# Mineração de Dados em Biologia Molecular

## Planejamento e Análise de Experimentos Preditivos

Docente: André C. P. L. F. de Carvalho
PAE: Victor Hugo Barella

## Principais tópicos

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



## Tarefa preditiva

- Indução de modelo com bom desempenho preditivo para novos dados
  - Escolher algoritmo com viés (bias) adequado para o conjunto de dados utilizado
    - Necessário avaliar o desempenho preditivo de algoritmos
      - Avaliar desempenho preditivo de modelos induzidos pelos algoritmos
      - Modelo induzido pelo algoritmo deve estimar corretamente o rótulo de novos exemplos



## Algoritmos e modelos

- Desempenho preditivo deve ser avaliado para
  - Algoritmos de AM
    - Saída de algoritmos de AM:
      - Modelos, funções, hipóteses preditivos
  - Modelos preditivos
    - Saída de modelos preditivos
      - Classe ou valor para um novo exemplo



## Desempenho de modelos preditivos

- Depende da tarefa:
  - Classificação: considera taxa de exemplos incorretamente classificados
    - Acurácia ou outras similares
  - Regressão: considera diferença entre valor previsto e valor correto
    - MSE ou outras similares
- Média das taxas de erro obtidas em diferentes execuções



## Tarefa de classufucação

- Objetivo de um modelo de classificação:
  - Estimar corretamente o rótulo de novos exemplos (errar o mínimo possível)
    - Minimizar taxa de erro de classificação
  - Geralmente não é possível medir com exatidão essa taxa de erro
    - Ela deve ser estimada
    - Amostragem de dados

## Amostragem de dados

- Divide conjunto de dados em subconjuntos
- Subconjunto de treinamento
  - Ajuste de híper-parâmetros de um algoritmo
  - Comparação de algoritmos ou de híper-parâmetros de um algoritmo
    - Para dados de treinamento
- Subconjunto de teste
  - Comparação de algoritmos/modelos para novos dados
  - Nunca para escolha

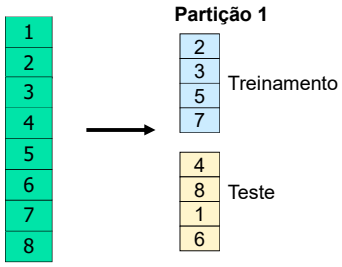


## Amostragem de dados

- Permite melhor avaliação do desempenho de um classificador
  - Pode induzir um ou mais modelos
- Alternativas
  - Amostragem única
    - *Hold-out*
  - Re-amostragem

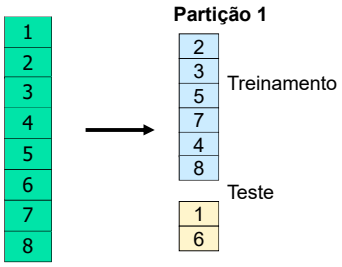


## *Hold-out*






## *Hold-out*



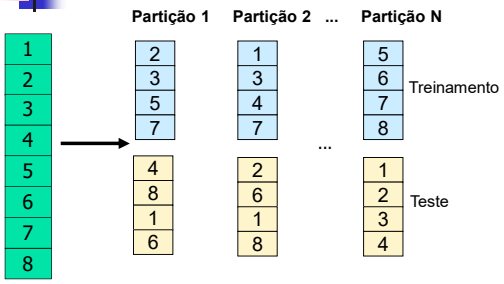



## Métodos de re-amostragem

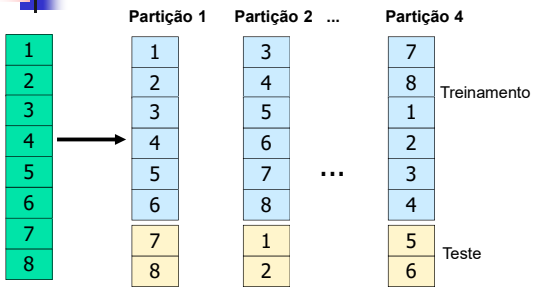
- Utilizam várias partições para os conjuntos de treinamento e teste
  - *Random subsampling*
  - *K-fold Cross-validation*
    - *Leave-one-out*
  - *Bootstrap*



