PSI3471 – Fundamentos de Sistemas Eletrônicos Inteligentes

Regressão Linear

Magno T. M. Silva e Renato Candido

Escola Politécnica da USP

1 Regressão Linear Univariada

Seja

$$\{(x_1,d_1),(x_2,d_2),\cdots,(x_N,d_N)\},$$

um conjunto de N pontos conhecidos previamente. Vamos obter a *melhor* reta que se ajusta a esses pontos segundo o critério dos mínimos quadrados. Assim, deseja-se obter a relação

$$d = b + wx$$

em que

- d representa o sinal desejado ou rótulo
- x representa a entrada
- w representa o peso (a determinar) e
- b representa o viés ou *bias* (a determinar)

2 Regressão Linear Univariada

Quando os pontos são colineares, a reta passa exatamente por todos os N pontos e as constantes desconhecidas w e b satisfazem

$$\begin{aligned} d_1 &= b + w\, x_1 \\ d_2 &= b + w\, x_2 \\ &\vdots \\ d_N &= b + w\, x_N, \end{aligned}$$

ou seja

$$\underbrace{\begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix}}_{\mathbf{d}} = \underbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} b \\ w \end{bmatrix}}_{\mathbf{w}}$$

e

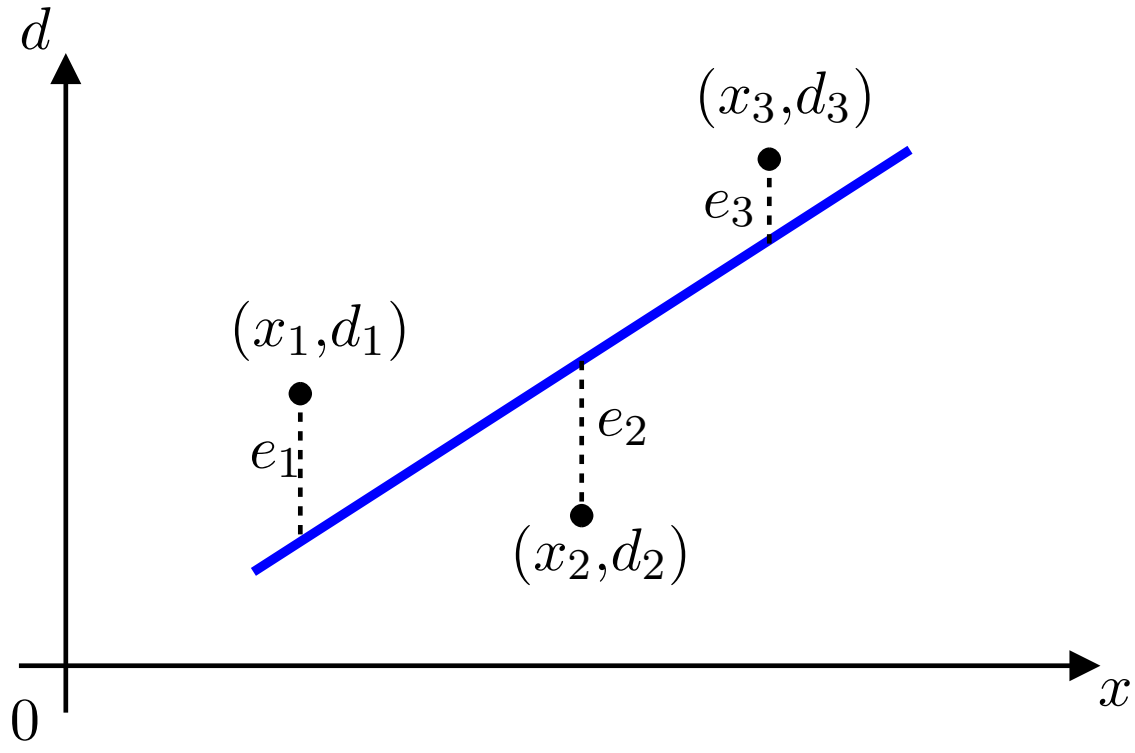
$$\mathbf{d} - \mathbf{X}\mathbf{w} = \mathbf{0}$$

3 Regressão Linear Univariada

Quando os pontos são não colineares, $\mathbf{d} - \mathbf{X}\mathbf{w} \neq \mathbf{0}$ e então define-se o vetor de erros

$$\mathbf{e} = \mathbf{d} - \mathbf{X}\mathbf{w},$$

cujos elementos $d_i - b - wx_i$ para $i = 1, \cdots, N$ representam as distâncias verticais da reta $d = wx + b$ aos pontos (x_i, d_i).



