Aprendizagem 2024/25

Homework III – Group 060

(ist1106970 – Francisco Heleno, ist1106494 – Mafalda Dias)

I. Pen-and-paper

1) Utilizamos a função/parametrização $\Phi(y_1, y_2)$ que transforma o espaço original num novo espaço de uma dimensão:

D	y_1	y_2	$\Phi(y_1, y_2)$
x_1	1	1	1
x_2	1	3	3
x_3	3	2	6
x_4	3	3	9
x_5	2	4	8

De seguida, juntamos uma coluna de 1's (bias) à coluna com os valores de $\Phi(y_1, y_2)$ e obtemos a matriz X. Definimos também o vetor coluna $y = y_{num}$:

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 6 \\ 1 & 9 \\ 1 & 8 \end{pmatrix} \qquad y = \begin{pmatrix} 1.25 \\ 7.0 \\ 2.7 \\ 3.2 \\ 5.5 \end{pmatrix}$$

Calculamos então os coeficientes do nosso modelo utilizando a solução fechada de OLS:

$$W = (X^T X)^{-1} X^T y$$

$$W = \left(\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 6 & 9 & 8 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 6 \\ 1 & 9 \\ 1 & 8 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 6 & 9 & 8 \end{pmatrix} \begin{pmatrix} 1.25 \\ 7.0 \\ 2.7 \\ 3.2 \\ 5.5 \end{pmatrix} = \begin{pmatrix} 3.31593 \\ 0.11372 \end{pmatrix}$$

$$W = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} 3.31593 \\ 0.11372 \end{pmatrix}$$

Assim, o modelo de regressão linear para prever y_{num} é:

$$\begin{aligned} \hat{y}_{\text{num}} &= 3.31593 + 0.11372 \cdot \phi(y_1, y_2) \\ &= 3.31593 + 0.11372 \cdot (y_1 \cdot y_2) \end{aligned}$$

Aprendizagem 2024/25

Homework III – Group 060

(ist1106970 – Francisco Heleno, ist1106494 – Mafalda Dias)

2) Os coeficientes da solução fechada de Ridge Regression com $\lambda = 1$ são calculados da seguinte forma:

$$W = \left(X^T X + \lambda I\right)^{-1} X^T y \quad \text{com} \quad \lambda = 1$$

$$W = \left(\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 6 & 9 & 8 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 6 \\ 1 & 9 \\ 1 & 8 \end{pmatrix} + 1 \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 6 & 9 & 8 \end{pmatrix} \begin{pmatrix} 1.25 \\ 7.0 \\ 2.7 \\ 3.2 \\ 5.5 \end{pmatrix} = \begin{pmatrix} 1.81809 \\ 0.32376 \end{pmatrix}$$

$$W = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} 1.81809 \\ 0.32376 \end{pmatrix}$$

Desta maneira, o nosso modelo de Ridge Regression para prever y_{num} é:

$$\begin{aligned} \hat{y}_{\text{num}} &= 1.81809 + 0.32376 \cdot \phi(y_1, y_2) \\ &= 1.81809 + 0.32376 \cdot (y_1 \cdot y_2) \end{aligned}$$

É sabido que a regularização imposta pela Ridge Regression (com λ=1) tende a "encolher" os coeficientes em relação à regressão linear simples, mas analisando os valores obtidos neste caso específico, notamos uma redução do valor de w_0 (de 3.31593 para 1.81809) e um aumento do valor de w_1 (de 0.11372 para 0.32376).

Este aumento de w_1 pode parecer contraditório mas, na verdade, reflete um equilíbrio entre o ajuste aos dados e a penalização por complexidade, ou seja, este tipo de regularização suaviza os parâmetros, de modo a evitar ajustes excessivos e melhorar a generalização do modelo, especialmente em casos onde há risco de *overfitting*.