## Random subsampling



## K-fold cross-validation






## Leave-one-out

- Estimativa de erro é praticamente não tendenciosa
  - Média das estimativas tende a taxa de erro verdadeira
- 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
- 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



## Bootstrap

- Se conjunto original tem N exemplos
  - Amostra de tamanho N tem ≈ 63,2% dos exemplos originais
    - Viés semelhante a 2-fold cross-validation
- Processo é repetido k vezes
  - Resultado final = média dos k experimentos
- Existem diversas variações



## Medidas de avaliação

- Acurácia
  - Trata as classes igualmente
  - Pode não ser adequada para dados desbalanceados
    - Pode prejudicar desempenho para classe minoritária
    - Classe rara é geralmente mais interessante que a majoritária
    - Alternativa: 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

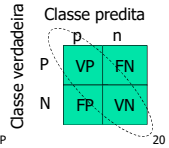
| Classe verdadeira \ Classe predita | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 25 | 0 | 5 |
| 2 | 10 | 40 | 0 |
| 3 | 0 | 0 | 20 |



## Exemplo

- Matriz de confusão para 200 exemplos divididos em 2 classes

| Classe verdadeira \ Classe predita | p | n |
|---|---|---|
| P | 70 | 30 |
| N | 40 | 60 |






## Medidas de avaliação

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

Erro do tipo I

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

Erro do tipo II



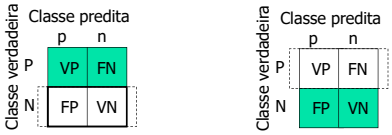

## Medidas de avaliação

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

Custo

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

Benefício

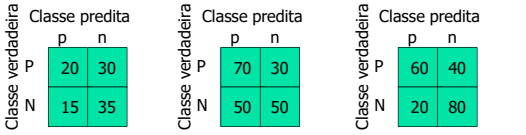





## Exemplo

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

- Avaliação de 3 classificadores



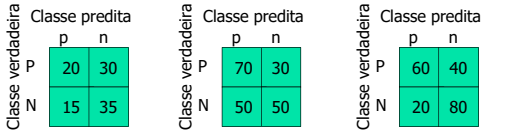

| Classificador 1 | Classificador 2 | Classificador 3 |
|---|---|---|
| TVP = | TVP = | TVP = |
| TFP = | TFP = | TFP = |



## Exemplo

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

- Avaliação de 3 classificadores




| Classificador 1 | Classificador 2 | Classificador 3 |
|---|---|---|
| TVP = 0.4 | TVP = 0.7 | TVP = 0.6 |
| TFP = 0.3 | TFP = 0.5 | TFP = 0.2 |

## Medidas de avaliação

- Medidas frequentemente utilizadas

TFP = (Erro tipo I) $\frac{FP}{VN+FP}$

TFN = (Erro tipo II) $\frac{FN}{VP+FN}$

Precisão = $\frac{VP}{VP+FP}$

TVP = $\frac{VP}{VP+FN}$
Sensibilidade
Revocação (Recall)

Especificidade = $\frac{VN}{VN+FP}$ = 1–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 $\frac{VP}{VP+FN}$
    - Nenhum exemplo positivo é deixado de fora
- Precisão
  - Porcentagem de exemplos classificados como positivos que são realmente positivos
    - Nenhum exemplo negativo é incluído $\frac{VP}{VP+FP}$



## Sensibilidade X Especificidade

- Sensibilidade $\frac{VP}{VP+FN}$
  - Porcentagem de exemplos positivos classificados como positivos
    - Igual a revocação
- Especificidade $\frac{VN}{VN+FP}$
  - Porcentagem de exemplos negativos classificados como negativos
    - Nenhum exemplo negativo é deixado de fora