4 Regressão Linear Univariada

Segundo o critério dos mínimos quadrados, a melhor reta deve minimizar

$$\|\mathbf{e}\|^2 = \sum_{i=1}^{N} e_i^2 = \|\mathbf{d} - \mathbf{X}\mathbf{w}\|^2 = \sum_{i=1}^{N} (d_i - b - wx_i)^2.$$

Derivando em relação a w e b obtemos

$$\begin{aligned}
\frac{\partial \sum_{i=1}^{N} e_i^2}{\partial w} &= 2\sum_{i=1}^{N} e_i \frac{\partial e_i}{\partial w} &= -2\sum_{i=1}^{N} e_i x_i \\
\frac{\partial \sum_{i=1}^{N} e_i^2}{\partial b} &= 2\sum_{i=1}^{N} e_i \frac{\partial e_i}{\partial b} &= -2\sum_{i=1}^{N} e_i,
\end{aligned}$$

5 Regressão Linear Univariada

As derivadas podem ser escritas de forma compacta como

$$\frac{\partial \sum_{i=1}^{N} e_i^2}{\partial \mathbf{w}} = -2 \sum_{i=1}^{N} \begin{bmatrix} 1 \\ x_i \end{bmatrix} e_i = -2\mathbf{X}^T\mathbf{e} = -2\mathbf{X}^T(\mathbf{d} - \mathbf{X}\mathbf{w}),$$

em que $(\cdot)^T$ representa a operação de transposição.
Igualando ao vetor nulo, obtemos

$$\mathbf{X}^T\mathbf{X}\mathbf{w}^{\mathrm{o}} = \mathbf{X}^T\mathbf{d},$$

sendo $\mathbf{w}^{\mathrm{o}} = [\, b^{\mathrm{o}} \;\; w^{\mathrm{o}} \,]^T$ o vetor que minimiza $\|\mathbf{e}\|^2$ e

$$y = w^{\mathrm{o}}x + b^{\mathrm{o}} \approx d$$

é a melhor reta que se ajusta aos pontos segundo o critério dos mínimos quadrados.

6 Regressão Linear Univariada

Se $\mathbf{X}^T\mathbf{X}$ for invertível,

$$\boxed{\mathbf{w}^{\mathrm{o}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{d}.}$$

- Existe uma única reta que se ajusta aos pontos segundo o critério dos mínimos quadrados

7 Regressão Linear Univariada

Observações:

- O modelo $y = w^{\text{o}}x + b^{\text{o}}$ é de fato linear apenas quando $b^{\text{o}} = 0$, pois neste caso $x = 0$ leva a $y = 0$. No entanto, o termo *linear* é frequentemente usado neste caso para se referir ao modelo dado por uma reta.
- Os dados

$$\{(x_1,d_1),(x_2,d_2),\cdots,(x_N,d_N)\},$$

  foram totalmente usados aqui para se obter o modelo da reta. Neste caso, eles são chamados de dados de treinamento do modelo.
- A matriz $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ é conhecida como a pseudoinversa de $\mathbf{X}$.

8 Regressão Linear Multivariada

Seja

$$\{(x_{11}, x_{21}, \cdots, x_{M1}, d_1), (x_{12}, x_{22}, \cdots, x_{M2}, d_2), \cdots,$$

$$(x_{1N}, x_{2N}, \cdots, x_{MN}, d_N)\}$$

um conjunto de dados composto por M valores de x, seguido do valor de d. Deseja-se agora obter a melhor função linear segundo o critério dos mínimos quadrados que se ajusta a esses dados, i.e.,

$$y = b + w_1 x_1 + w_2 x_2 + \cdots + w_M x_M \approx d.$$

9 Regressão Linear Multivariada

Considerando os N conjuntos de dados, obtém-se

$$\underbrace{\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}}_{\mathbf{e}} = \underbrace{\begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix}}_{\mathbf{d}} - \underbrace{\begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{M1} \\ 1 & x_{12} & x_{22} & \cdots & x_{M2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1N} & x_{2N} & \cdots & x_{MN} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} b \\ w_1 \\ \vdots \\ w_M \end{bmatrix}}_{\mathbf{w}}$$

Como no caso da reta, o *melhor* hiperplano deve minimizar

$$\|\mathbf{e}\|^2 = \|\mathbf{d} - \mathbf{X}\mathbf{w}\|^2.$$

Generalizando, chega-se a

$$\boxed{\mathbf{w}^{\mathrm{o}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{d}}$$

em que $\mathbf{w}^{\mathrm{o}} = [\, b^{\mathrm{o}} \; w_1^{\mathrm{o}} \; w_2^{\mathrm{o}} \; \cdots \; w_M^{\mathrm{o}} \,]^T$ é o vetor que minimiza $\|\mathbf{e}\|^2$.