3) Começamos por calcular $\Phi(y_1, y_2)$ para as novas observações:

D	y_1	y_2	$\Phi(y_1, y_2)$
x_6	2	2	4
x_7	1	2	2
x_8	5	1	5

Homework III – Group 060

(ist1106970 – Francisco Heleno, ist1106494 – Mafalda Dias)

De seguida, utilizamos os modelos calculados anteriormente para prever y_{num} para as novas observações:

$$\text{OLS:} \begin{cases} \hat{y}_6 = 3.31593 + 0.11372 \cdot 4 \approx 3.77081 \\ \hat{y}_7 = 3.31593 + 0.11372 \cdot 2 \approx 3.54337 \\ \hat{y}_8 = 3.31593 + 0.11372 \cdot 5 \approx 3.88453 \end{cases}$$

$$\text{Ridge:} \begin{cases} \hat{y}_6 = 1.81809 + 0.32376 \cdot 4 \approx 3.11313 \\ \hat{y}_7 = 1.81809 + 0.32376 \cdot 2 \approx 2.46561 \\ \hat{y}_8 = 1.81809 + 0.32376 \cdot 5 \approx 3.43689 \end{cases}$$

Calculamos então os RMSE's de teste:

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_{num_i})^2}$$

$$\text{RMSE}_{\text{OLS}} = \sqrt{\frac{1}{3}\left((3.77081 - 0.7)^2 + (3.54337 - 1.1)^2 + (3.88453 - 2.2)^2\right)} \approx 2.46560$$

$$\text{RMSE}_{\text{Ridge}} = \sqrt{\frac{1}{3}\left((3.11313 - 0.7)^2 + (2.46561 - 1.1)^2 + (3.43689 - 2.2)^2\right)} \approx 1.75290$$

Agora, voltamos a utilizar os modelos de regressão para prever y_{num}, mas desta vez para as observações de treino:

$$\text{OLS:} \begin{cases} \hat{y_1} = 3.31593 + 0.11372 \cdot 1 \approx 3.42965 \\ \hat{y_2} = 3.31593 + 0.11372 \cdot 3 \approx 3.65709 \\ \hat{y_3} = 3.31593 + 0.11372 \cdot 6 \approx 3.99825 \\ \hat{y_4} = 3.31593 + 0.11372 \cdot 9 \approx 4.33941 \\ \hat{y_5} = 3.31593 + 0.11372 \cdot 8 \approx 4.22569 \end{cases}$$

$$\text{Ridge:} \begin{cases} \hat{y_1} = 1.81809 + 0.32376 \cdot 1 \approx 2.14185 \\ \hat{y_2} = 1.81809 + 0.32376 \cdot 3 \approx 2.78937 \\ \hat{y_3} = 1.81809 + 0.32376 \cdot 6 \approx 3.76065 \\ \hat{y_4} = 1.81809 + 0.32376 \cdot 9 \approx 4.73193 \\ \hat{y_5} = 1.81809 + 0.32376 \cdot 8 \approx 4.40817 \end{cases}$$

Podemos assim calcular os RMSE's de treino:

$$\text{RMSE}_{\text{OLS}} = \sqrt{\frac{1}{5}\left((3.42965 - 1.25)^2 + (3.65709 - 7)^2 + (3.99825 - 2.7)^2 + (4.33491 - 3.2)^2 + (4.22569 - 5.5)^2\right)} \approx$$
$$\approx 2.02650$$

$$\text{RMSE}_{\text{Ridge}} = \sqrt{\frac{1}{5}\left((2.14185 - 1.25)^2 + (2.78937 - 7)^2 + (3.76065 - 2.7)^2 + (4.73193 - 3.2)^2 + (4.40817 - 5.5)^2\right)} \approx$$
$$\approx 2.15354$$

Homework III – Group 060

(ist1106970 – Francisco Heleno, ist1106494 – Mafalda Dias)

No conjunto de treino, o modelo de OLS apresentou um melhor desempenho do que o modelo de Ridge Regression, com um RMSE de 2.02650 comparado ao 2.15354 do Ridge. Esta diferença é esperada, pois a regularização do Ridge penaliza coeficientes elevados, introduzindo um bias que leva a um ajuste menos preciso aos dados de treino.

