Lab I - End-to-end ML process

module I, Introduction to Machine Learning

DT4031, *Applied Machine Learning*
Jens Lundström

Introduction

In this lab exercise you will finalize an end-to-end machine learning development pipeline (except the deployment and maintenance steps). The challenge is to do supervised learning to automatically recognize hand-written digits. The dataset that will be used is the MNIST database of handwritten digits [1] consisting of 70k images where each image depicts a digit 0-9, see Figure 1 for an example. This dataset is a classical dataset for benchmarking learning algorithms and therefore also very well documented. Despite the academic attention, this dataset is linked to real-world engineering through the application of automatic recognition of digits in the US postal service [2].

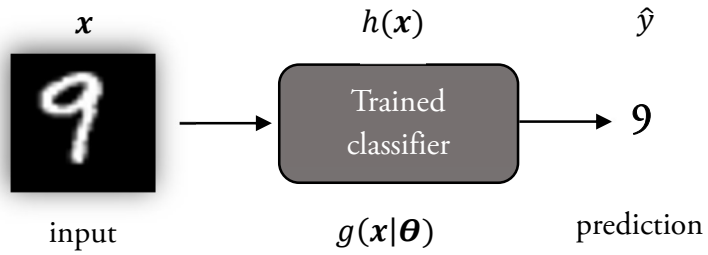

Figure 1 - Predictive modelling of handwritten digits, correctly predicting 9 given the input by using a learnt model hypothesis $g(\boldsymbol{x}|\boldsymbol{\theta})$

Lab assignment

In this lab assignment, you will play the role of a data scientist taking over an unfinished ML process pipeline and the main task is to fill in the pipeline gaps by reading the instructions from your data scientist colleague. The pipeline is in the form of a Jupyter Notebook[1] (handed over to you by a colleague) and is intended to contain the following steps:

1. Download the MNIST dataset
2. Read dataset into memory
3. Plot examples of the digits
4. Extract features from the hand-written digits by HOG (Histogram of Oriented Gradients [3])
5. Divide the data in memory into training and test sets
6. Train one (or several) models using the training dataset
7. Apply the model(s) by using the test dataset as input
8. Assess the performance of the model(s)
9. Plot evaluation metrics graphs and other figures to explore the model(s)

The Jupyter Notebook contains "TODO" messages from the colleague (most at the top of the document), the lab is about solving these "todo's" and to write a lab report, see next section *Lab report instructions*. A suggestion to start the lab is to reflect on which of the steps 1 to 9 that are implemented in the notebook and where (feel free to annotate). So, please make sure you grasp the overview of the pipeline before starting to code.

[1] Jupyter Notebook was introduced in the video material, if not available on your machine the software could be installed from Jupyter webpage or by installing the whole Anaconda suite.

Lab report instructions

The lab solution shall be documented in the form of a written report. As a data scientist it is your task to explain what you have done in a detail such that the solution is reproducible, the reasoning behind the decisions, the outcome (results) and the conclusions you have drawn from such results. It is suggested to have the following lab report structure: Introduction & Motivation; Methods & Solutions; Results; Conclusions & Discussion; Appendix. Keep the number of pages in the lab report limited, no more than five pages (excluding *Appendix* which should contain Notebook code and figures not directly linked to the result).

Lab assistance

Feel free to reach out to Kunru or Jens if you have questions regarding the lab.

Lab Resources & References

This lab requires some software to be downloaded and installed:

Some Notebook reader/editor, such as Jupyter Notebook: https://jupyter.org/install
Python module scikit-learn: https://scikit-learn.org/stable/install.html
Python module scikit-image: https://scikit-image.org/docs/stable/install.html
Python module matplotlib: https://matplotlib.org/

[1] http://yann.lecun.com/exdb/mnist/
[2] LeCun, Yann, et al. "Backpropagation applied to handwritten zip code recognition." *Neural computation* 1.4 (1989): 541-551.
[3] https://en.wikipedia.org/wiki/Histogram_of_oriented_gradients