10 Regressão Linear Multivariada

- Calcular a inversa de $\mathbf{X}^T\mathbf{X}$ diretamente pode causar problemas numéricos
- A matriz $\mathbf{X}^T\mathbf{X}$ é uma estimativa da matriz de autocorrelação dos dados de entrada x
- O vetor $\mathbf{X}^T\mathbf{d}$ é uma estimativa da correlação cruzada entre os dados de entrada x e o sinal desejado d

11 Regressão Linear Multivariada

- Quando se deseja ajustar um polinômio de grau M aos dados

$$\{(x_1,d_1),(x_2,d_2),\cdots,(x_N,d_N)\},$$

basta usar o resultado anterior, considerando

$$\{(x_1,x_1^2,\cdots,x_1^M,d_1),(x_2,x_2^2,\cdots,x_2^M,d_2),\cdots,$$

$$(x_N,x_N^2,\cdots,x_N^M,d_N)\},$$

o que leva a

$$y = b + w_1 x + w_2 x^2 + \cdots + w_M x^M \approx d.$$

Neste caso a matrix $\mathbf{X}$ se torna

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^M \\ 1 & x_2 & x_2^2 & \cdots & x_2^M \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^M \end{bmatrix}.$$

12 Regressão Linear Multivariada

- Podemos aproximar os dados não só por polinômios, mas também por outras funções.
  Por exemplo, poderíamos calcular as seguintes aproximações para os dados

$$y = b + w_1 \ln(x + 5) + w_2 \exp(x - 2) \approx d$$

ou

$$y = b + w_1 \cos(2\pi f_0 x) + w_2 \text{sen}(2\pi f_0 x) \approx d$$

em que f_0 é uma frequência pré-determinada.

13 Regressão Linear Multivariada

- Um outro exemplo útil em Engenharia Elétrica é aproximar uma função $f(t)$ periódica com período $T_0 = 1/f_0$ por uma soma de senos e cossenos, ou seja,

$$\begin{aligned} f(t) \approx\ & b + w_{11}\cos(2\pi f_0 t) + w_{12}\text{sen}\,(2\pi f_0 t) \\ & + w_{21}\cos(2\pi 2 f_0 t) + w_{22}\text{sen}\,(2\pi 2 f_0 t) \\ & + \cdots \\ & + w_{M1}\cos(2\pi f_0 M t) + w_{M2}\text{sen}\,(2\pi f_0 M t). \end{aligned}$$

Os coeficientes $b, w_{11}, w_{12}, w_{21}, w_{22}, \cdots, w_{M1}, w_{M2}$ são conhecidos como coeficientes da série de Fourier e os dados usados para obter essa aproximação são obtidos a partir da amostragem da função $f(t)$.

14 Sobreajuste (*Overfitting*)

Considere o seguinte problema de regressão multivariada:

> Deseja-se prever o valor de venda de um automóvel usado, utilizando diferentes informações como (1) ano de fabricação, (2) modelo, (3) estado de conservação, (4) valor da tabela Fipe, (5) valor médio de venda no mercado, etc.

Poderíamos usar todos os dados disponíveis para gerar o modelo.

- Se fizéssemos isso, como conseguiríamos avaliar se o modelo obtido é bom?
- Como saber se o modelo é capaz de prever adequadamente o valor de venda de um determinado carro que não aparece no banco de dados?

15 *Overfitting*

É importante reservar uma parte dos dados para avaliar a qualidade do modelo. É comum separar os dados de forma aleatória em dois conjuntos disjuntos:

- **treinamento** (ou aprendizado) usado para gerar o modelo e
- **teste** usado para avaliar a qualidade do modelo gerado.

Os dados usados para avaliação não devem aparecer no treinamento e vice-versa.
Se o modelo se sair bem no teste, costuma-se dizer que ele tem uma boa capacidade de **generalização**.