Já no conjunto de teste, o OLS teve um RMSE de 2.46559, indicando *overfitting*, ou seja, ajustou-se demasiado aos dados de treino e não generalizou bem para novos dados. Por outro lado, o modelo de Ridge obteve um RMSE de 1.75289, significativamente mais baixo, o que demonstra uma melhor generalização. A regularização permitiu ao modelo Ridge evitar o sobreajuste, resultando num desempenho mais robusto em dados não vistos.

Concluindo, os resultados estão de acordo com o esperado, pois o Ridge apresentou uma melhor generalização ao reduzir o *overfitting* em comparação ao OLS, conforme evidenciado pelo RMSE mais baixo no conjunto de teste.

4) Através dos dados fornecidos para os pesos e biases, conseguimos perceber a estrutura da rede neuronal em questão: camada de entrada com 2 neurónios, uma camada oculta com 3 neurónios e camada de saída com 3 neurónios.

Começamos então a fazer a Propagação (Propagation):

$$\mathbf{z}[i] = W[i] \cdot \mathbf{x}[i-1] + b[i]$$

Note-se que consideramos $\mathrm{x}[0] = \mathrm{x}_1$.

$$\mathbf{z}[1] = W[1] \cdot \mathbf{x}[0] + b[1] = \begin{pmatrix} 0.1 & 0.1 \\ 0.1 & 0.2 \\ 0.2 & 0.1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 0.1 \\ 0 \\ 0.1 \end{pmatrix} = \begin{pmatrix} 0.3 \\ 0.3 \\ 0.4 \end{pmatrix}$$

Como não há ativações nas camadas ocultas:

$$\mathbf{x}[1] = \mathbf{z}[1] = \begin{pmatrix} 0.3 \\ 0.3 \\ 0.4 \end{pmatrix}$$

Usando *softmax* como função de ativação na camada de saída:

$$\mathbf{z}[2] = W[2] \cdot \mathbf{x}[1] + b[2] = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0.3 \\ 0.3 \\ 0.4 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2.7 \\ 2.3 \\ 2 \end{pmatrix}$$

$$\mathbf{x}[2] = \mathrm{softmax}(\mathbf{z}[2]) = \frac{1}{e^{2.7} + e^{2.3} + e^{2}} \begin{pmatrix} e^{2.7} \\ e^{2.3} \\ e^{2} \end{pmatrix} \approx \begin{pmatrix} 0.46149 \\ 0.30934 \\ 0.22917 \end{pmatrix}$$

Homework III – Group 060

(ist1106970 – Francisco Heleno, ist1106494 – Mafalda Dias)

Passamos para a Retropropagação (Back-propagation):

$$\frac{\partial E}{\partial W[2]} = \frac{\partial E}{\partial \mathbf{x}[2]} \circ \frac{\partial \mathbf{x}[2]}{\partial \mathbf{z}[2]} \cdot \left(\frac{\partial \mathbf{z}[2]}{\partial W[2]}\right)^T = \frac{\partial E}{\partial \mathbf{z}[2]} \cdot \left(\frac{\partial \mathbf{z}[2]}{\partial W[2]}\right)^T = \delta[2] \cdot \left(\frac{\partial \mathbf{z}[2]}{\partial W[2]}\right)^T$$

$$\mathbf{t} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \text{ já que } y_{\text{class}} = B \text{ para } x_1$$

$$\delta[2] = \frac{\partial E}{\partial \mathbf{z}[2]} = \frac{\partial}{\partial \mathbf{z}[2]} \left(-\sum_{l=1}^{|C|} t_l \log\left(x_l[2]\right) \right) = \ldots = \mathbf{x}[2] - \mathbf{t}$$

