Aprendizagem 2021/22

Homework I – Group 009

I. Pen-and-paper

1) Cálculos da propagação:

$$x^{[0]} = \begin{pmatrix}1\\1\\1\\1\\1\end{pmatrix}$$

$$z^{[1]} = W^{[1]}x^{[0]} + b^{[1]} = \begin{pmatrix}1 & 1 & 1 & 1 & 1\\0 & 0 & 0 & 0 & 0\\1 & 1 & 1 & 1 & 1\end{pmatrix}\begin{pmatrix}1\\1\\1\\1\\1\end{pmatrix} + \begin{pmatrix}1\\1\\1\end{pmatrix} = \begin{pmatrix}6\\1\\6\end{pmatrix} \qquad x^{[1]} = \tanh\begin{pmatrix}6\\1\\6\end{pmatrix} = \begin{pmatrix}0.99999\\0.76159\\0.99999\end{pmatrix}$$

$$z^{[2]} = W^{[2]}x^{[1]} + b^{[2]} = \begin{pmatrix}1 & 1 & 1\\1 & 1 & 1\end{pmatrix}\begin{pmatrix}0.99999\\0.76159\\0.99999\end{pmatrix} + \begin{pmatrix}1\\1\end{pmatrix} = \begin{pmatrix}3.76157\\3.76157\end{pmatrix} \qquad x^{[2]} = \tanh\begin{pmatrix}3.76157\\3.76157\end{pmatrix} = \begin{pmatrix}0.99892\\0.99892\end{pmatrix}$$

$$z^{[3]} = W^{[3]}x^{[2]} + b^{[3]} = \begin{pmatrix}0 & 0\\0 & 0\end{pmatrix}\begin{pmatrix}0.99892\\0.99892\end{pmatrix} + \begin{pmatrix}0\\0\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix} \qquad x^{[3]} = \tanh\begin{pmatrix}0\\0\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix}$$

Cálculo dos deltas:

$$tanh'(x) = \left(\frac{e^x - e^{-x}}{e^x + e^{-x}}\right)' = \frac{(e^x-e^{-x})'(e^x+e^{-x}) - (e^x-e^{-x})(e^x+e^{-x})'}{(e^x+e^{-x})^2} = \frac{(e^x+e^{-x})^2 - (e^x-e^{-x})^2}{(e^x+e^{-x})^2} = 1 - \tanh^2(x)$$

$$\frac{\partial E}{\partial x^{[3]}} = \frac{\partial}{\partial x^{[3]}}\left(\frac{1}{2}\sum_{i=1}^{2}\left(t_i - x_i^{[3]}\right)^2\right) = \frac{\partial}{\partial x^{[3]}}\left(\frac{1}{2}\sum_{i=1}^{3}\left(t_i^2 - 2x_i^{[3]}t_i + {x_i^{[3]}}^2\right)\right) = x^{[3]} - t$$

$$\delta^{[3]} = \frac{\partial E}{\partial z^{[3]}} = \frac{\partial E}{\partial x^{[3]}} \circ \frac{\partial x^{[3]}}{\partial z^{[3]}} = (x^{[3]} - t\,)\circ(1 - \tanh^2(z^{[3]})) = \left(\begin{pmatrix}0\\0\end{pmatrix} - \begin{pmatrix}1\\-1\end{pmatrix}\right)\circ\left(1 - \begin{pmatrix}0\\0\end{pmatrix}\right) = \begin{pmatrix}-1\\1\end{pmatrix}$$

$$\delta^{[2]} = \left(\frac{\partial z^{[3]}}{\partial x^{[2]}}\right)^T \cdot \delta^{[3]} \circ \frac{\partial x^{[2]}}{\partial z^{[2]}} = W^{[3]^T} \cdot \delta^{[3]} \circ (1 - \tanh^2(z^{[2]})) = \begin{pmatrix}0 & 0\\0 & 0\end{pmatrix}^T \cdot \begin{pmatrix}-1\\1\end{pmatrix} \circ \left(1 - \begin{pmatrix}0.99784\\0.99784\end{pmatrix}\right) = \begin{pmatrix}0\\0\end{pmatrix}$$

