Tipos de Erros

Douglas Rodrigues

Universidade Federal Fluminense

Errors: In Sample x Out of Sample

Quando realizamos uma predição utilizando aprendizado de máquina, nos deparamos com dois tipos de erros de predição:

- **Erro Amostral (In Sample Error):** é a taxa de erro que aparece nos dados que utilizamos para construir o algoritmo preditor. Também chamado de *resubstitution error*.
- **Erro fora da Amostra (Out of Sample Error):** é a taxa de erro quando utilizamos novos bancos de dados, distintos dos que utilizamos para construir o preditor. Também conhecido como *generalization error*.

Errors: In Sample x Out of Sample

- No exemplo da aula anterior, o *In Sample Error* foi 24,86%.

	Não SPAM	SPAM
Não SPAM	0,4590306	0,101717
SPAM	0,1469246	0,292328

Taxa de acerto $\approx 0.4590306 + 0.2923278 \approx 75{,}14\%$.

- Em geral, se constrói o preditor tentando **minimizar** o erro amostral (*In Sample Error*), mas é preciso tomar muito cuidado com problemas de sobreajuste (*overfitting*).
- Isso ocorre quando realizamos excesso de "preciosismo" para minimizar o *In Sample Error*.

Errors: In Sample x Out of Sample

Algumas ideias-chave

1. Quase sempre o erro fora da amostra é o que interessa.
2. Geralmente, o erro amostral é menor que o erro fora da amostra.
3. Um erro frequente é ajustar muito o algoritmo ao dados que temos. Em outras palavras, criar um modelo sobreajustado.
4. Em muitos casos, é preferível aumentar o vício da função preditora, para reduzir a variância.

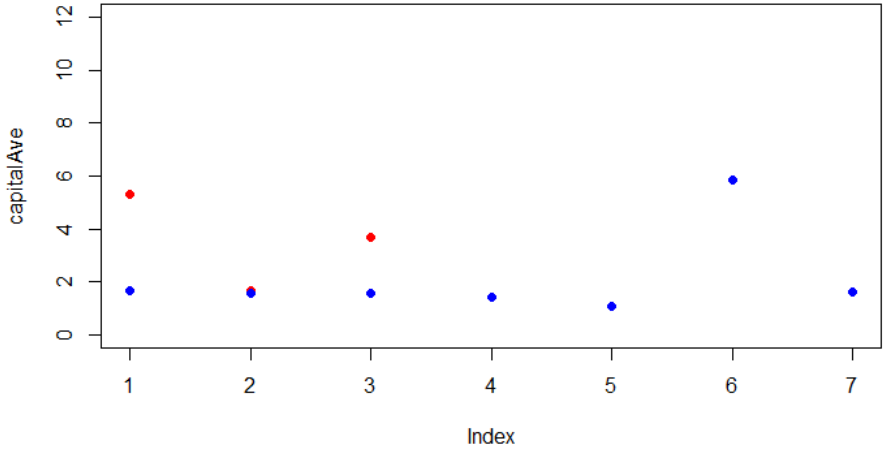
Errors: In Sample x Out of Sample

- Vamos sortear 10 observações do banco de dados `spam`, e tentar construir um preditor que minimize o *In Sample Error*.

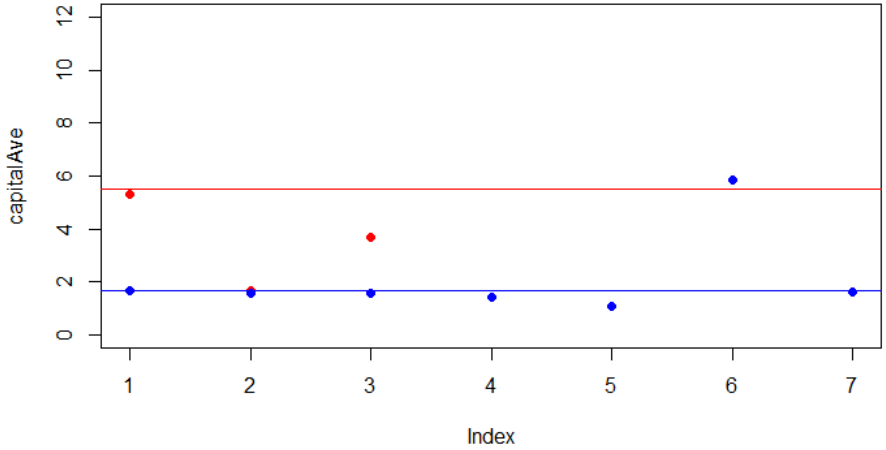
```
> set.seed(100)
> sorteio <- sample(dim(spam)[1],size=10)
> small.spam<-spam[sorteio,]
```

- Vamos plotar os dados, e tentar criar um critério que minimize o *In Sample Error*.