16 *Overfitting*

- Um modelo com muitos parâmetros pode ter um ótimo desempenho do treinamento, mas uma baixa capacidade de generalização, o que leva a um erro elevado na fase de teste gerando **overfitting**.
- Modelos com baixa capacidade de generalização não são desejáveis, uma vez que na prática serão apresentados a dados que não foram usados no treinamento e deveriam ser capazes de realizar uma predição ou classificação de maneira adequada.

17 Exemplo de *Overfitting*

Considere 10 valores igualmente espaçados de x em $[0{,}1,\ 1{,}5]$.
Os valores de d são gerados por

$$d = 0{,}5 + 0{,}25\cos(2\pi x) + v,$$

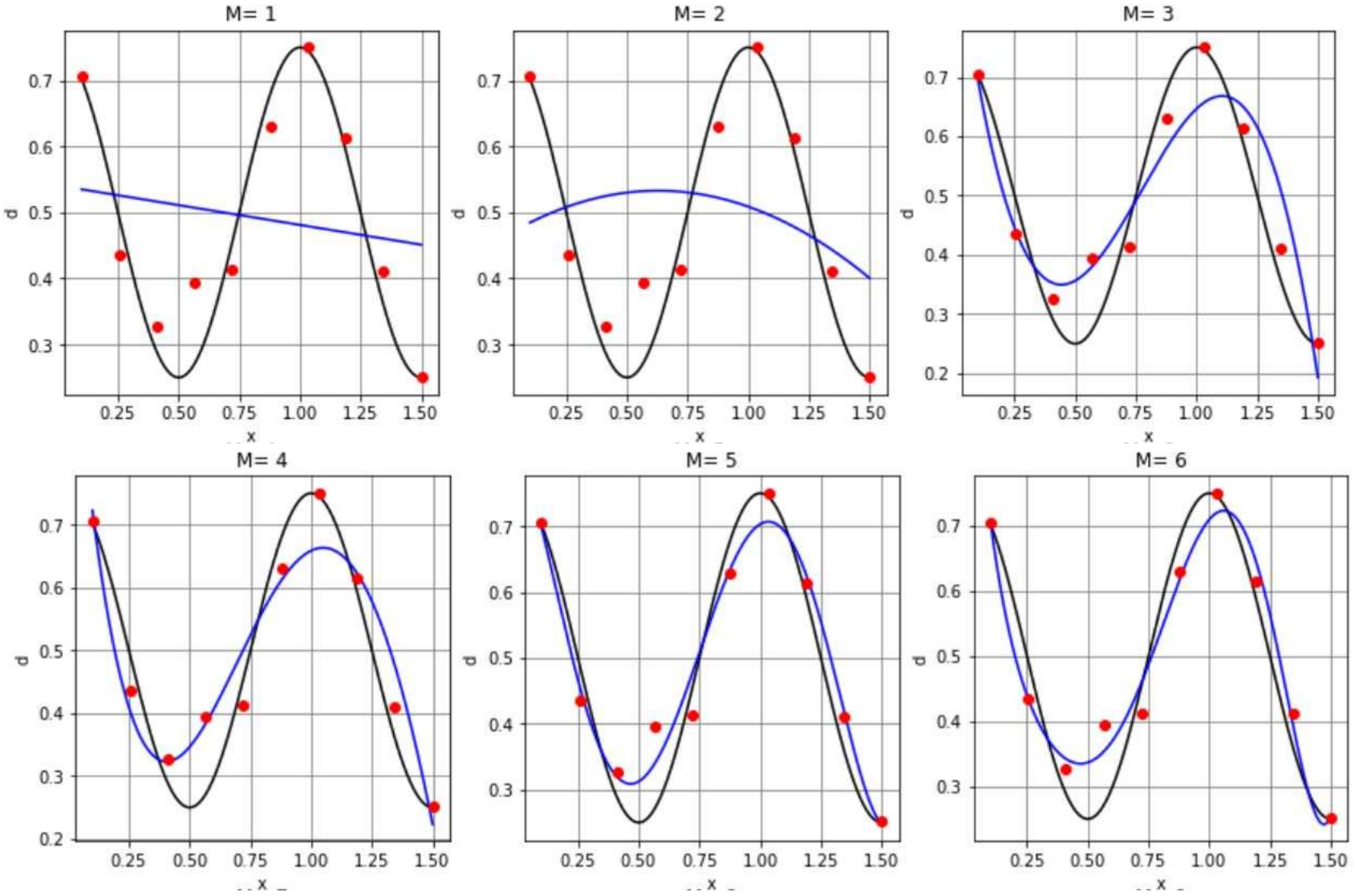
em que v é um ruído branco gaussiano com média zero e desvio padrão 0,06.
Por exemplo, poderíamos ter o seguinte conjunto de treinamento

$$\{(0{,}1000,\ 0{,}7055),\ (0{,}2556,\ 0{,}4357),\ (0{,}4111,\ 0{,}3264),\ \cdots,$$

