TÉCNICO LISBOA

Aprendizagem 2021/22

Homework III – Group 057

I. Pen-and-paper

1)

a)

Forward Propagation:

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

$$z^{[1]} = \underset{3\times 5}{W^{[1]}} \underset{5\times 1}{x^{[0]}} + \underset{3\times 1}{b^{[1]}} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 5 \\ 0 \\ 5 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 6 \\ 1 \\ 6 \end{pmatrix}$$

$$x^{[1]} = f\begin{pmatrix} 6 \\ 1 \\ 6 \end{pmatrix} = \tanh\begin{pmatrix} 6 \\ 1 \\ 6 \end{pmatrix} = \begin{pmatrix} 0{,}9999877 \\ 0{,}7615942 \\ 0{,}9999877 \end{pmatrix}$$

$$z^{[2]} = \underset{2\times 3}{W^{[2]}} \underset{3\times 1}{x^{[1]}} + \underset{2\times 1}{b^{[2]}} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0{,}9999877 \\ 0{,}7615942 \\ 0{,}9999877 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2{,}7615696 \\ 2{,}7615696 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3{,}7615696 \\ 3{,}7615696 \end{pmatrix}$$

$$x^{[2]} = f\begin{pmatrix} 3{,}7615696 \\ 3{,}7615696 \end{pmatrix} = \begin{pmatrix} 0{,}9989197 \\ 0{,}9989197 \end{pmatrix}$$

$$z^{[3]} = \underset{2\times 2}{W^{[3]}} \underset{2\times 1}{x^{[2]}} + \underset{2\times 1}{b^{[3]}} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0{,}9989197 \\ 0{,}9989197 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$x^{[3]} = f\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Back Propagation:

$$\delta^{[3]} = \frac{\partial E}{\partial x^{[3]}} \circ \frac{\partial x^{[3]}}{\partial z^{[3]}} \qquad \delta^{[i]} = \left(\frac{\partial z^{[i+1]}}{\partial x^{[i]}}\right)^T \cdot \delta^{[i+1]} \circ \frac{\partial x^{[i]}}{\partial z^{[i]}}$$

$$\frac{\partial E}{\partial x^{[L]}}(x^{[L]}, z) = \frac{\partial E}{\partial (x^{[L]}-z)^2} \frac{\partial (x^{[L]}-z)^2}{\partial (x^{[L]}-z)} \frac{\partial (x^{[L]}-z)}{\partial x^{[L]}} =$$

$$= \frac{1}{2}\left(2(x^{[L]}-z)\right) = x^{[L]} - z$$

$$\frac{\partial \tanh(x)}{\partial x} = \frac{\partial \frac{e^x - e^{-x}}{e^x + e^{-x}}}{\partial x} = \frac{(e^x+e^{-x})(e^x+e^{-x}) - (e^x-e^{-x})(e^x-e^{-x})}{(e^x-e^{-x})^2} =$$

$$= 1 - \tanh^2(x)$$

$$\frac{\partial x^{[L]}}{\partial z^{[L]}} = \left(\tanh(z^{[L]})\right)' = 1-\tanh(z^{[L]})^2 \qquad \frac{\partial z^{[L]}}{\partial b^{[L]}} = 1$$

$$\frac{\partial z^{[L]}}{\partial x^{[L-1]}} = W^{[L]} \qquad \frac{\partial z^{[L]}}{\partial W^{[L]}} = x^{[L-1]}$$

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

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

$$= \begin{pmatrix}0 & 0\\0 & 0\end{pmatrix}\begin{pmatrix}-1\\1\end{pmatrix} \circ \left(\begin{pmatrix}1\\1\end{pmatrix} - \tanh^2\begin{pmatrix}3{,}7615696\\3{,}7615696\end{pmatrix}\right) = \begin{pmatrix}0\\0\end{pmatrix} \circ \begin{pmatrix}0{,}0021594\\0{,}0021594\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix}$$