Estamos em condições de calcular os deltas:

$$\delta[2] = \mathbf{x}[2] - \mathbf{t} = \begin{pmatrix} 0.46149 \\ 0.30934 \\ 0.22917 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.46149 \\ -0.69066 \\ 0.22917 \end{pmatrix}$$

$$\delta[1] = (W[2])^T \cdot \delta[2] = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}^T \cdot \begin{pmatrix} 0.46149 \\ -0.69066 \\ 0.22917 \end{pmatrix} = \begin{pmatrix} 0 \\ -0.22917 \\ 0.46149 \end{pmatrix}$$

Podemos agora proceder à atualização de todos os pesos e biases pela descida do gradiente estocástico (stochastic gradient descent update):

$$W[2] = W[2] - \eta \cdot \frac{\partial E}{\partial W[2]} = W[2] - \eta \cdot \left(\delta[2] \cdot (x[1])^T\right) =$$

$$= \begin{pmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix} - 0.1 \cdot \left(\begin{pmatrix} 0.46149 \\ 0.69066 \\ 0.22977 \end{pmatrix} \cdot \begin{pmatrix} 0.3 & 0.3 & 0.4 \end{pmatrix} \right) \approx$$

$$\approx \begin{pmatrix} 0.98616 & 1.98616 & 1.98154 \\ 1.02072 & 2.02072 & 1.02763 \\ 0.99312 & 0.99312 & 0.99083 \end{pmatrix}$$

$$b[2] = b[2] - \eta \cdot \frac{\partial E}{\partial b[2]} = b[2] - \eta \cdot \delta[2] =$$

$$= \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - 0.1 \cdot \begin{pmatrix} 0.46149 \\ -0.69066 \\ 0.22917 \end{pmatrix} \approx$$

$$\approx \begin{pmatrix} 0.95385 \\ 1.06907 \\ 0.97708 \end{pmatrix}$$

Homework III – Group 060

(ist1106970 – Francisco Heleno, ist1106494 – Mafalda Dias)

$$W[1] = W[1] - \eta \cdot \frac{\partial E}{\partial W[1]} = W[1] - \eta \cdot \left(\delta[1] \cdot (x[0])^T\right) =$$

$$= \begin{pmatrix} 0.1 & 0.1 \\ 0.1 & 0.2 \\ 0.2 & 0.1 \end{pmatrix} - 0.1 \cdot \left(\begin{pmatrix} 0 \\ -0.22917 \\ 0.46149 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \end{pmatrix} \right) \approx$$

$$\approx \begin{pmatrix} 0.1 & 0.1 \\ 0.12292 & 0.22292 \\ 0.15385 & 0.05385 \end{pmatrix}$$

$$b[1] = b[1] - \eta \cdot \frac{\partial E}{\partial b[1]} = b[1] - \eta \cdot \delta[1] =$$

$$= \begin{pmatrix} 0.1 \\ 0 \\ 0.1 \end{pmatrix} - 0.1 \cdot \left(\begin{pmatrix} 0 \\ -0.22917 \\ 0.46149 \end{pmatrix} \right) \approx$$

$$\approx \begin{pmatrix} 0.1 \\ 0.02292 \\ 0.05385 \end{pmatrix}$$

Resumindo, os novos pesos e biases após a atualização são:

$$W[1] = \begin{pmatrix} 0.1 & 0.1 \\ 0.12292 & 0.22292 \\ 0.15385 & 0.05385 \end{pmatrix} \qquad b[1] = \begin{pmatrix} 0.1 \\ 0.02292 \\ 0.05385 \end{pmatrix}$$

