Exam Assignment

Machine Learning (BSc DS) Fall 2024
IT University of Copenhagen

1 Introduction and formalities

This is the project description for the exam project in the Machine Learning course for the BSc program in Data Science at the IT University of Copenhagen. The project must be submitted electronically via LearnIT no later than 14.00 on 6$^{\text{th}}$ January 2025.

Groups. Please register as a group in LearnIT by the end of the day on Wednesday 9$^{\text{th}}$ October. The project is designed for groups of 2-3 students. Therefore it is strongly recommended to work in groups of 2–3 persons. If, for a very good reason, you want to do the project individually, you need to contact the lecturers first.

The course manager reserves the right to modify the grouping if necessary. Only one person for each group should submit the project.

1.1 What should be handed in?

You must hand in both a report and the source code you have developed during the project. Note that the project's evaluation will be based almost entirely on the report; the source code should be seen as a supporting document.

Make sure to use correct references to works of other people in your report, including references to any of the course textbooks. This also applies to code; if you copy or take inspiration from code developed by others, this should be stated clearly in your report. Note that any work based on a previous exam submission for this course – even if it is your own work – needs to be clearly marked as such and cited.

Report. The report should be submitted as a single PDF file. There is a strict limit of 10 pages, including figures, tables, code snippets, references, and appendixes, but excluding the front page. The project must be typeset with at least 11pt font size and margins of at least 2 cms. The report must be in PDF format and have a front page that meets the ITU requirements.[1]

Implementation and code. Your implementation has to be in Python. The code must be handed in as a single file (either a zip or tar archive). Except where explicitly stated, there are no restrictions as to which Python libraries you may use.

Your code should be organised such that it is easy to read, i.e. you have to use descriptive names for files, functions, variables, etc. The code may be organised in regular Python source files (.py files) or Jupyter notebooks.

[1] Found at `https://itustudent.itu.dk/study-administration/exams/submitting-written-work`

2 Problem and data set

In this project we will explore different methods for determining the type of clothing from an image of the item.

The data for the project consists of 15,000 labelled images of clothing based on images from the Zalando website (Xiao et al., 2017). Each image is a grayscale 28x28 picture of either a t-shirt/top, trousers, a pullover, a dress, or a shirt (see Figure 1).

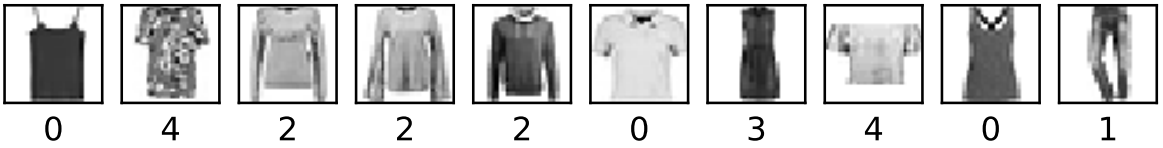

Figure 1: A random sample of 10 images from the training dataset.

Table 1: Categories of clothes

Type of clothing	T-shirt/top	Trouser	Pullover	Dress	Shirt
Label	0	1	2	3	4

The images are divided into a training set of 10,000 images and a test set of 5,000 images. The images and associated labels are available in NPY format as:

`fashion_train.npy` and `fashion_test.npy`.

Each line describes a piece of clothing. The first 784 columns are the pixel values of the 28x28 grayscale image, each taking an integer value between 0 and 255. The last column, number 785, is the category of clothing and takes values in $\{0, 1, 2, 3, 4\}$ (see Table 1).

3 Scientific requirements to the project

This project aims to investigate methods for determining the type of clothing from an image. You should carry out and report on the investigations making sure that you cover all of the tasks set out below.

3.1 Exploratory data analysis and visualisation

You should carry out an analysis of the fashion data by illustrating selected aspects of the data. You should also present a visualization of the dataset based on dimensionality reduction obtained from a principal components analysis.

You are recommended to consider applying some kind of feature scaling to the pixel values as part of your analyses of the data.

3.2 Classification

You should explore at least three classification methods:

M1. Decision trees

M2. Feed-forward neural networks

M3. One or more classification methods of your own choice.

The first two methods, M1, and M2 should be implemented in two versions:

1. An implementation from scratch (training and prediction) using only Python standard libraries and the numerical libraries NumPy and SciPy.

   Thus, you cannot use machine learning libraries such as TensorFlow, PyTorch, or Keras. Note that the restriction only applies to the implementation of the method itself (training and prediction), but not any pre-processing before you feed the input to the method or the further interpretations of the results, such as visualisations.

2. A "reference implementation" using any Python library, which can assist in asserting the correctness of your own implementation.

For the third method you may use any library you wish with no restrictions.

3.3 Report

Exploratory data analysis. Your report should introduce the reader to the data by illustrating selected aspects of the data.

Visualization of data. As part of the exploratory data analysis, your report should present a visualization of the dataset based on dimensionality reduction obtained from a principal components analysis. Remember to include some comments on what the reader can learn from the visualization.

Details on implementations. Your report should describe and discuss the key points of how you implemented the two methods, M1 and M2. Please also include a discussion of how you have asserted your implementation's correctness.

Details on machine learning methods. For each method please make sure to include

- A description of how you applied the method to the data, including details needed for an independent reproduction of your results.
- A discussion on how you have gone about selecting any hyperparameters for the method.

Interpretation and discussion of the results. Your report should include a thorough discussion of the performance of each of the methods applied. In particular, you should compare the methods' performance and guide the reader in interpreting the results. Use your expert knowledge to explain the results; for instance, why do particular methods perform better than others?

References

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms.