
Web Data Mining com R: design de projetos para criação de modelos preditivos

Fabrício Jailson Barth

Faculdade BandTec e VAGAS Tecnologia

Junho de 2013

Sumário e Objetivos

- Etapas em estudos preditivos
- Escolha dos dados
- Medidas de erro

Etapas em estudos preditivos

- Escolher o conjunto de dados corretos.
- Dividir os dados em:
 - ★ Treinamento.
 - ★ Teste.
 - ★ Validação (opcional).

-
- Selecionar atributos que devem formar o conjunto de treinamento.
 - Identificar modelos preditivos usando o conjunto de treinamento.

-
- Aplicar *cross-validation* sobre o conjunto de treinamento.
 - Se não existe conjunto de validação então aplicar o modelo 1x no conjunto de teste.
 - Se existe conjunto de validação então aplicar o modelo no conjunto de teste e refinar o modelo.
 - Se existe conjunto de validação então aplicar o modelo 1x no conjunto de validação.

Identificando o conjunto de dados corretos

- Em alguns casos é fácil (avaliação de filmes → novas avaliações de filmes).
- Em outros pode ser mais difícil (dados genéticos → doenças).
- Geralmente, quanto maior a quantidade de dados, melhor são os modelos.
- Conhecer *bench marks* ajuda!
- **Sempre começamos com dados brutos e precisamos processá-los.**

Definição de Erro

Table 1: Conjunto de teste

Exemplo	Classe real	Classe inferida
1	Positivo	Positivo
2	Positivo	Negativo
3	Negativo	Negativo
4	Negativo	Negativo
5	Negativo	Negativo
6	Positivo	Positivo
7	Positivo	Negativo
8	Negativo	Negativo

$$\text{erro}(\text{modelo}) = \frac{qtd_incorretos}{qtd_exemplos} \quad (1)$$

onde:

- *qtd_exemplos*: quantidade de exemplos do conjunto de teste.
- *qtd_corretos*: quantidade de exemplos do conjunto de teste incorretamente classificados.

Neste exemplo:

Table 2: Conjunto de teste

Exemplo	Classe real	Classe inferida
1	Positivo	Positivo
2	Positivo	Negativo
3	Negativo	Negativo
4	Negativo	Negativo
5	Negativo	Negativo
6	Positivo	Positivo
7	Positivo	Negativo
8	Negativo	Negativo

$$erro(modelo) = \frac{2}{8} = 0.25 \quad (2)$$

Definição de Verdadeiro e Falso Positivo

- Verdadeiro Positivo = identificado corretamente.
- Falso Positivo = identificado incorretamente.
- Verdadeiro Negativo = rejeitado corretamente.
- Falso Negativo = rejeitado incorretamente.

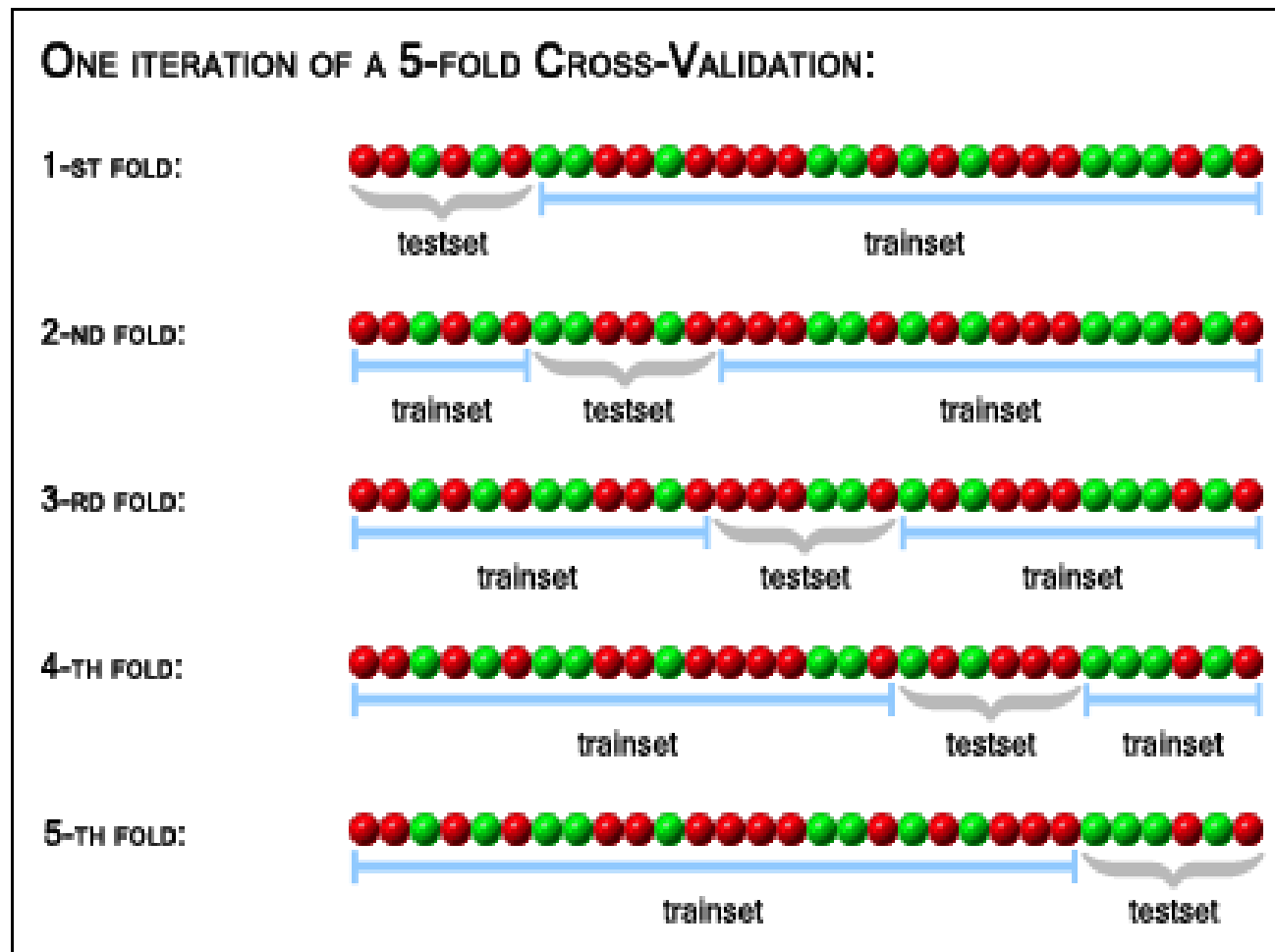
Exemplo de teste médico:

- Verdadeiro Positivo = Pessoa doente corretamente classificada como doente.
- Falso Positivo = Pessoa saudável incorretamente classificada como doente.
- Verdadeiro Negativo = Pessoa saudável corretamente classificada como saudável.
- Falso Negativo = Pessoa doente incorretamente classificada como saudável.

Matriz de precisão e cobertura

	Positivo de fato	Negativo de fato	Precisão
Classificados pelo modelo como positivo	Verdadeiro Positivo (VP)	Falso Positivo (FP)	$VP / (VP + FP)$
Classificados pelo modelo como negativo	Falso Negativo (FN)	Verdadeiro Negativo (VN)	$VN / (VN + FN)$
Cobertura	$VP / (VP + FN)$	$VN / (FP + VN)$	Acurácia: $(VP + VN) / (FP + FN)$

Cross-validation



Material de **consulta**

- Tom Mitchell. Machine Learning, 1997. (Capítulo 5).
- Iah H. Witteh and Eibe Frank. Data Mining, 2000. (Capítulo 5).
- *Prediction study design. Data Analysis Course.*
Coursera.org
- Imagens retiradas de
<http://genome.tugraz.at/proclassify/help/pages/XV.html>