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

$$= \begin{pmatrix}1 & 1\\1 & 1\\1 & 1\end{pmatrix} \cdot \begin{pmatrix}0\\0\end{pmatrix} \circ \left(\begin{pmatrix}1\\1\\1\end{pmatrix} - \tanh^2\begin{pmatrix}6\\1\\6\end{pmatrix}\right) = \begin{pmatrix}0\\0\\0\end{pmatrix} \circ \begin{pmatrix}0{,}0000245765\\0{,}4199743416\\0{,}0000245765\end{pmatrix} = \begin{pmatrix}0\\0\\0\end{pmatrix}$$

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

$$= \begin{pmatrix}1 & 1 & 1 & 1 & 1\\0 & 0 & 0 & 0 & 0\\1 & 1 & 1 & 1 & 1\end{pmatrix} - 0{,}1 \begin{pmatrix}0\\0\\0\end{pmatrix} \begin{pmatrix}1 & 1 & 1 & 1 & 1\end{pmatrix} =$$

$$= \begin{pmatrix}1 & 1 & 1 & 1 & 1\\0 & 0 & 0 & 0 & 0\\1 & 1 & 1 & 1 & 1\end{pmatrix} - 0{,}1 \begin{pmatrix}0 & 0 & 0 & 0 & 0\\0 & 0 & 0 & 0 & 0\\0 & 0 & 0 & 0 & 0\end{pmatrix} = \begin{pmatrix}1 & 1 & 1 & 1 & 1\\0 & 0 & 0 & 0 & 0\\1 & 1 & 1 & 1 & 1\end{pmatrix}$$

$$b^{[1]} = b^{[1]} - \eta \frac{\partial E}{\partial b^{[1]}} = \begin{pmatrix}1\\1\\1\end{pmatrix} - 0{,}1\ \delta^{[1]} = \begin{pmatrix}1\\1\\1\end{pmatrix}$$

Aprendizagem 2021/22

Homework III – Group 057

$$W^{[2]} = W^{[2]} - 0{,}1\,\delta^{[2]}\left(\frac{\partial z^{[2]}}{\partial W^{[2]}}\right)^T = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} - 0{,}1\begin{pmatrix} 0 \\ 0 \end{pmatrix}\begin{pmatrix} 0{,}9999877 & 0{,}7615942 & 0{,}9999877 \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

$$b^{[2]} = b^{[2]} - 0{,}1\,\delta^{[2]} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - 0{,}1\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$W^{[3]} = W^{[3]} - 0{,}1\,\delta^{[3]}\left(\frac{\partial z^{[3]}}{\partial W^{[3]}}\right)^T =$$

$$= \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} - 0{,}1\begin{pmatrix} -1 \\ 1 \end{pmatrix}\begin{pmatrix} 0{,}9989197 & 0{,}9989197 \end{pmatrix} =$$

$$= \begin{pmatrix} 0{,}1 \\ -0{,}1 \end{pmatrix}\begin{pmatrix} 0{,}9989197 & 0{,}9989197 \end{pmatrix} = \begin{pmatrix} 0{,}09989197 & 0{,}09989197 \\ -0{,}09989197 & -0{,}09989197 \end{pmatrix}$$

$$b^{[3]} = b^{[3]} - 0{,}1\,\delta^{[3]} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - 0{,}1\begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0{,}1 \\ -0{,}1 \end{pmatrix} = \begin{pmatrix} 0{,}1 \\ -0{,}1 \end{pmatrix}$$

b)

$$E(x^{[3]}, z) = -\sum_{i=1}^{d} z_i \log\left(x_i^{[3]}\right)$$

$$x^{[3]} = \text{softmax}(z^{[3]}) = \text{softmax}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$$

Os valores das restantes camadas mantêm-se iguais aos da alínea anterior.

