University of Southern Denmark

Statistical machine learning

KNN

Author:
Keerthikan Ratnarajah
kerat12@student.sdu.dk
Mikael Westermann
miwes12@student.sdu.dk

Supervisor:
Norbert Krüger

Dato: February 15, 2016

Contents

0.1 Introduction

This report documents the results of the first exercise within a project concerned with digit recognition. The purpose of the project is to develop a system capable of recognizing (classifying) hand written digits (0 - 9) using differend machine learning techniques, and documenting the performance of the system. The first exercise, documented here, was to measure the impact of image density (DPI), smoothing kernel size and parameter k of the k-NN algorithm on the misclassification error rate. The datasets used for training and testing of the system were hand written by two students (the authors), and each consist of 400 examples of each digit (4000 digits in each dataset). See 1 for example data.

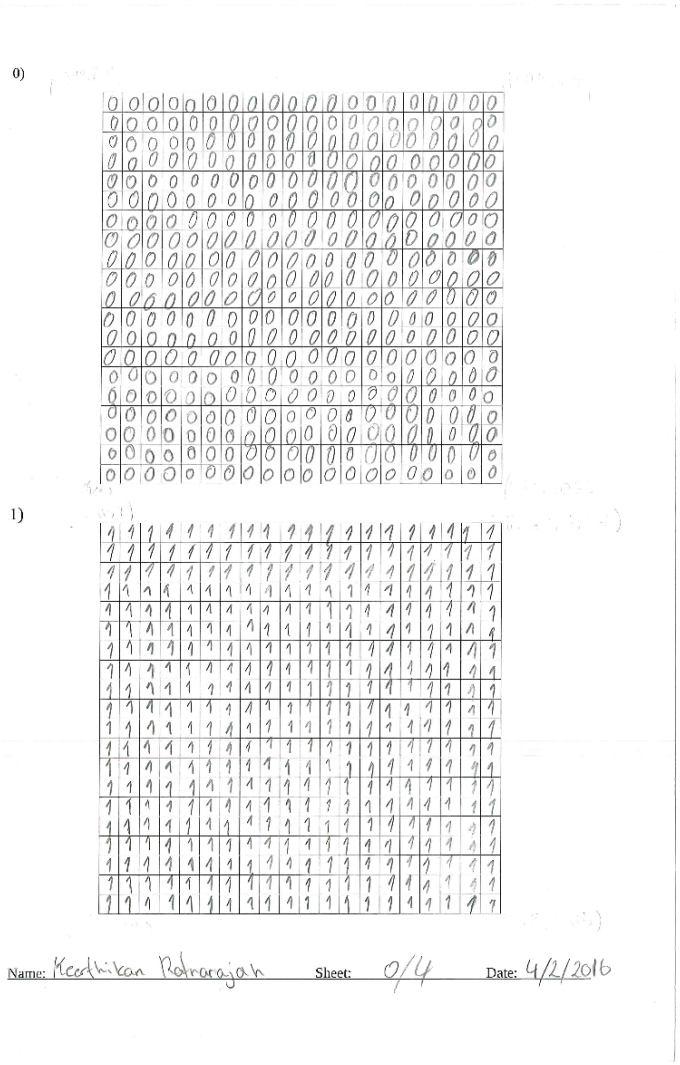

Figure 1: Example of the dataset

The process of recognizing the digits can be divided into 3 steps.

- Preprocessing - Extracting the data, and discarding irrelevant information.

- Feature extraction - Extracting relevant features.
- Classification - Using the extracted features for digit classification.

0.2 kNN

K - nearest - neighbor or kNN, is a method for classifying objects based on the distance to features extracted from ones training data. The idea in kNN is to identiy k observation which is closest to the new observation trying to be classified. The class with most vote determines which class the new observation will be classified as. K is arbitrarily choosen by the user, and has an effect of the peformance of the algorithm. k-nn has an optimal classification rate when k becomes very large, but will computation wise take more time and vice versa.

0.3 Data processing

0.3.1 Knn

The dataprocessing will consist of finding the optimal k and smoothing and DPI which lowers the error rate This is found by applying knn on different training set, with different smoothing levels, and DPI. For each case will an contour plot be made, which shows that how each parameter effect each other, and thereby be useful for deciding which parameter gives the optimal peformance.