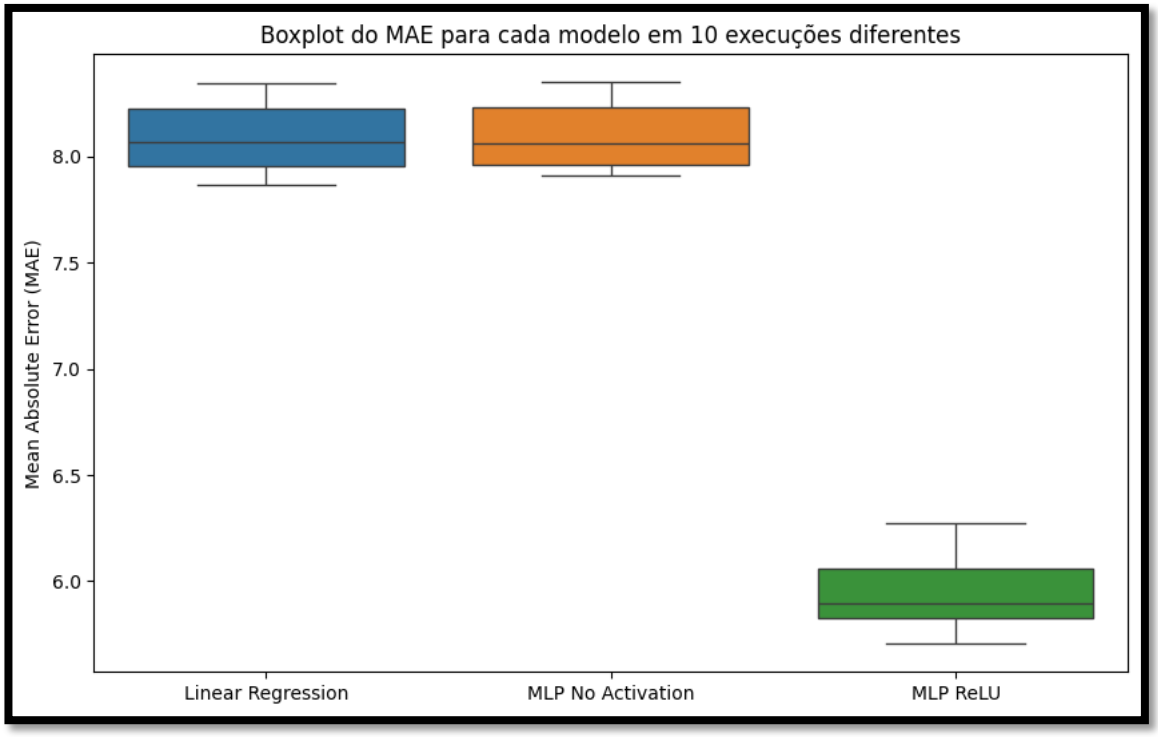
$$W[2] = \begin{pmatrix} 0.98616 & 1.98616 & 1.98154 \\ 1.02072 & 2.02072 & 1.02763 \\ 0.99312 & 0.99312 & 0.99083 \end{pmatrix} \qquad b[2] = \begin{pmatrix} 0.95385 \\ 1.06907 \\ 0.97708 \end{pmatrix}$$

Homework III – Group 060

(ist1106970 – Francisco Heleno, ist1106494 – Mafalda Dias)

II. Programming and critical analysis

5)



6)

A comparação entre a Regressão Linear e a MLP sem funções de ativação no boxplot mostra que o desempenho de ambas, medido pelo erro médio absoluto (MAE), é praticamente idêntico, rondando os 8.0. Esta semelhança deve-se ao facto de uma MLP sem funções de ativação se comportar como um modelo linear, isto porque sem a não-linearidade introduzida pelas funções de ativação, a MLP só consegue modelar relações lineares, o que a torna equivalente à Regressão Linear na sua capacidade de aprender a partir dos dados.

Por outro lado, a MLP com a função de ativação ReLU apresenta um MAE significativamente mais baixo, entre 5.0 e 6.0, o que demonstra uma clara melhoria de desempenho. Isto sublinha a importância das funções de ativação nas MLPs, pois estas introduzem não-linearidade, permitindo assim ao modelo captar padrões complexos que os modelos lineares não conseguem.

Concluímos que as funções de ativação são essenciais para desbloquear o verdadeiro potencial das MLPs, visto que lhes permitem aprender além das relações lineares simples.