$$\frac{\partial\, \text{softmax}(x)}{\partial x} = \begin{cases} \dfrac{\partial \frac{e^{x_i}}{\sum e^{x_k}}}{\partial x_j}, & i = j \\[2ex] \dfrac{\partial \frac{e^{x_i}}{\sum e^{x_k}}}{\partial x_j}, & i \neq j \end{cases}$$

$$\begin{cases} \dfrac{\sum e^{x_k} \frac{\partial e^{x_i}}{\partial x_j} - e^{x_i} \frac{\partial \sum e^{x_k}}{\partial x_j}}{(\sum e^{x_k})^2} \\[2ex] e^{x_i} \dfrac{\partial (\sum e^{x_k})^{-1}}{\partial \sum e^{x_k}} \cdot \dfrac{\partial \sum e^{x_k}}{\partial x_j} \end{cases}$$

$$\begin{cases} \dfrac{e^{x_i}}{\sum e^{x_k}} \cdot \dfrac{\sum e^{x_k} - e^{x_i}}{\sum e^{x_k}} = \text{softmax}(x_i)(1 - \text{softmax}(x_i)) \\[2ex] \dfrac{-e^{x_i}}{\sum e^{x_k}} \cdot \dfrac{e^{x_j}}{\sum e^{x_k}} = -\text{softmax}(x_i)\,\text{softmax}(x_j) \end{cases}$$

$$\delta_i^{[3]} = \frac{\partial E(x^{[3]}, z)}{\partial z_i} = -\frac{\partial}{\partial z_i}\sum_{k=1}^{d} z_k \log x_k^{[3]} =$$

$$= -\sum_{k=1}^{d} z_k \frac{\partial}{\partial z_i} \log x_k^{[3]} = -\sum_{k=1}^{d} z_k \frac{1}{x_k^{[3]}} \frac{\partial x_k^{[3]}}{\partial z_i} =$$

$$= -\sum_{k=i} z_k \frac{1}{x_k^{[3]}} \frac{\partial x_k^{[3]}}{\partial z_i} - \sum_{k\neq i} z_k \frac{1}{x_k^{[3]}} \frac{\partial x_k^{[3]}}{\partial z_i} =$$

$$= -\sum_{k=i} z_k \frac{1}{x_k^{[3]}}\left(x_i^{[3]}(1 - x_i^{[3]})\right) - \sum_{k\neq i} z_k \frac{1}{x_k^{[3]}}\left(-x_k^{[3]} x_i^{[3]}\right) =$$

$$= -z_i \frac{1}{x_i^{[3]}}\left(x_i^{[3]}(1 - x_i^{[3]})\right) - \sum_{k\neq i} z_k \frac{1}{x_k^{[3]}}\left(-x_k^{[3]} x_i^{[3]}\right) =$$

$$= -z_i(1 - x_i^{[3]}) + \sum_{k\neq i} t_k x^{[3]} =$$

$$= -z_i + z_i x_i^{[3]} + \sum_{k\neq i} z_k x_i^{[3]} =$$

$$= -z_i + x_i^{[3]}\left(z_i + \sum_{k\neq i} z_k\right) =$$

$$= -z_i + x_i^{[3]}\left(\sum_{k=1}^{d} z_k\right) = -z_i + x_i^{[3]} =$$

$$= x_i^{[3]} - z_i$$

$$\delta^{[3]} = \begin{pmatrix} \delta_1^{[3]} \\ \delta_2^{[3]} \end{pmatrix} = \begin{pmatrix} x_1^{[3]} - z_1 \\ x_2^{[3]} - z_2 \end{pmatrix} = \left(x^{[3]} - z \right) = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} - \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1/2 \\ 1/2 \end{pmatrix}$$

$$\delta^{[2]} = \left(W^{[3]}\right)^T \cdot \delta^{[3]} \circ \left(1 - \tanh\left(z^{[2]}\right)^2\right) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} -1/2 \\ 1/2 \end{pmatrix} \circ \begin{pmatrix} 0{,}0021594 \\ 0{,}0021594 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \circ \begin{pmatrix} 0{,}0021594 \\ 0{,}0021594 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\delta^{[1]} = \left(W^{[2]}\right)^T \cdot \delta^{[2]} \circ \left(1 - \tanh\left(z^{[1]}\right)^2\right) = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \end{pmatrix} \circ \begin{pmatrix} 0{,}0000245765 \\ 0{,}4199743416 \\ 0{,}0000245765 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