$$\delta^{[1]} = \left(\frac{\partial z^{[2]}}{\partial x^{[1]}}\right)^T \cdot \delta^{[3]} \circ \frac{\partial x^{[1]}}{\partial z^{[1]}} = W^{[2]^T} \cdot \delta^{[2]} \circ (1 - \tanh^2(z^{[1]})) = \begin{pmatrix}1 & 1 & 1\\1 & 1 & 1\end{pmatrix}^T \cdot \begin{pmatrix}0\\0\end{pmatrix} \circ \left(1 - \begin{pmatrix}0.99998\\0.58003\\0.99998\end{pmatrix}\right) = \begin{pmatrix}0\\0\\0\end{pmatrix}$$

Atualização dos pesos e dos bias:

$$\frac{\partial E}{\partial W^{[i]}} = \frac{\partial E}{\partial z^{[i]}}\frac{\partial z^{[i]}}{\partial W^{[i]}} = \delta^{[i]}x^{[i-1]^T}$$

$$W^{[3]} = W^{[3]} - \eta\frac{\partial E}{\partial W^{[3]}} = W^{[3]} - \eta\left(\delta^{[3]}x^{[2]^T}\right) = \begin{pmatrix}0 & 0\\0 & 0\end{pmatrix} - 0.1\left(\begin{pmatrix}-1\\1\end{pmatrix}\begin{pmatrix}0.99892 & 0.99892\end{pmatrix}\right) = \begin{pmatrix}0.09989 & 0.09989\\0.09989 & 0.09989\end{pmatrix}$$

$$W^{[2]} = W^{[2]} - \eta\frac{\partial E}{\partial W^{[2]}} = W^{[2]} - \eta\left(\delta^{[2]}x^{[1]^T}\right) = \begin{pmatrix}1 & 1 & 1\\1 & 1 & 1\end{pmatrix} - 0.1\left(\begin{pmatrix}0\\0\end{pmatrix}\begin{pmatrix}0.99999 & 0.76159 & 0.99999\end{pmatrix}\right)$$
$$= \begin{pmatrix}1 & 1 & 1\\1 & 1 & 1\end{pmatrix}$$

$$W^{[1]} = W^{[1]} - \eta\frac{\partial E}{\partial W^{[1]}} = W^{[1]} - \eta\left(\delta^{[1]}x^{[0]^T}\right) = \begin{pmatrix}1 & 1 & 1 & 1 & 1\\0 & 0 & 0 & 0 & 0\\1 & 1 & 1 & 1 & 1\end{pmatrix} - 0.1\left(\begin{pmatrix}0\\0\\0\end{pmatrix}\begin{pmatrix}1 & 1 & 1 & 1 & 1\end{pmatrix}\right)$$
$$= \begin{pmatrix}1 & 1 & 1 & 1 & 1\\0 & 0 & 0 & 0 & 0\\1 & 1 & 1 & 1 & 1\end{pmatrix}$$

$$\frac{\partial E}{\partial b^{[i]}} = \frac{\partial E}{\partial z^{[i]}}\frac{\partial z^{[i]}}{\partial b^{[i]}} = \delta^{[i]}$$

$$b^{[3]} = b^{[3]} - \eta \frac{\partial E}{\partial b^{[3]}} = b^{[3]} - \eta\left(\delta^{[3]}\right) = \begin{pmatrix}0\\0\end{pmatrix} - 0.1\begin{pmatrix}-1\\1\end{pmatrix} = \begin{pmatrix}0.1\\-0.1\end{pmatrix}$$

$$b^{[2]} = b^{[2]} - \eta \frac{\partial E}{\partial b^{[2]}} = b^{[2]} - \eta\left(\delta^{[2]}\right) = \begin{pmatrix}1\\1\end{pmatrix} - 0.1\begin{pmatrix}0\\0\end{pmatrix} = \begin{pmatrix}1\\1\end{pmatrix}$$

$$b^{[1]} = b^{[1]} - \eta \frac{\partial E}{\partial b^{[1]}} = b^{[1]} - \eta\left(\delta^{[1]}\right) = \begin{pmatrix}1\\1\\1\end{pmatrix} - 0.1\begin{pmatrix}0\\0\\0\end{pmatrix} = \begin{pmatrix}1\\1\\1\end{pmatrix}$$

