Project Proposal **ECE/CS 532**

Nikita Bhatt | nbhatt3@wisc.edu

Project Dataset

The dataset I plan on investigating is the MNIST Database of Handwritten digits. Information about the dataset can be found here: http://yann.lecun.com/exdb/mnist/

The data set is a multi-label data set with a training set of 60,000 examples and a test set of 10,000 examples. Each image/example is a grayscale 28 x 28 pixel images of handwritten numbers 0 to 9, where each pixel represents a feature, and so the data set has 784 features associated with it. Each feature contains a value between one and zero and describes the intensity of each pixel.

The classification problem to be explored will be pertaining to building a model which will , as accurately as possible, identify and classify each example in the test set to a correct label from one of the 9 labels (1-9). This process will be narrowed down with the help of certain pre-processing steps and algorithms.

Algorithms to Investigate

Pre-analysis steps will include utilizing PCA or other subsequently needed techniques to carry out dimensionality reduction, with the main goal of bringing our data from a higher-dimensional space to a lower one while keeping most of the relevant information.

Subsequently, the following algorithms will be applied to arrive at a conclusion-

1)Linear Regression techniques will be used for predictive modeling such as lasso & least squares. Linear regression techniques assume a linear relationship between inputs and target variable. Least squares regression can be used to evaluate a linear model whereas lasso can be used to shrink the coefficients for input variables that do not contribute much to the prediction. Since these techniques are primarily for binary classification, I will convert binary classifiers to multi-class classifiers using scikit-learn.

2) The KNN algorithm will be used to first find the distance between a query and all the examples in the training set, choosing a value for ‘k’ which is nearest to the query, and then classifying to the nearest label. In order to select the right value of ‘k’, the KNN algorithm will be run several times with different K values and then the value of K that gives us the least errors will be used in the end to arrive at an model that accurately makes predictions.

3) A Neural network based algorithm will be developed with the help of the training set and used to evaluate the test data set.

The three algorithms will be evaluated by determining the confusion matrix values for the models generated using each of the three types of algorithms, and then comparing the values of False Positives and Negatives to evaluate the model efficiency. Essentially, this means that we will be evaluating the errors using each of the models and using that as a metric of evaluation.

Project Github

A GitHub repo for the project has been setup here: https://github.com/nbhatt15397/ECE-CS532-Final-Project.git

The repo is currently public but will be switched to private(and accessible only with link) as I start pushing code.

Project Timeline

Deadline	***Tasks to Accomplish***
10/21 - 10/30	Data Cleaning + Introductory Graph/Image generation + Understanding Dataset
11/01 – 11/7	Finish Analyzing data using Algo 1
11/7 – 11/14	Finish up Algo 1 and start Algo 2
11/14 - 11/17	Work on first Update Document
11/17	First Update Due
11/17-11/20	Complete work on Algo 2 and add to report
11/20-11/30	Start working on Algo 3
11/30-12/1	Work on Second Update Document
12/1	Second Update Due
12/1 – 12/5	Finish working on Algo3 and wrap up
12/5-12/10	Last Minute Wrap ups and wiggle room
12/10 – 12/12	Final Report Writing
12/12	Final Report Due
12/13 - 12/17	Peer Review of Projects
12/17	Peer Review Due