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

$$= \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} - 0{,}1 \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

$$b^{[1]} = b^{[1]} - 0{,}1 \, \delta^{[1]} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - 0{,}1 \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$W^{[2]} = W^{[2]} - 0{,}1 \, \delta^{[2]} \left(x^{[1]}\right)^T = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

$$b^{[2]} = b^{[2]} - 0{,}1 \, \delta^{[2]} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$W^{[3]} = W^{[3]} - 0{,}1 \, \delta^{[3]} \left(x^{[2]}\right)^T = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} - 0{,}1 \begin{pmatrix} -0{,}5 \\ 0{,}5 \end{pmatrix} \begin{pmatrix} 0{,}9989197 & 0{,}9989197 \end{pmatrix} =$$

$$= 0{,}1 \begin{pmatrix} 0{,}499 & 0{,}499 \\ -0{,}499 & -0{,}499 \end{pmatrix} = \begin{pmatrix} 0{,}0499 & 0{,}0499 \\ -0{,}0499 & -0{,}0499 \end{pmatrix}$$

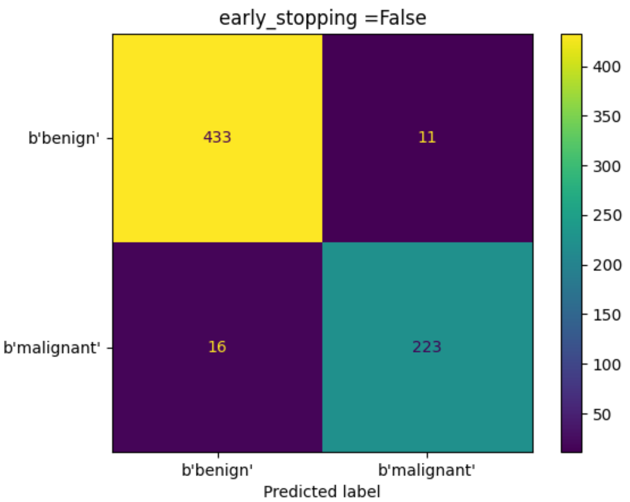
$$b^{[3]} = b^{[3]} - 0{,}1 \, \delta^{[3]} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - 0{,}1 \begin{pmatrix} -0{,}5 \\ 0{,}5 \end{pmatrix} = \begin{pmatrix} 0{,}05 \\ -0{,}05 \end{pmatrix}$$

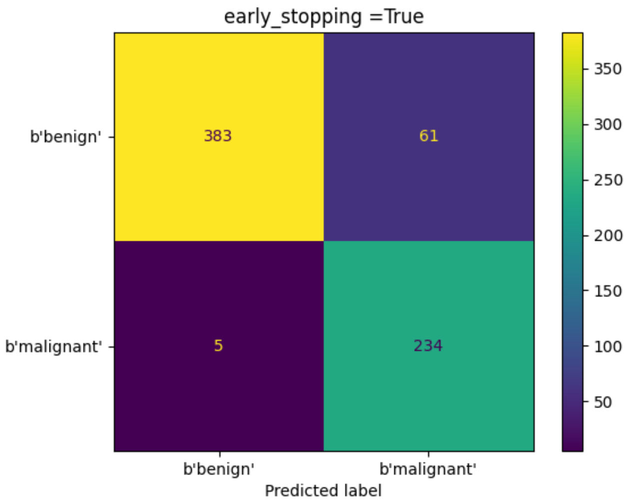
Aprendizagem 2021/22

Homework III – Group 057

II. Programming and critical analysis

2)



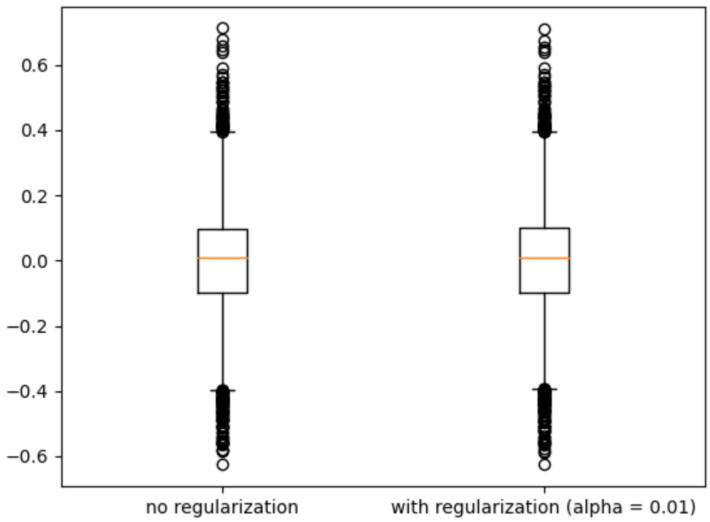


Calculando a accuracy de ambas as matrizes, concluímos que obtemos melhores resultados sem early stopping (Accuracy = (TP+TN)/(FP+FN+TP+TN), sem early stopping = 0,9605 e com early stopping = 0,9034).