2) Cálculos da propagação:

$$x^{[0]} = \begin{pmatrix}1\\1\\1\\1\\1\end{pmatrix}$$

$$z^{[1]} = W^{[1]}x^{[0]} + b^{[1]} = \begin{pmatrix}1&1&1&1&1\\0&0&0&0&0\\1&1&1&1&1\end{pmatrix}\begin{pmatrix}1\\1\\1\\1\\1\end{pmatrix} + \begin{pmatrix}1\\1\\1\end{pmatrix} = \begin{pmatrix}6\\1\\6\end{pmatrix} \qquad x^{[1]} = \tanh\begin{pmatrix}6\\1\\6\end{pmatrix} = \begin{pmatrix}0.99999\\0.76159\\0.99999\end{pmatrix}$$

$$z^{[2]} = W^{[2]}x^{[1]} + b^{[2]} = \begin{pmatrix}1&1&1\\1&1&1\end{pmatrix}\begin{pmatrix}0.99999\\0.76159\\0.99999\end{pmatrix} + \begin{pmatrix}1\\1\end{pmatrix} = \begin{pmatrix}3.76157\\3.76157\end{pmatrix} \qquad x^{[2]} = \tanh\begin{pmatrix}3.76157\\3.76157\end{pmatrix} = \begin{pmatrix}0.99892\\0.99892\end{pmatrix}$$

$$z^{[3]} = W^{[3]}x^{[2]} + b^{[3]} = \begin{pmatrix}0&0\\0&0\end{pmatrix}\begin{pmatrix}0.99892\\0.99892\end{pmatrix} + \begin{pmatrix}0\\0\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix} \qquad x^{[3]} = \text{softmax}\begin{pmatrix}0\\0\end{pmatrix} = \begin{pmatrix}0.5\\0.5\end{pmatrix}$$

Cálculo dos deltas:

$$\frac{\partial x_i^{[3]}}{\partial z_i^{[3]}} = \frac{\partial}{\partial z_i^{[3]}}\left(\frac{e^{z_i^{[3]}}}{\sum_{k=1}^{2} e^{z_k^{[3]}}}\right) = \frac{e^{z_i^{[3]}}\left(\sum_{k=1}^{2} e^{z_k^{[3]}}\right) - \left(e^{z_i^{[3]}}\right)^2}{\left(\sum_{k=1}^{2} e^{z_k^{[3]}}\right)^2} = x_i^{[3]} - \left(x_i^{[3]}\right)^2 = x_i^{[3]}\left(1 - x_i^{[3]}\right)$$

$$\frac{\partial x_i^{[3]}}{\partial z_j^{[3]}} = \frac{\partial}{\partial z_j^{[3]}}\left(\frac{e^{z_i^{[3]}}}{\sum_{k=1}^{2} e^{z_k^{[3]}}}\right) = \frac{-e^{z_i^{[3]}}e^{z_j^{[3]}}}{\left(\sum_{k=1}^{2} e^{z_k^{[3]}}\right)^2} = -x_i^{[3]}x_j^{[3]}$$

$$\delta_i^{[3]} = \frac{\partial E}{\partial z_i^{[3]}} = \frac{\partial}{\partial z_i^{[3]}}\left(-\sum_{j=1}^{2} t_j \log\left(x_j^{[3]}\right)\right) = -\sum_{j=1}^{2}\frac{t_j}{x_j^{[3]}}\frac{\partial x_j^{[3]}}{\partial z_i^{[3]}} = -\frac{t_i}{x_i^{[3]}}\frac{\partial x_i^{[3]}}{\partial z_i^{[3]}} - \sum_{j\neq i}\frac{t_j}{x_j^{[3]}}\frac{\partial x_j^{[3]}}{\partial z_i^{[3]}} = -\frac{t_i}{x_i^{[3]}}x_i^{[3]}\left(1 - x_i^{[3]}\right) - \sum_{j\neq i}\left(\frac{t_j}{x_j^{[3]}}\left(-x_j^{[3]}x_i^{[3]}\right)\right) = -t_i\left(1 - x_i^{[3]}\right) + \sum_{j\neq i}\left(t_j x_i^{[3]}\right) = -t_i + x_i^{[3]}\left(t_i + \sum_{j\neq i} t_j\right) = -t_i + x_i^{[3]}\sum_{l=1}^{2} t_l = x_i^{[3]} - t_i$$