Aprendizagem 2024/25

Homework III – Group 060

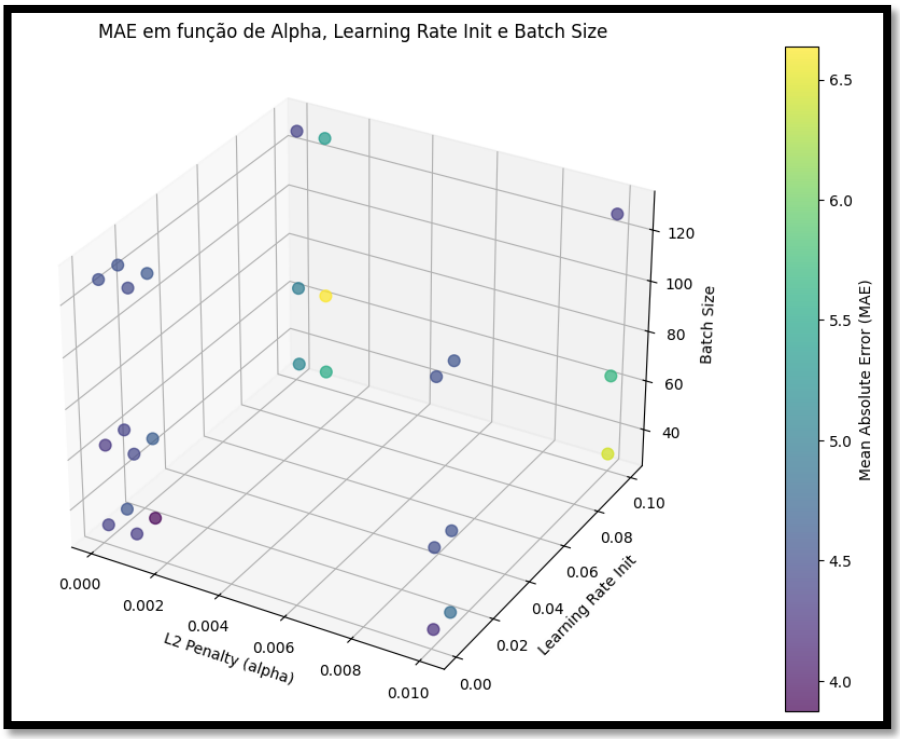
(ist1106970 – Francisco Heleno, ist1106494 – Mafalda Dias)

7)

Melhor combinação de hiperparâmetros:

- L2 Penalty (alpha) = 0.001;
- Batch size = 32;
- Learning Rate = 0.01.

Melhor MAE no conjunto de teste: 3.8731



Análise dos Hiperparâmetros:

1. **Batch Size:** Tamanhos de 32 e 64 apresentaram os melhores resultados, com MAEs entre 4.0 e 4.5 (representados pelas cores escuras, como roxo e azul escuro). O batch de 128 teve MAEs um pouco maiores, variando entre 5.0 e 6.5 (tons mais claros, como verde claro e amarelo).
2. **Learning Rate:** A taxa de aprendizagem de 0.001 resultou nos menores MAEs, com valores entre 4.0 e 4.5 (cores escuras). Taxas mais altas, como 0.01 e 0.1, levaram a um aumento do MAE, variando entre 5.0 e 6.5 (cores mais claras).
3. **L2 Penalty (Alpha):** Valores baixos de alpha, como 0.0001, produziram os melhores MAEs (pontos roxos). À medida que o alpha aumentou para 0.01, os MAEs subiram para 6.0 a 6.5 (tons mais claros de verde claro e amarelo).

Através da análise do gráfico, é possível notar que combinações com batch sizes menores, taxas de aprendizagem mais baixas e valores de alpha menores resultam num melhor desempenho (cores mais escuras). Isto sugere uma boa robustez do modelo em relação a essas variações. Também é evidente que há diversos grupos de pontos com MAEs muito semelhantes, indicando que a diferença de desempenho entre pontos desses grupos não é muito significativa.

END