Isto é possível pois o código utiliza entropia cruzada para determinar quando parar; não decide pela accuracy.

Além disso, um dos problemas com early stopping é que o modelo não usa todos os dados de treino disponíveis, e particularmente neste caso, a quantidade de dados disponíveis é muito limitada, pelo que poderá ser preferível treinar em todos os dados possíveis e evitar assim overfitting.

3)



Para este exercício, selecionámos um alpha de 0,01 pois após testarmos vários valores diferentes foi este o que nos deu melhor accuracy, ligeiramente superior à de sem regularização:

Accuracy without regularization= 0.6455

Accuracy with regularization= 0.6466.

Contudo, a diferença observada é muito pequena, pelo que concluímos que neste caso a regularização não tem um impacto estatísticamente significativo.

Portanto outras estratégias que poderíamos utilizar para minimizarmos o erro observado do regressor MLP são:

1. Obter mais dados, pois ao utilizarmos uma rede neuronal, overfitting é um problema comum
2. Normalizar/escalar os dados
3. Aplicar early stopping
4. Utilizar outros valores para a learning rate.

III. APPENDIX

```
# Ex 2

from scipy.io import arff
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import ConfusionMatrixDisplay
import matplotlib.pyplot as plt
from sklearn.model_selection import cross_val_predict, cross_val_score

data = arff.loadarff('breast.w_modified.arff')
df = pd.DataFrame(data[0])

X = df.iloc[:, 0:9]
y = df.iloc[:, -1]
y = y.astype('string')

def make_cm(es):
    clf = MLPClassifier(hidden_layer_sizes=(3,2),activation="relu",early_stopping=es,random_state=0,
                        alpha=0.0001,max_iter=2000).fit(X.values, y)
    y_pred = cross_val_predict(clf, X.values, y, cv=5)
    scores = cross_val_score(clf, X.values, y, cv=5)
    print("Accuracy when es =", es, ":", scores.mean())

    ConfusionMatrixDisplay.from_predictions(y, y_pred)
    plt.title("early_stopping =" + str(es))

make_cm(False)
make_cm(True)
plt.show()


# Ex 3

from scipy.io import arff
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPRegressor
from sklearn.metrics import ConfusionMatrixDisplay
import matplotlib.pyplot as plt
from sklearn.model_selection import cross_val_predict, cross_val_score

data = arff.loadarff('kin8nm.arff')
df = pd.DataFrame(data[0])

X = df.iloc[:, 0:8]
y = df.iloc[:, -1]
y = y.astype("float64")

fig, ax = plt.subplots()

clf_noreg = MLPRegressor(hidden_layer_sizes=(3,2),activation="relu",random_state=0,
                         alpha=0).fit(X.values, y.values)
y_pred_noreg = cross_val_predict(clf_noreg, X.values, y.values, cv=5)
scores = cross_val_score(clf_noreg, X.values, y.values, cv=5)
print("Accuracy no regularization= ", scores.mean())
residues_noreg = y-y_pred_noreg
```

```
clf_reg = MLPRegressor(hidden_layer_sizes=(3,2),activation="relu",random_state=0,
                       alpha=0.01).fit(X.values, y.values)
y_pred_reg = cross_val_predict(clf_reg, X.values, y.values, cv=5)
scores = cross_val_score(clf_reg, X.values, y.values, cv=5)
print("Accuracy with regularization= ", scores.mean())
residues_reg = y-y_pred_reg

ax.boxplot([residues_noreg, residues_reg])
ax.set_xticklabels(['no regularization', 'with regularization (alpha = 0.01)'])

plt.show()
```

END