Machine Learning Engineer Nanodegree

Capstone Proposal – Digit Recognition

Domain Background

Recognition of the character is fundamental, but the most difficult in the field of pattern recognition with a large number of useful applications. This was an intensive field of research, since because it's a natural way interactions between computers and people. More precise character recognition is the process of detecting and recognition of symbols from the input image and conversion of it in ASCII or other equivalent machine editorial form.

The technique by which a computer system can recognize characters and other characters written manually in natural handwriting. If the handwriting is scanned, and then understood by a computer, it is called autonomous handwriting recognition. If the handwriting is recognized when recording through the touch panel using a stylus, it is called online handwriting recognition. From the point of view of the character recognition classifier are classified into two main categories, i.e. without segmentation (global) and based on segmentation (analytical). Segmentation free, also known as a holistic approach to recognition symbol, without breaking it into subunits or symbols. Each word is represented as a set of global characteristics. While on the basis of segmentation approach. each word / ligature is segmented into subunits, both homogeneous and heterogeneous, and subunits independently.

Handwritten character processing systems are a domain and specific application, for example, it is impossible to create a common system that can handle all kinds of handwritten scripts and language. However, machine learning algorithms are bring new approaches this kind of problems.

Problem Statement

In this project, I am working for Digit Recognition problem.

The project for digit recognition refers to the classification of data from Semeion Handwritten Digit Data Set 1593 handwritten digits from around 80 persons were scanned, stretched in a rectangular box 16x16 in a gray scale of 256 values.Then each pixel of each image was scaled into a bolean (1/0) value using a fixed threshold.

Each person wrote on a paper all the digits from 0 to 9, twice. The commitment was to write the digit the first time in the normal way (trying to write each digit accurately) and the second time in a fast way (with no accuracy).

Datasets and Inputs

This data set consists of 1593 records (rows) and 256 attributes (columns).

Each entry is a handwritten digit originally scanned with a 256 gradient scale resolution (28).

Each pixel of each original scanned image was first stretched, and after scaling between 0 and 1 (setting to 0 for each pixel whose value was under the 127 value of the gray scale (127 enabled) and setting to 1 each pixel whose value in the gray scale was more 127).

Solution Statement

The solution will be based on SVM in the data set, reducing the size using PCA to a number that it can be managed efficiently. It is clearly not the best approach to solving the problem of digit recognition. but it is choosen because of efficiency (consumed time / accuracy ratio)

Benchmark Model

The benchmark model here is again SVM without applying PCA

Evaluation Metrics

It will be given as a simple measurement of accuracy as a general measure of models performance. This will give an estimate that takes into account all the right and wrong classifications. It will be as given below,

accuracy = # correctly predicted digits / # total number of digits

A high score will reflect a large number of correctly predictions and correctly predicted as a percentage of the total number of attempts at predictions.

Project Design

1. Data preparation

- Load data
- Separate the dataset into images and labels
- Visualization

2. SVM-PCA

- Splitting data for training and testing
- Define normalize function for normalizing the data
- Dimensionality Reduction using PCA
- Define the model

3. **Evaluate the model**

- Validate the classifier
- Get accuracy scores