MTU
Ollscoil Teicneolaíochta na Mumhan
Munster Technological University

MACHINE LEARNING

Assignment 2: Supervised Learning

DUE DATE

This assignment should be submitted to Canvas before 11:59pm on **Friday 18/11/2022**.

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 it refers to.
- All Python code you developed in a single .py file that can be executed and that generates the outputs you are referring to in your evaluation. The file needs to be readable in a plain text editor, please do NOT submit a notebook file or link. Please also 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.

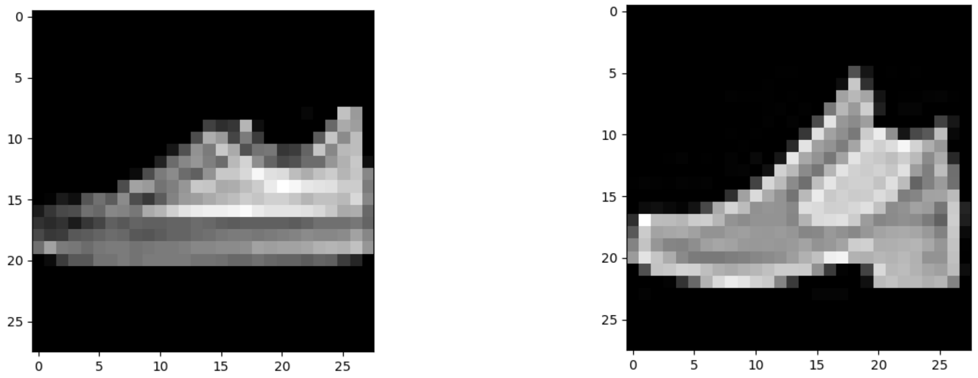

Figure 1 28x28 feature vector for an example sneaker (left) and an example ankle boot (right).

TASK 1 (pre-processing and visualisation, 5 points)

Load the product image dataset and separate the labels [1 point] from the feature vectors [1 point]. How many samples are images of sneakers, how many samples are images of ankle boots [1 point]? Display at least one image for each class [2 point].

TASK 2 (evaluation procedure, 9 points)

Create a k-fold cross-validation procedure to split the data into training [1 point] and evaluation subsets [1 point]. Parameterise the number of samples to use from the dataset 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.

Make the function flexible to accommodate different types of classifiers as required in tasks 3-6. Measure for each split of the cross-validation procedure the processing time required for training [1 point], the processing time required for prediction [1 point] and determine the confusion matrix [1 point] and accuracy score of the classification. 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].

TASK 3 (Perceptron, 3 points)

Use the procedure developed in task 2 to train and evaluate the Perceptron classifier [1 point]. What is the mean prediction accuracy of this classifier [1 point]? Vary the number of samples and plot the relationship between input data size and runtimes for the classifier [1 point].

TASK 4 (Support Vector Machine, 5 points)

Use the procedure developed in task 2 to train and evaluate the Support Vector Machine classifier [1 point]. Use a radial basis function kernel and try different choices for the parameter γ [1 point]. Determine a good value for γ based on mean prediction accuracy [1 point]. What is the best achievable mean prediction accuracy of this classifier [1 point]? Vary the number of samples and plot the relationship between input data size and runtimes for the optimal classifier [1 point].

TASK 5 (k-nearest Neighbours, 5 points)

Use the procedure developed in task 2 to train and evaluate the k-nearest neighbour classifier [1 point]. Try different choices for the parameter k [1 point] and determine a good value based on mean prediction accuracy [1 point]. What is the best achievable mean prediction accuracy of this classifier [1 point]? Vary the number of samples and plot the relationship between input data size and runtimes for the optimal classifier [1 point].

TASK 6 (Decision trees, 3 points)

Use the procedure developed in task 2 to train and evaluate the Decision tree classifier [1 point]. What is the mean prediction accuracy of this classifier [1 point]? Vary the number of samples and plot the relationship between input data size and runtimes for the classifier [1 point].

TASK 7 (comparison, 5 points)

Compare the training and prediction times of the four classifiers. What trend do you observe for each of the classifiers and why [4 points]? Also taking the accuracy into consideration, how would you rank the four classifiers and why [1 point]?