Binary classification

Alexandre Allauzen
Nicolas Pécheux

2014

1 Overview

The goal of this assignment is to implement and evaluate two algorithms of binary classification: the minimum distance classifier and the perceptron. We will use the MNIST dataset, which is a standard dataset for machine learning. This dataset of handwritten digits has a training set of 60,000 examples, and a test set of 10,000 examples. The digits have been size-normalized and centered in a fixed-size image. This dataset is designed for multiclass classification and not for binary classification. However, we can convert this data set for binary classification purpose. For instance, for this assignment, the goal is to build a binary classifer that can answer the following question: does an image correspond to the digit "7" or not. Keep in mind that this targeted digit could be changed and be sure that it corresponds to a variable in your code.

2 Data handling

First of all take a look at the data and see what we can easily do with python: `http://perso.limsi.fr/Individu/allauzen/webpages/pmwiki.php?n=Cours.Main#toc7`

Write a program to load the dataset and take a look at some images.

Question 1. What is the structure of the dataset? How is represented each image?

3 Minimum distance classifier

Question 2. Recall the principle of this classifier.

Question 3. What is the distance between two images? How can we compute it?

Implement that function.

Question 4. Write the pseudo-code of the Minimum distance classifier for the two steps: training and inference. The goal is to efficiently design the code. Thus, you can describe each step as a sequence of functions and then describe each function as a sequence of sub-functions and so on.

Question 5. Implement the classifier in two functions: *train* and *predict*.

For experimentation purpose, take 1 000 training examples and 1 000 test examples. Compute the error rate on these both datasets. Finally re-run the experiments with another targeted digit: "0".

4 The perceptron algorithm

Do the same with this classifier and compare the results.