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 14, 2016

Contents

0.1 Introduction

Something...

The purpose of this report is to develop a system capable of recognizing hand writting characters such as digits (0 - 9). This report will contain different approaches of classifying this, and their performance. The dataset used for training and testing the performance of this, has be been made by the students of the statical machine learning class as seen in 1. The dataset consist of 400 $\times$ 10 individually handwritten numbers, provided by each student of the class.

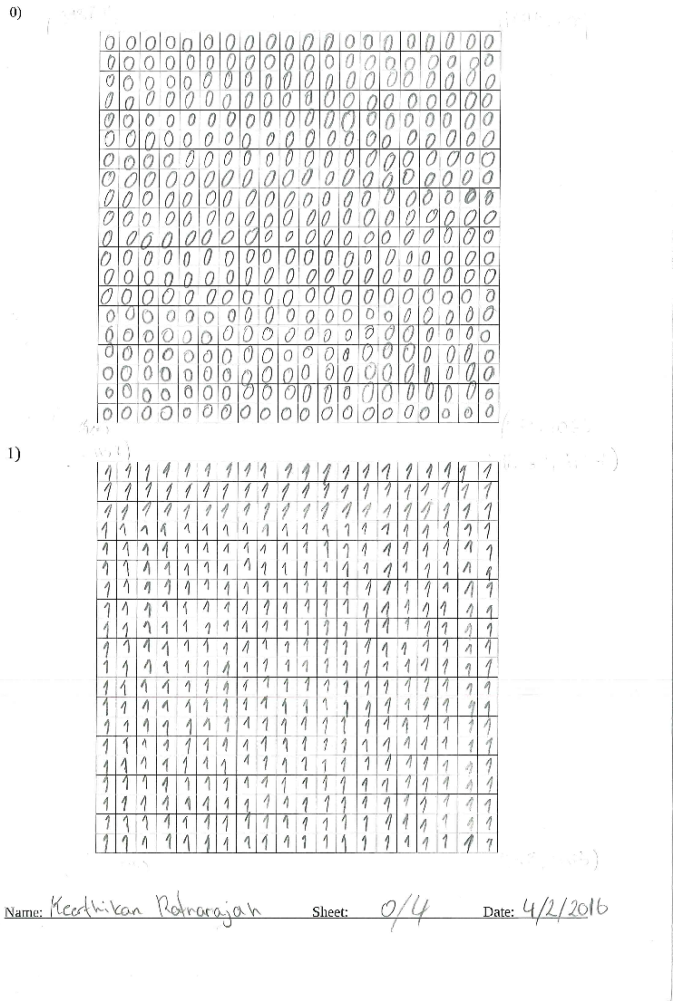

Figure 1: Example of the dataset

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

- Preprocessing - Extracting the data, and discard irrelevant information.
- Feature extraction - Extracting relevant features

- Classification - Use the extracted to classify the digits.

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.