COMP8043 - MACHINE LEARNING

ASSIGNMENT 2 – Supervised Learning

SUBMISSION & DUE DATE

This assignment should be submitted to Canvas before 11:59pm on **Saturday, 16/11/2024**. The usual late submission penalties apply in accordance with MTU's marks and standards policy.

Please submit a single ZIP file containing **exactly 2 files**:

- A detailed documentation of all code you developed, including the tests and evaluations you carried out. Please 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 must be readable in a plain text editor; please do NOT submit a notebook file or link. 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.

EVALUATION PROCEDURE

You can achieve a total of 35 points for the submission as indicated in the tasks. For each subtask you are given full marks for correct answers in your submission, 70% for minor mistakes, and 35%, 15%, or 0% for major mistakes or omissions depending on severity.

To mitigate against any form of plagiarism, including but not limited to the use of generative AI tools, **you MUST also present and explain your solution and confidently answer questions** about your submission to be graded. This oral assessment will take place during the scheduled labs after the submission deadline (or the week after, should we run out of time during these sessions):

SDH4-A	Wednesday, 20/11/2024	2pm-4pm	IT2.3
SDH4-B	Thursday, 21/11/2024	11am-1pm	IT1.3
SDH4-C	Wednesday, 20/11/2024	11am-1pm	IT1.4

If you do not show up your submission will not be graded unless there are approved IECs!

Your ability to fully explain your own submission will be crucial. You will only receive full marks if you are able to confidently present and explain all your work. Your assignment grade will be capped at 70% or 50% in case of minor issues with your explanations, and at 35%, 15%, or 0% in case of major problems with your explanations during the oral assessment.

The oral assessment will cover everything you uploaded onto Canvas, so be very careful what you include into your submission and make sure that you **only submit what you fully understand** to avoid your final marks being capped; in particular, do NOT include anything, including snippets taken from labs or lectures, unless you are able to explain it in detail or risk the whole rest of the assignment being marked down.

OBJECTIVE

The Excel file "fashion-mnist_train.csv" on Canvas contains down-sampled product images from Zalando.com. Every row consists of a label and 28x28 8-bit grayscale pixel values of the product image (see Figure 1). The labels are

LABEL	PRODUCT CATEGORY
0	T-shirt/top
1	Trouser
2	Pullover
3	Dress
4	Coat
5	Sandal
6	Shirt
7	Sneaker
8	Bag
9	Ankle boot

The goal of this assignment is to evaluate and optimise the performance of different classifiers for their suitability to classify this dataset.

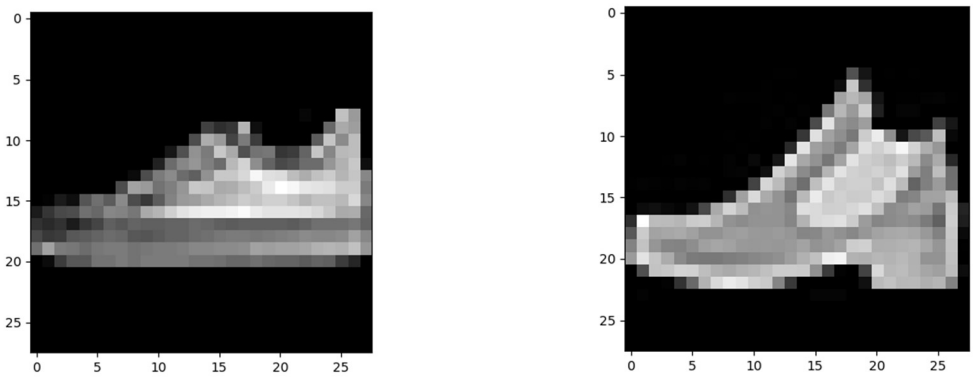

Figure 1 28x28 feature vector for an example sneaker (left) and an example ankle boot (right).

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

Load all sandals, sneakers, and ankle boots from the dataset and separate the labels [1 point] from the feature vectors [1 point]. Display one image for each class [3 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 (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 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 (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 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]?