$$\delta^{[3]} = \begin{pmatrix}\delta_1^{[3]}\\\delta_2^{[3]}\\\delta_3^{[3]}\end{pmatrix} = x^{[3]} - t = \begin{pmatrix}0.5\\0.5\end{pmatrix} - \begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}-0.5\\0.5\end{pmatrix}$$

$$\delta^{[2]} = \left(\frac{\partial z^{[3]}}{\partial x^{[2]}}\right)^T \cdot \delta^{[3]} \circ \frac{\partial x^{[2]}}{\partial z^{[2]}} = W^{[3]^T} \cdot \delta^{[3]} \circ \left(1 - \tanh^2\left(z^{[2]}\right)\right) = \begin{pmatrix}0&0\\0&0\end{pmatrix}^T \cdot \begin{pmatrix}-0.5\\0.5\end{pmatrix} \circ \left(1 - \begin{pmatrix}0.99784\\0.99784\end{pmatrix}\right) = \begin{pmatrix}0\\0\end{pmatrix}$$

$$\delta^{[1]} = \left(\frac{\partial z^{[2]}}{\partial x^{[1]}}\right)^T \cdot \delta^{[3]} \circ \frac{\partial x^{[1]}}{\partial z^{[1]}} = W^{[2]^T} \cdot \delta^{[2]} \circ \left(1 - \tanh^2\left(z^{[1]}\right)\right) = \begin{pmatrix}1&1&1\\1&1&1\end{pmatrix}^T \cdot \begin{pmatrix}0\\0\end{pmatrix} \circ \left(1 - \begin{pmatrix}0.99998\\0.58003\\0.99998\end{pmatrix}\right) = \begin{pmatrix}0\\0\\0\end{pmatrix}$$

Atualização dos pesos e dos bias:

$$\frac{\partial E}{\partial W^{[i]}} = \frac{\partial E}{\partial z^{[i]}}\frac{\partial z^{[i]}}{\partial W^{[i]}} = \delta^{[i]} x^{[i-1]^T}$$

$$W^{[3]} = W^{[3]} - \eta\frac{\partial E}{\partial W^{[3]}} = W^{[3]} - \eta\left(\delta^{[3]} x^{[2]^T}\right) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} - 0.1\left(\begin{pmatrix} -0.5 \\ 0.5 \end{pmatrix}\begin{pmatrix} 0.99892 & 0.99892 \end{pmatrix}\right)$$
$$= \begin{pmatrix} 0.04995 & 0.04995 \\ -0.04995 & -0.04995 \end{pmatrix}$$

$$W^{[2]} = W^{[2]} - \eta\frac{\partial E}{\partial W^{[2]}} = W^{[2]} - \eta\left(\delta^{[2]} x^{[1]^T}\right) = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} - 0.1\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}\begin{pmatrix} 0.99999 & 0.76159 & 0.99999 \end{pmatrix}\right)$$
$$= \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

$$W^{[1]} = W^{[1]} - \eta\frac{\partial E}{\partial W^{[1]}} = W^{[1]} - \eta\left(\delta^{[1]} x^{[0]^T}\right) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} - 0.1\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \end{pmatrix}\right)$$
$$= \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

$$\frac{\partial E}{\partial b^{[i]}} = \frac{\partial E}{\partial z^{[i]}}\frac{\partial z^{[i]}}{\partial b^{[i]}} = \delta^{[i]}$$

