Project 1: Supervised Learning

T-504: Introduction to Machine Learning

Fall, 2020

1 Introduction

The goal of this project is to apply your theoretic knowledge about supervised learning algorithms and the workflow of solving a supervised learning problem in practice. In particular, the task is to use a realistic dataset for a supervised learning task, design the process of preprocessing the data, selecting a model and its hyperparameters and implement a program that automates this process. You need to hand in a report outlining the process, justifying all design decisions and presenting the results.

This project can be done in groups of up to 3 students. (The groups do not have to be the same as for the labs.) The intended workload for this project is about 15-20 hours per student (not counting computation time, that is, assuming you set up the program, let it run and do something else while you wait for the results). Start working on this project in time. There will be some waiting for results of your program.

2 Data Set

You need to select the problem you want to solve, that is, which data set to learn from. The data needs to be labeled (that is, have a defined output) such that it is suitable for supervised learning and should have a decent size (at least some thousand instances). Be aware that some available data sets require additional preprocessing before they can be used for training a model.

You can find datasets to use online, e.g., on these sites:

- `https://www.kaggle.com/datasets`
- `https://archive.ics.uci.edu/`

You can also use data from other sources that you have access to.

3 Tasks

1. Design a process for finding a good model for the data set. This includes deciding
   - how to split the data into training / test set
   - which models have a good chance of working well on this data set

- what preprocessing needs to be done to use these models
- how to set the hyperparameters of these models and/or which range of values for the hyperparameters should be tried
- how to evaluate the different models to decide on the best one

Justify each one of your decisions! (E.g., if you decide on a set of potential models / values for hyperparameters then say why.) Don't forget to set data aside for testing or you won't be able to report on how well the best model you found actually performs in the end. I suggest for this part to look at publications (e.g., scientific papers) on the respective data set or on similar problems to see which models with which parameters experts used and also how they designed their experiments.

Don't forget to make a rough estimate how much time it will take to execute that process for deciding how many different configurations you can test. Maybe, you will need to do some trials by hand to see how long it takes to train some of the models for the given data set.

2. Automate the process you designed. That is, implement a program or a collection of programs that go through the process of splitting the data, preprocessing it, training and validating different models with different parameter settings to find the best model. The program should print out results for each of the trials and report on the performance of the best model.

3. Write a report in the style of a research paper on your findings. The report should be roughly structured into:

   - Introduction: describing the problem and the data
   - Process: describing the process for finding the best model and justifying all decision. If you used other papers/sources for your decisions you need to reference them here.
   - Results: report on the performance for the different models and hyperparameters. Essentially, take the numbers you got as output of your program and put them into a nice form (plots, tables, ...) that makes them easier to interpret and shows, for example, how the choice of values for different hyperparameters influences the performance.
   - Conclusions: Interpret the results and compare with results in the literature (if possible). What can be said about the performance of different models on the problem? How important is the choice of the hyperparameters for the performance of the different models?
   - Future Work: Suggest how your results could be improved. What did you learn from the results you got and what would you do differently now that you have learned this?