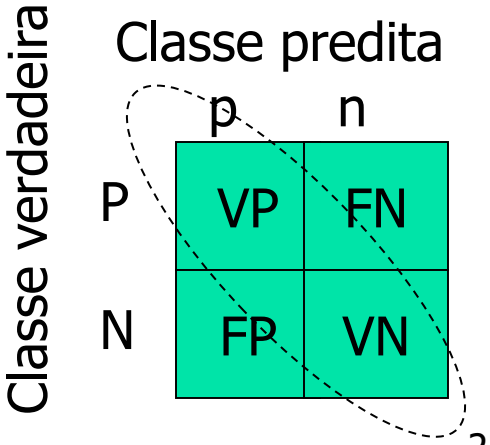
- Matriz de confusão para 200 exemplos divididos em 2 classes

Classe verdadeira

Classe predita

	p	n
P	70	30
N	40	60



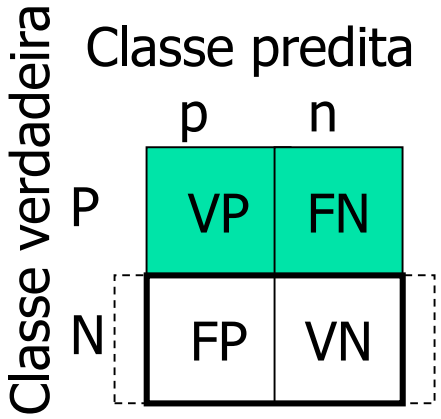


Medidas de avaliação

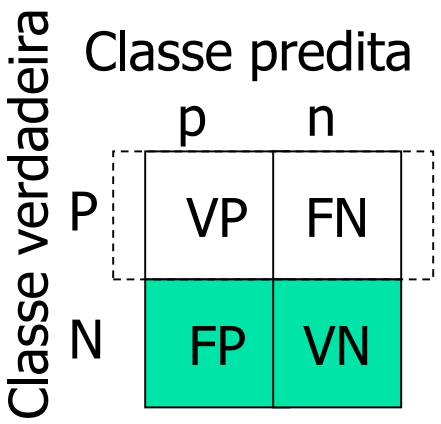
Taxa de FP (TFP) = $\frac{FP}{FP + VN}$
(Alarmes falsos)

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

Erro do tipo I



Erro do tipo II



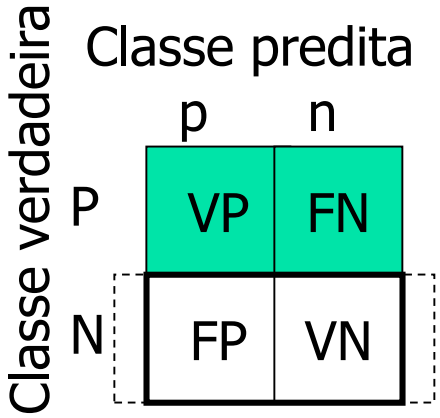
Medidas de avaliação

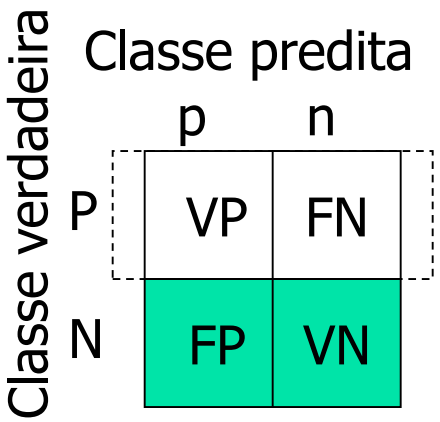
Taxa de FP (TFP) = (Alarmes falsos) $\frac{FP}{FP+VN}$

Custo

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

Benefício





Exemplo

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

- Avaliação de 3 classificadores

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

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

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

Classificador 1
TVP =
TFP =

Classificador 2
TVP =
TFP =

Classificador 3
TVP =
TFP =

Exemplo

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

- Avaliação de 3 classificadores

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

Classificador 1
TVP = 0.4
TFP = 0.3

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

Classificador 2
TVP = 0.7
TFP = 0.5

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

Classificador 3
TVP = 0.6
TFP = 0.2

Medidas de avaliação

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

Taxa de falso positivo (TFP) = 1-TVN

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

Taxa de falso negativo (TFN) = 1-TVP

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

Valor predito positivo (VPP), precisão

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

Valor predito negativo (VPN)

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

Taxa de verdadeiro positivo (TVP), Sensibilidade ou Revocação (Recall)

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

Taxa de verdadeiro negativo (TVN), especificidade

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

Acurácia

$$\frac{2}{1/\,prec.+1/\,revoc.}$$

Medida-F1

Medidas de avaliação

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

False positive rate (FPR) = 1-TNR

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

False negative rate (FNR) = 1-TPR

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

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

True positive rate (TPR), also known as recall or sensitivity

True negative rate (TNR), also known as specificity

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

Positive predictive value (PPV), also known as precision

$$\frac{TN}{TN+FN}$$

Negative predictive value (NPV)

$$\frac{TP+TN}{TP+TN+FP+FN}$$

Accuracy

$$\frac{2}{1/\,precision+1/\,recall}$$

F1-measure