$$b^{[3]} = b^{[3]} - \eta\frac{\partial E}{\partial b^{[3]}} = b^{[3]} - \eta\left(\delta^{[3]}\right) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - 0.1\begin{pmatrix} -0.5 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 0.05 \\ -0.05 \end{pmatrix}$$

$$b^{[2]} = b^{[2]} - \eta\frac{\partial E}{\partial b^{[2]}} = b^{[2]} - \eta\left(\delta^{[2]}\right) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - 0.1\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$b^{[1]} = b^{[1]} - \eta\frac{\partial E}{\partial b^{[1]}} = b^{[1]} - \eta\left(\delta^{[1]}\right) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - 0.1\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

II. Programming and critical analysis

3) Após testar o modelo pretendido para valores de 0.1, 1 e 10 no parâmetro de regularização (alpha), obtivemos os seguintes valores de eficácia:

	alpha=0.1	alpha=1	alpha=10
Sem *early stopping*	93.71%	93.56%	95.76%
Com *early stopping*	89.30%	89.16%	91.06%

Decidimos assim fixar um alpha de valor 10, uma vez que foi a opção que permitiu uma maior eficácia, na presença e ausência de *early stopping.*

Obtemos, de seguida, as matrizes de confusão para ambos os casos através da soma das matrizes de confusão de cada um dos 5 *folds* utilizados.

Treino sem *early stopping*

		Valores Reais	
		P	N
Valores Previstos	P	426	18
	N	11	228

Treino com *early stopping*

		Valores Reais	
		P	N
Valores Previstos	P	384	60
	N	1	238

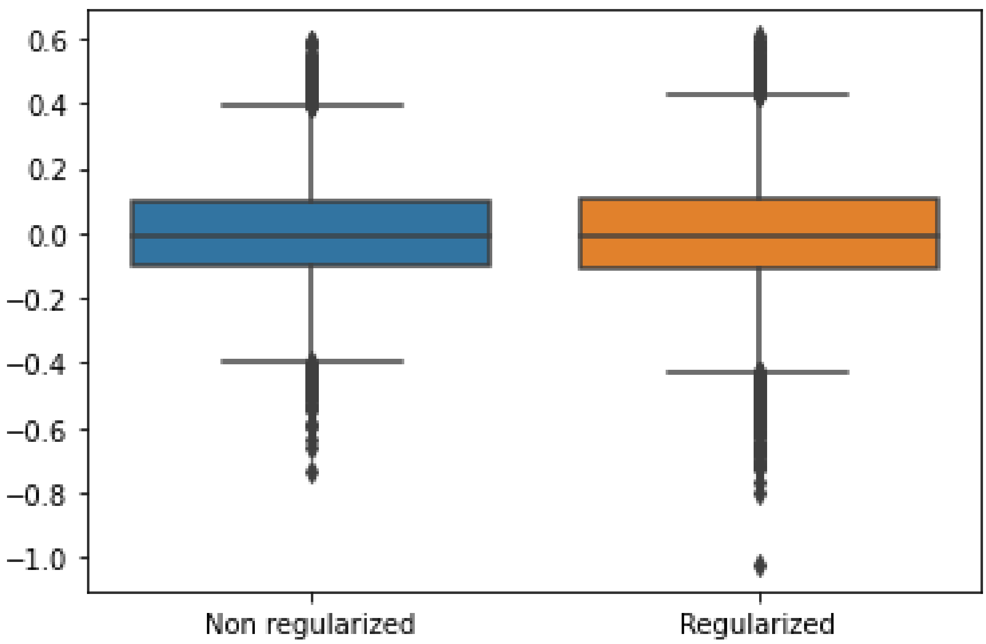
Podemos observar que ao aplicar a técnica de *early stopping,* a eficácia do modelo acaba por piorar sobretudo devido à menor quantidade de verdadeiros positivos e maior quantidade de falsos positivos, comparativamente à solução sem *early stopping*.

