Project Proposal - Digit Classification

Ce Zhang, Student No. 55542639
Yang Ji, Student No. 56064832

1 Problem Statement

In this project we would like to have a hands on experience of the MNIST handwritten digital classification task. Given 4,000 images of handwriting digits, the task is to learn images' corresponding digits from 2,000 training images and evaluate machine learning methods with the left 2,000 test images. The number of training set is much smaller than the traditional MNIST benchmark's, where the training set contains 60,000 images. This makes our problem more challenging. One of the state-of-the-art of solving the digital classification task with high accuracy is to leverage the Convolutional Neural Network [1] and the error rate is reduced to 0.23%. We would like to utilize the classical machine learning techniques which we have learned in class to test how high accuracy we can reach. Also, we will try to implement a CNN network to see whether we can get close to the state-of-the-art's accuracy.

2 Milestones

In this project, we would like to go through the entire machine learning task, from feature selection to result evaluation. First, we will use the dimension reduction techniques, like PCA, kPCA, LDA to reduce the dimension. Also, some other techniques like normalization or image processing will be considered. These can exclude some irrelevant features from our training set, which can help improve the performance. Several learning algorithms will be tested. We consider to test: (1) logistic regression, (2) k-nearest-neighbor, (3) Bayes classifier, (4) SVM. These algorithms have been implemented in several machine learning libraries that we can use. Since some classifiers are binary classifiers, there are mainly two strategies that we can use: one versus all and one versus one. For one versus all method, the i^{th} classifier is to separate the i^{th} class with other $k-1$ classes, when we have k classes. One versus one method is based on training $\frac{k(k-1)}{2}$ classifiers, where each classifier distinguishes two classes only. The class with more votes is selected as output. Apart from the classical machine learning techniques, we will also try to implement a CNN using Keras or Pytorch to see how far we can go to get closed to the state-of-the-art accuracy. Since we haven't learned neural network in class, we will study the basic knowledge of neural network and CNN for better implementation and understanding. Meanwhile, we will give some analysis of the reason why some classifiers work well while others do not work well, based on the understanding of the algorithms. For the evaluation, we will use 2,000 out of 4,000 images for testing. Cross-validation will be used to select the optimal parameters. We will show the performance of each preprocessing method, each learning methods and each hyper-parameters.

References

[1] D. Cireşan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," *arXiv preprint arXiv:1202.2745*, 2012.