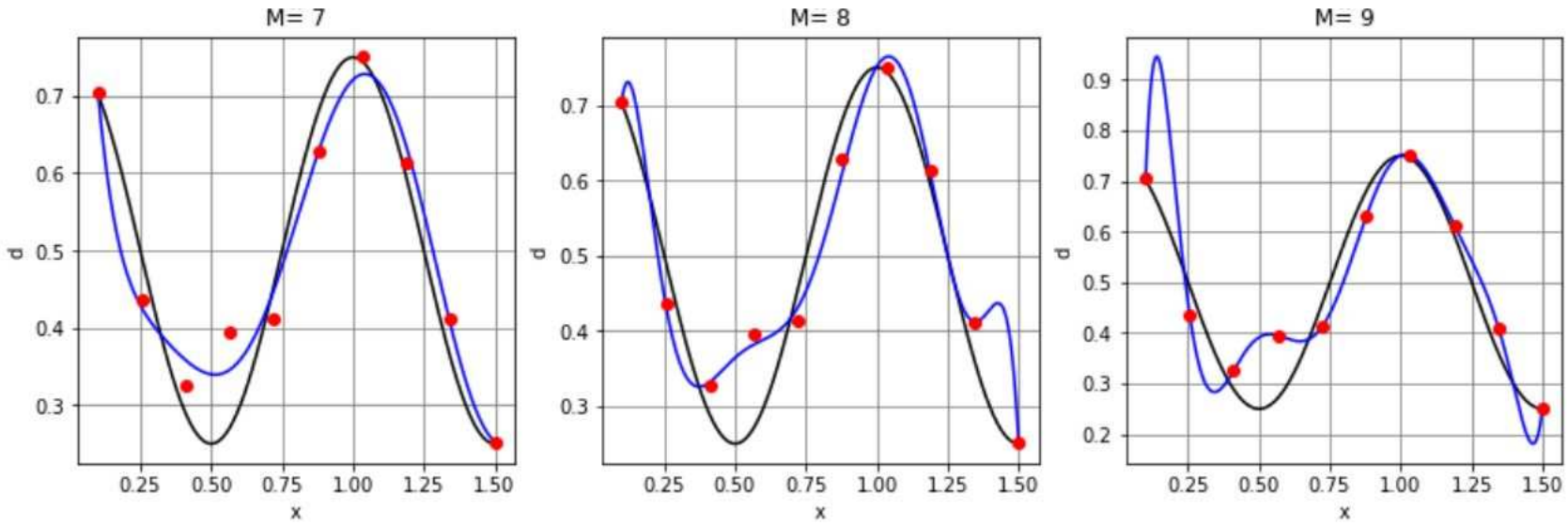
$$(1{,}5000,\ 0{,}2514)\}.$$

Objetivo: encontrar uma função (um modelo) polinomial de grau M que melhor se aproxima dos pontos do conjunto de treinamento, levando em conta a forma da cossenóide sem ruído.

18 Exemplo de *Overfitting*



19 Exemplo de *Overfitting*



- Para $M = 1$ e $M = 2$ ocorre *underfitting*
- À medida que o valor do grau do polinômio aumenta, observa-se um melhor ajuste entre os pontos vermelhos e as curvas azuis
- Para $M = 9$, o polinômio obtido passa exatamente em todos os pontos do treinamento, mas claramente a curva azul fica distante da cossenóide sem ruído em alguns trechos. Pode ter ocorrido *overfitting* devido ao número excessivo de parâmetros do modelo.

20 Exemplo de *Overfitting*

- Geramos um conjunto de teste com 1401 valores de x igualmente espaçados no intervalo $[0{,}1,\ 1{,}5]$ e calculamos o valor de d correspondente
- Para cada valor de M, medimos o valor absoluto médio do erro de predição, levando em conta o conjunto de treinamento e de teste