Machine Learning
Assignment 2: Supervised Learning

Due date

This assignment should be submitted to Canvas before 11:59pm on **Tuesday 14/12/2021**.

Please submit a single ZIP file with your student number and name in the filename. Your submission should contain exactly 2 files:

- A detailed documentation of all code you developed, including the tests and evaluations you carried out. Please make sure that you include a .pdf document with every result you produce referencing the exact subtask and lines of code.
- All Python code you developed in a Jupyter Notebook file that can be executed and that generates the outputs you are referring to in your evaluation. Please make sure that you clearly indicate in your comments the exact subtask every piece of code is referring to.
-

Please do NOT include the input files in your submission.

You can achieve a total of 35 points as indicated in the tasks.

Objective

The Excel file "product_images.csv" on Canvas contains processed product images of sneakers and ankle boots from Zalando.com. Every row consists of a label, being either 0 for a sneaker or 1 for an ankle boot, and 28x28 8-bit grayscale pixel values of the product image (see Figure 1).
The goal of this assignment is to evaluate and optimise the performance of different classifiers for their suitability to classify this dataset.

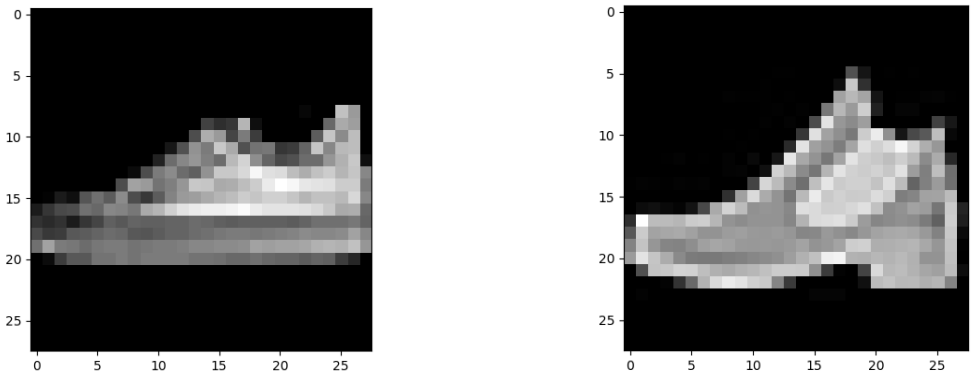

Figure 1 28x28 feature vector for an example sneaker (left) and an example ankle boot (right).

Task 1 (pre-processing and visualisation, 8 points)

Load the product images [1 point] and separate the labels [1 point] from the feature vectors [1 point]. How many samples are images of sneakers [1 point], how many samples are images of ankle boots [1 point]? Display at least one image for each class [2 points]. Parameterise the number of samples to use from the dataset in order to be able to control the runtime of the algorithm evaluation [1 point]. Start developing using a small number of samples, and increase for the final evaluation.

Task 2 (Perceptron, 11 points)

Create a k-fold cross validation procedure to split the data into training and evaluation subsets [1 point]. Train a perceptron classifier on the training subsets [1 point] and predict labels for the evaluation subsets [1 point]. Measure the processing time required for training [1 point], the processing time required for prediction [1 point], and determine the accuracy score of the classification [1 point] and the confusion matrix [1 point] for each split. Calculate the minimum, the maximum, and the average of the training time per training sample [1 point], the prediction time per evaluation sample [1 point] and the prediction accuracy [1 point]. Use a sufficient number of splits and vary the number of samples to observe the effect on runtime and accuracy [1 point].

Task 3 (Support Vector Machine, 14 points)

Create a k-fold cross validation procedure to split the data into training and evaluation subsets [1 point]. Train a support vector machine classifier on the training subsets. Try a linear kernel [1 point] and a radial basis function kernel for different choices of the parameter γ [2 points]. Predict the labels for the evaluation subsets [1 point]. Measure the time required for training [1 point], the time required for prediction [1 point], and determine the accuracy score of the classification [1 point] and the confusion matrix [1 point] for each split. Calculate the minimum, the maximum, and the average of the training time per training sample [1 point], the prediction time per evaluation sample [1 point] and

the prediction accuracy [1 point]. Determine a good value for γ based on the mean accuracies you calculated [1 point]. Use a sufficient number of splits and vary the number of samples to observe the effect on runtime and accuracy [1 point].

Task 4 (comparison, 2 points)
Compare the runtime and accuracy of the classifiers [1 point]. Which one would you choose and why? [1 point].