COMPUTING SUBJECT:	Machine Learning
TYPE:	WORK ASSIGNMENT
IDENTIFICATION:	Classification MNIST
COPYRIGHT:	*Michael Claudius*
DEGREE OF DIFFICULTY:	Medium
TIME CONSUMPTION:	1-2 hours
EXTENT:	< 150 lines
OBJECTIVE:	Basic understanding of binary classification. MNIST data set

COMMANDS:

IDENTIFICATION: Classification MINIST/MICL

The Mission
To understand the idea behind classification and performance metrics regression.

Precondition
You must have done the exercises on Linear Regression in chapter 2

The problem
Given a data set with images of digits (X) and the label, the correct digit value (Y), you are to train a binary classification (the digit 5) and evaluate different performance metrics. You are to use the MNIST data set with 70.000 handwritten digits downscaled to 10.000 digits.

Figure 3-1. Digits from the MNIST dataset

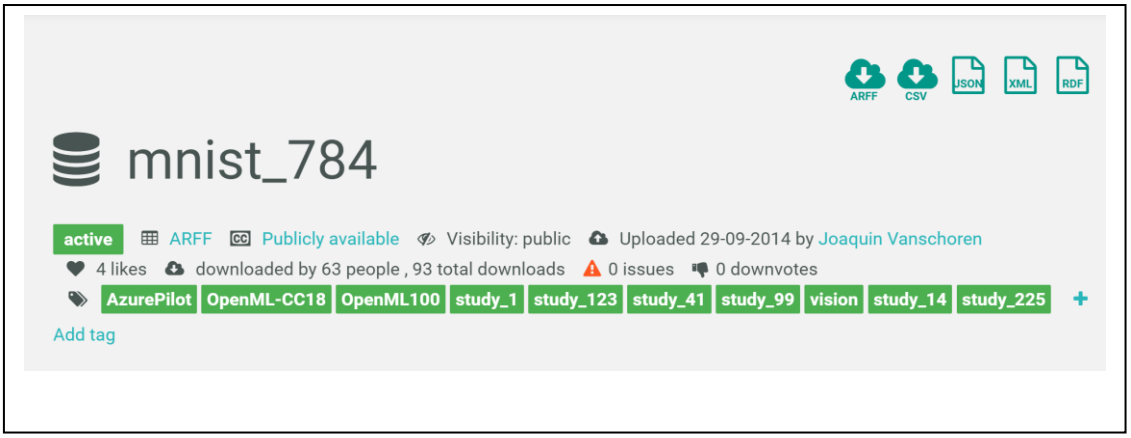
As performance measure for the classification, we will use:

Correlation matrix
Confusion matrix
Precision vs. recall
ROC-AUC

Useful links

https://www.openml.org/d/554
https://matplotlib.org/3.1.0/tutorials/introductory/pyplot.html
http://onlineconfusionmatrix.com
https://www.openml.org/d/554

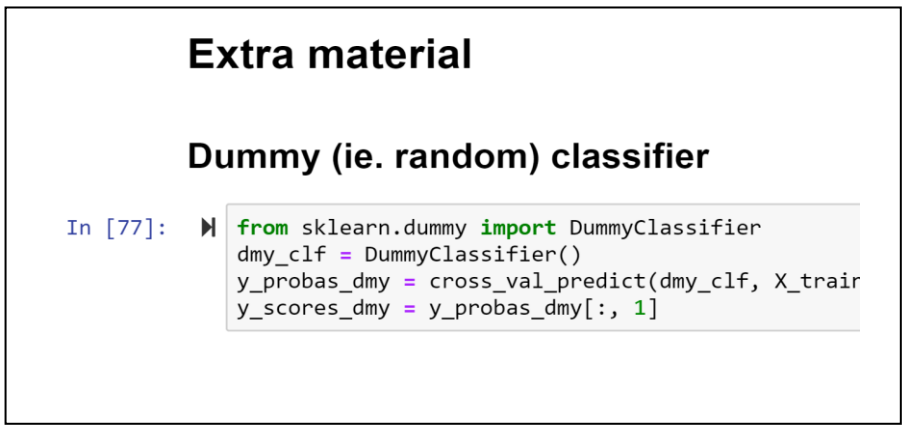
Assignment 1: Download data set
Download the data file as a .csv file from https://www.openml.org/d/554



and save it in your folder for solutions (Machine Learning/Solutions)
Copy the Chapter 3 Jupyter program, "*03-classification.ipynb*", into the same folder. Rename it "*MiniClassfication.ipynb*"

DON'T RUN THE PROGRAM,
IF YOU ARE A PATIENT DARE DEVIL, YOU CAN TRY---

Assignment 2: Application program, adjusting the program
Start Jupyter and open the file. You will now delete and out-comment some lines/blocks.
First, scroll down to the heading "**Extra**" app. Cell [77].



Delete all cells from cell [77] and down to the end. They are superfluous and some of them will take hours if not days to run on a normal laptop !

Furthermore, in order to speed up the execution time, lets down scale the number of digits from 60.000 to 10.000 (:10000) and the test set to 2.000 (68000:), by making changes to Cell [13]:

```
In [13]: X_train, X_test, y_train, y_test = X[:60000], X[60000:], y[:60000], y[60000:]
```

Finally, when I was running the program, the Standardscaler took too so long time (20 minutes on the full 70.000 set) and I got a Convergence Warning about reaching the max number of iterations:

```
In [62]: cross_val_score(sgd_clf, X_train, y_train, cv=3, scoring="accuracy")
Out[62]: array([0.87082583, 0.87089354, 0.88628294])

In [*]: from sklearn.preprocessing import StandardScaler
        scaler = StandardScaler()
        X_train_scaled = scaler.fit_transform(X_train.astype(np.float64))
        cross_val_score(sgd_clf, X_train_scaled, y_train, cv=3, scoring="accuracy")

        C:\Users\EASJ\Anaconda3\lib\site-packages\sklearn\linear_model\stochastic_gradient.py:561: Converg
        enceWarning: Maximum number of iteration reached before convergence. Consider increasing max_iter
        to improve the fit.
          ConvergenceWarning)
```

If you get the same then either raise the iteration number or out-comment Cell[62] and Cell[63] utilizing *Standardscaler*.

Now we can start to execute the cells.

Assignment 3: Binary classifier

Run the cells one by one and on the way discuss the topics and write down the answers to the following questions:

a. What is a binary classifier?
b. Why is the data set split into training and test sets?
c. How to use the SDGClassifier ? (Show the code)
d. What is K-fold cross validation?
e. How to use K-fold validation?
f. Are you satisfied with accuracy of your cross validation with 3-folds?
g. What happens if you use 4-folds?
h. What is a confusion matrix?
i. Which are the values (TP, TN, FP, FN) in your confusion matrix?
j. State the values of Precision, Recall and FalsePositiveRate?
k. Draw the ROC curves.
l. What is the ROC-AUC for SDGClassifer compared to RandomForestClassifier?
m. Describe and analyse the confusion matrix for a multi class!