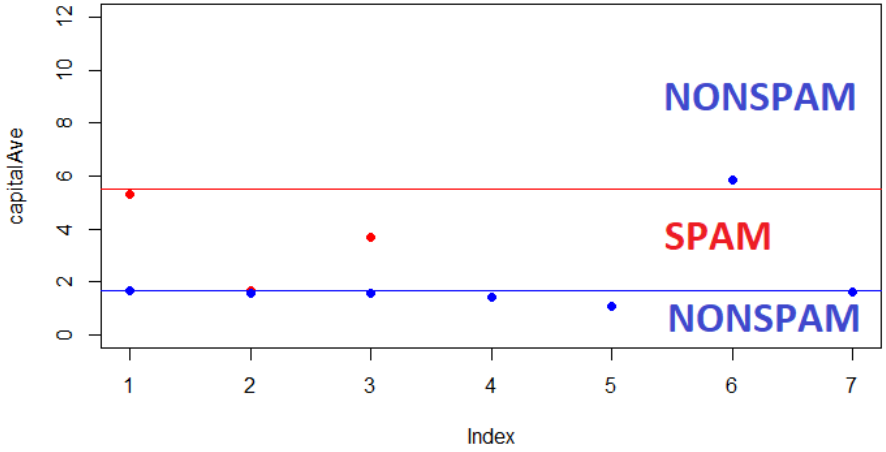
Plot dos Dados



Plot dos Dados



Plot dos Dados - Regra 2

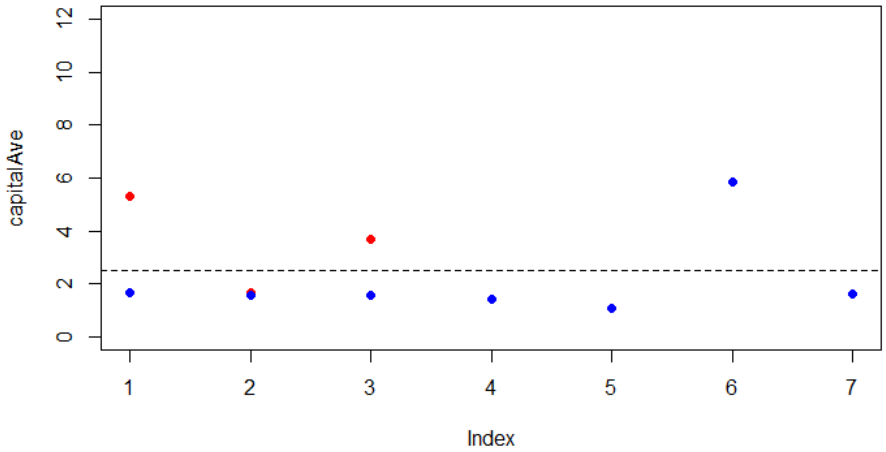


Regra 1

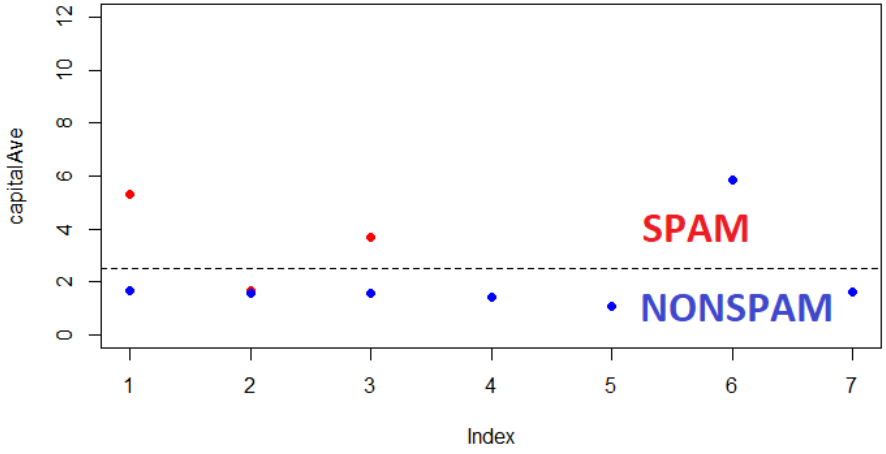
Para minimizar o *in Sample Error*, vamos criar um preditor que avalia a variável `capitalAve`, e classifica os emails da seguinte forma:

- capitalAve $\leq 1.66 \Rightarrow$ "nonspam"
- $1.66 <$ capitalAve $\leq 5.5 \Rightarrow$ "spam"
- capitalAve $> 5.5 \Rightarrow$ "nonspam"

Plot dos Dados - Regra 2



Plot dos Dados - Regra 2



Regra 2

Para minimizar o *in Sample Error*, porém evitando o *overfitting*, vamos criar um preditor que avalia a variável `capitalAve`, e classifica os emails da seguinte forma:

- capitalAve $\leq 2.5 \Rightarrow$ "nonspam"
- capitalAve $> 2.5 \Rightarrow$ "spam"