

## Exercício Computacional

---

Profs. Cristiano Leite de Castro e André Paim Lemos

24 de março de 2017

### 1 TAREFAS PRÉ-LIMINARES

1. Ler o Capítulo 2 do Livro *An Introduction to Statistical Learning* dos autores Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani.

### 2 DECOMPOSIÇÃO VIÉS E VARIÂNCIA DO ERRO DE TESTE

Seja uma problema de regressão de uma única variável tal que os valores observados da variável de saída ( $y_i$ ) são gerados de acordo com a seguinte expressão

$$Y = f(X) + \epsilon \quad (2.1)$$

onde  $f(X)$  representa a função de regressão (desconhecida) entre a saída  $Y$  e a entrada  $X$ ; e  $\epsilon$  é uma v.a.  $\sim \mathcal{N}(\mu, \sigma^2)$ .

Assumindo que a função de regressão  $f(X)$  é conhecida, o objetivo deste exercício é estimar  $MSE_{Test}$  para 5 métodos de aprendizagem com diferentes flexibilidades (graus de liberdade). Sugestão de métodos a serem utilizados:

- **Método 1:**  $\hat{f}(x_i) = c \forall x_i$ , onde  $c$  é a média amostral dos valores observados de  $y_i$ .
- **Método 2:**  $\hat{f}(x_i)$ : regressão linear de  $Y$  em  $X$ . (modelo linear)

- **Método 3:**  $\hat{f}(x_i)$ : método de flexibilidade intermediária, sendo maior que o método 2 e menor que o método 4.
- **Método 4:**  $\hat{f}(x_i)$ : método de elevada flexibilidade, sendo maior que o método 3 e menor que o método 5.
- **Método 5:**  $\hat{f}(x_i) = y_i \forall x_i$ . Este método interpola os dados de treinamento.

**Observação:** a escolha dos métodos 2, 3 e 4 fica a gosto do cliente.

A estimativa do  $MSE_{Test}$  passa pelo cálculo dos termos  $Var(\hat{f}(x_0))$ ,  $[Bias(\hat{f}(x_0))]^2$  e  $Var(\epsilon)$  para uma dada observação do conjunto de teste ( $x_0$ ), conforme Equação a seguir

$$E \left[ \left( y_0 - \hat{f}(x_0) \right)^2 \right] = \underbrace{E \left[ \left( f(x_0) - E \left[ \hat{f}(x_0) \right] \right)^2 \right]}_{Bias^2(f(\hat{x}_0))} + \underbrace{E \left[ \left( E \left[ \hat{f}(x_0) \right] - \hat{f}(x_0) \right)^2 \right]}_{Var(f(\hat{x}_0))} + Var(\epsilon) \quad (2.2)$$

Finalmente, pode se calcular o  $MSE_{Test}$  (global) tomando-se a média sobre todas as observações ( $x_0$ ) pertencentes ao conjunto de teste.

$$MSE_{test}(\hat{f}) = Ave \left( E \left[ \left( y_0 - \hat{f}(x_0) \right)^2 \right] \right) \forall (x_0, y_0).$$

Passos para obtenção de  $MSE_{test}$ :

1. gerar 20 realizações da expressão (2.1), correspondendo a 20 diferentes conjuntos de treinamento  $Tr_k = \{x_i, y_i\}_{i=1}^n$  com  $n = 40$ . Para cada realização de  $Tr_k$  os valores de  $x_i$  devem ser amostrados de forma aleatória segundo uma distribuição uniforme no intervalo equivalente ao domínio da variável de entrada  $X$ .
2. treinar os métodos 1,2,3,4 e 5 com os 20 conjuntos de treinamento gerados. Para cada método devem ser gerados 20 diferentes modelos.
3. após o treinamento, todos os modelos produzidos (de cada método) devem ser avaliados sobre um conjunto de teste  $Te$  que corresponde a uma grande quantidade de valores de  $X$  amostrados no intervalo  $[X_{min} \leq X \leq X_{max}]$ . Por exemplo, supondo que  $[-8 \leq X \leq 12]$ , então um conjunto de teste amostrado neste intervalo, com passo de 0.1, seria igual a  $Te = \{-8, -7.9, -7.8, \dots, 11.9, 12\}$ .
4. A partir da função de regressão  $f(X)$  e das saídas dos modelos  $\hat{f}(X)$  para as observações do conjunto  $Te$ , obtenha os valores de  $Var(\hat{f})$ ,  $[Bias(\hat{f})]^2$  e  $MSE_{test}(\hat{f})$  para cada método de aprendizagem.

5. gere um gráfico mostrando os valores de  $Var(\epsilon)$ ,  $Var(\hat{f})$ ,  $[Bias(\hat{f})]^2$  e  $MSE_{test}(\hat{f})$  (eixo Y) em função dos métodos 1,2,3,4 e 5 (eixo X), tal que o eixo X cresce com a flexibilidade do método.

O procedimento descrito anteriormente deve ser testado com as seguintes relações  $X, Y$ :

- $Y = \frac{(X-2)(2X+1)}{1+X^2} + \epsilon \sim N(0, \sigma^2)$  com  $[-8 \leq X \leq 12]$ .
- $Y = \sin(X) + \epsilon \sim N(0, \sigma^2)$  com  $[0 \leq X \leq 2\pi]$ .

Deve ser entregue um relatório (pdf) no Moodle.