Capstone Proposal
Digit Recognizer

Author: Allison Senden

Domain Background:
Digit recognition is being studying by those practicing their computer vision and machine learning skills. I specifically chose this project topic because I would like to get into some work at my current job where they are helping to design a model that can handle taking in and interpreting the information on faxes. This would include handwritten text and digits. I think that many companies can benefit by having a model that can take handwritten notes and interpret them into useable information. This would help to alleviate someone manually completing this task which I could imagine would take hours. If the company can use a model that can do the same work, then they can pay that person to do some more meaningful work that help get them to their goal in better way.

Problem Statement:
The goal is to be able to correctly identify the handwritten digit by applying algorithms that learn based on the training set, which is thousands of pictures of handwritten digits. I want to be able to create a model/algorithm that performs better than the baseline models. I will have to do research into neural networks and which algorithms and layers perform best for image classification.

Datasets and inputs:
The data I will be using is a very commonly known dataset called MNIST. It contains hundreds of thousands of pictures of handwritten digits that I will be able to use to solve the problem. More information about the MNIST dataset can be found here. The training data consists of 60,000 images and the test set contains 10,000 images.

I was also able to find a similar Kaggle competition. Here you can see many people taking the challenge to classify the handwritten digits properly.

Solution Statement:
In order to solve this problem, I will implement a k-nearest neighbor model that's tuned so that it can classify the images properly. I feel that it best suites the problem at hand. I think choosing a k=10 as there are 10 numbers (0-9) makes the most sense, but we will see once we do more analysis.

Benchmark Model:
Some models that are historically use to evaluate this dataset are SVM and k-nearest neighbors. I will use these 2 models as a baseline for my work, and then hopefully I create and fine tune a model that will perform even better than these baselines. K-nearest neighbor is good fit for this problem since its goal is to essentially cluster the alike images into the same cluster. This helps me because it's choosing the images that most likely are the same as those in that same cluster.

For me, I can make those clusters each of the different digits 0-9. Then when it's predicting, it will choose the cluster that the digit resembles the most.

Evaluation Metrics:
In order to evaluate the performance of the model I create, I will evaluate the accuracy. This will tell me the percentage of photos that were correctly classified as the right digit. We only care about when our model is correctly classifying images so that's why I chose accuracy as the evaluation metric.

Project Design:

- Environment: Jupyter notebook with python3
- Language: Python 3
- Libraries: pandas, numpy, keras, matplotlib, sklearn
- Strategy:
    - Analyze some of the images and decide what preprocessing steps are necessary if any
    - Implement any preprocessing that needs to be done on the images for better results
    - Ensure all images are sized the same so that the model interprets all in the same manner
    - Define the architecture of the model
    - Split the data into training and test sets
    - Train the model with the training data
    - Test the trained model with the remaining data
    - Evaluate the performance of the model using accuracy as the metric