## Medidas de avaliação

- Medida-F
  - Média harmônica ponderada da precisão e da revocação

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

  - Medida-F1
    - Precisão e revocação têm o mesmo peso

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



## Exemplo

- Seja um classificador com a seguinte matriz de confusão, definir:
  - Acurácia
  - Precisão
  - Revocação
  - Especificidade

| Classe verdadeira | Classe predita p | Classe predita n |
|---|---|---|
| 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 = $\frac{VN}{VN+FP}$

| Verdadeiro | Predito p | Predito n |
|---|---|---|
| P | VP | FN |
| N | FP | VN |

| | p | n |
|---|---|---|
| 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 = $\frac{VP}{VP+FN}$ = 7/(70+30) = 0.70

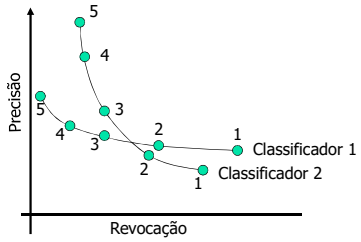
Especificidade = $\frac{VN}{VN+FP}$ = 60/(40+60) = 0.60

| Verdadeiro \ Predito | p | n |
|---|---|---|
| P | VP | FN |
| N | FP | VN |

| | p | n |
|---|---|---|
| P | 70 | 30 |
| N | 40 | 60 |



## Observação






## Outras medidas

- Média geométrica de taxas positivas
  - G-means

$$\sqrt{precisão \times revocação}$$

- Acurácia balanceada
- Kappa



## 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)



## Exemplo

- Colocar no gráfico ROC os 3 classificadores do exemplo anterior

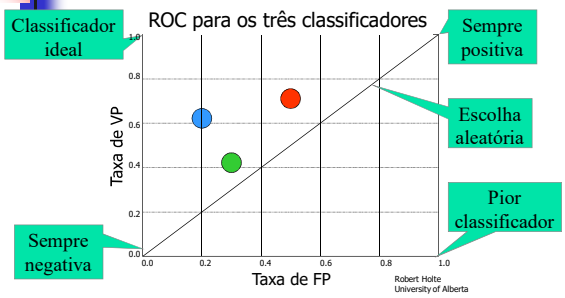
Classificador 1
TFP = 0.3
TVP = 0.4

Classificador2
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 escore
  - RNAs e NB
  - Permitem gerar vários pontos
- Curvas ROC permitem uma melhor comparação de classificadores
  - Insensíveis a mudanças na distribuição das classes

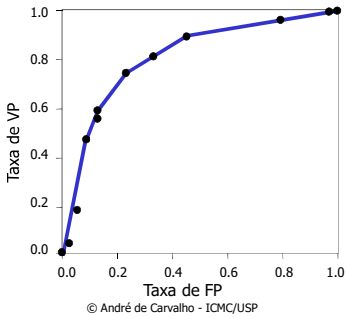


## Curvas ROC

- Mostram pontos no gráfico para diferentes variações de um classificador
- 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



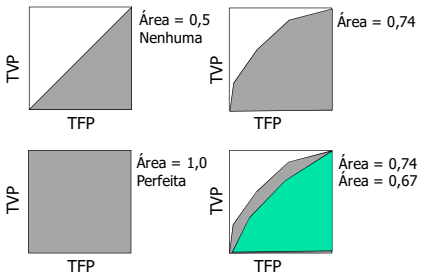



## Área sob a curva ROC (AUC)

- Fornece uma estimativa do desempenho de classificadores
- Gera um valor continuo no intervalo [0.0, 1.0]
  - Quanto maior melhor
  - Adição de áreas de sucessivos trapezóides
- É mais confiável utilizar médias de AUCs



## Área Sob Curvas ROC






## Avaliação de Desempenho

- 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
    - Testes paramétricos
    - Testes não paramétricos
  - Número de técnicas comparadas
    - Duas
    - Mais que duas

## Considerações Finais

- Desempenho preditivo
- Avaliação do desempenho
  - Erro
  - Tempo de resposta
  - Memória
  - Representação
- Medidas
- Gráficos e curvas ROC
- Teste de hipóteses



## Perguntas