Project 1: Classification

T-504: Introduction to Machine Learning

Fall, 2016

1 Introduction

The goal of this project is to apply your theoretic knowledge about classification algorithms and the workflow of solving a classification problem in practice. In particular, you will use a realistic dataset for a classification task, design the process of selecting a model and its hyperparameters and implement a program that automates the process of finding the right model and hyperparameters according to your process.

2 Data Set

I suggest you use `https://archive.ics.uci.edu/ml/datasets/Letter+Recognition`. This is a data set of printed characters in different fonts. The attributes are higher level features of the images (as opposed to just pixel values) and the output are the 26 (uppercase) characters in the English alphabet. You can work on a different problem / use a different data set, if you want to, just ask whether it is appropriate for the project. The dataset should be for a classification problem and should have a decent size (some thousand instances). Be aware that some available data sets require additional preprocessing before they can be used for training a classifier.

3 Tasks

1. Design a process for finding a good classifier for the data set. This includes deciding

   - how to split the data into training / test set
   - which classifiers have a good chance of working well on this data set
   - how to set the hyperparameters of these classifiers and/or which range of values for the hyperparameters should be tried
   - how to evaluate the different classifiers to decide on the best one

   Justify each one of your decisions! (If you decide on a set of potential classifiers / 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 classifier you found actually performs in the end. It might be helpful for this stage to look at the papers published on this data set or on similar

problems to see which classifiers with which parameters they used and also how the 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 classifiers 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 f splitting the data, training and validating different classifiers with different parameter settings to find the best classifier. The program should print out results for each of the trials and report on the performance of the best classifier.

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 classifier and justifing 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 classifiers and hyperparameters. Essentially, take the numbers you got as output of your program and put them into a nice form (tables, graphs, ...) that makes them easier to interpret.
   - Conclusions: Interpret the results and compare with results in the literature you looked at. What can be said about the performance of different classifiers on the problem? How important is the choice of the hyperparameters for the performance of different classifiers?