Uma das razões que pode levar à existência destas diferenças é o facto de, ao aplicar *early stopping*, a estabilização do erro no conjunto de validação se poder tratar de um mínimo local para o erro do modelo, o que leva à execução de um número de épocas inferior ao necessário para atingir o mínimo absoluto do mesmo erro.

Para além disso, estas diferenças podem ainda dever-se ao facto de, na aplicação de *early stopping*, parte do conjunto de dados de treino ser utilizado como conjunto de validação o que leva a uma menor dimensão do conjunto de dados de treino, o que consequentemente pode levar a uma diminuição da eficácia do modelo.

4) Após testar o modelo de regressão com regularização para valores de 0.1, 1 e 10 no parâmetro de regularização (alpha), verificamos que a soma dos erros quadrados do modelo no conjunto de treino era menor para o valor 0.1.

Logo, fixando esse parâmetro para o modelo com regularização, obtivemos a seguinte distribuição dos resíduos para um modelo com regularização e para um modelo sem regularização.



De forma a minimizar o erro dos modelos podemos adotar estratégias tais como: aumentar a dimensão do conjunto de treino permitindo uma maior abrangência de dados e um menor erro; selecionar as variáveis mais correlacionadas com a variável de output; alterar o número de camadas internas da rede bem como o número de percetrões de forma a descobrir a arquitetura que mais se adequa ao problema em questão; e, tal como feito neste exercício, testar diferentes níveis de regularização de maneira a encontrar um nível que permita um ajustamento razoável aos dados evitando ainda o *overfitting*.

III. APPENDIX

```
import numpy as np

import seaborn as sns
from sklearn.neural_network import MLPClassifier, MLPRegressor
from sklearn import model_selection
from sklearn.metrics import confusion_matrix
import pandas as pd
from scipy.io import arff

data = arff.loadarff('breast.w.arff')
df = pd.DataFrame(data[0])
df.dropna(inplace=True)
df.replace(b'benign', 0, inplace=True)
df.replace(b'malignant', 1, inplace=True)
data = df.drop(["Class"],axis=1).values
target = df["Class"].values
fold = model_selection.KFold(n_splits=5, shuffle=True, random_state=0)
conf_matrix1 = np.matrix([[0,0],[0,0]])
conf_matrix2 = np.matrix([[0,0],[0,0]])
for train_filter, test_filter in fold.split(data):
    data_train, data_test, target_train, target_test = data[train_filter],
data[test_filter], target[train_filter], target[test_filter]
    mlp1 =
MLPClassifier(hidden_layer_sizes=[3,2],alpha=10,shuffle=True,random_state=0).fit(data_train
, target_train)
    mlp2 =
MLPClassifier(hidden_layer_sizes=[3,2],alpha=10,shuffle=True,random_state=0,early_stopping=
True).fit(data_train, target_train)
    conf_matrix1 = conf_matrix1 + confusion_matrix(target_test, mlp1.predict(data_test))
    conf_matrix2 = conf_matrix2 + confusion_matrix(target_test, mlp2.predict(data_test))
print(conf_matrix1)
print(conf_matrix2)

data = arff.loadarff('kin8nm.arff')
df = pd.DataFrame(data[0])
df.dropna(inplace=True)
data = df.drop(["y"],axis=1).values
target = df["y"].values
fold = model_selection.KFold(n_splits=5, shuffle=True, random_state=0)
residuals1, residuals2 = [], []
for train_filter, test_filter in fold.split(data):
    data_train, data_test, target_train, target_test = data[train_filter],
data[test_filter], target[train_filter], target[test_filter]
    mlp1 =
MLPRegressor(hidden_layer_sizes=[3,2],alpha=0,shuffle=True,random_state=0).fit(dat
a_train, target_train)
    mlp2 =
MLPRegressor(hidden_layer_sizes=[3,2],alpha=0.1,shuffle=True,random_state=0).fit(d
ata_train, target_train)
    residuals1 += list(mlp1.predict(data_test) - target_test)
    residuals2 += list(mlp2.predict(data_test) - target_test)
df = pd.DataFrame(data={"Non regularized": residuals1, "Regularized": residuals2})
ax = sns.boxplot(data=df)
```

END