Machine Learning Progress Report

Jackson Pontsler, Micheal Bentley

Problem:

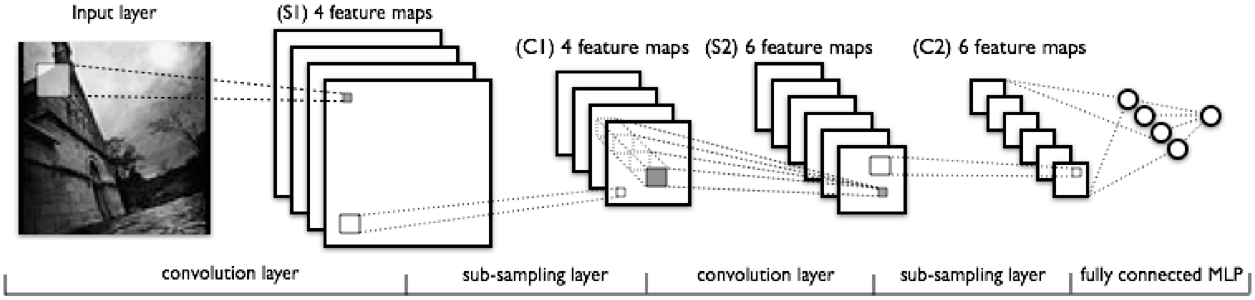
Develop a machine learning algorithm that will look at census records and be able to correctly identify the gender of an individual, ethnicity, and marital status.

What has been done:

We have contacted the Church of Jesus Christ of Latter Day Saint's Family Services and petition the usage of their data. They sent us a non-disclosure agreement requesting that we keep the data protected and confidential. We signed the papers and sent them back to the church and are now waiting for the data.

After talking with the PHD student Dustin Webb we were instructed to look into deep learning via a convolutional neural network. One of the reasons why convolutional neural networks are so powerful is because it is a deep learning algorithm that teaches itself what needs to be learned and then learns it. This had advantages because when reading common features of classify values only a few we available (reflective over an x axis, reflective over a y axis, which vertical half has the majority of the pixels, and which horizontal half has the majority of the pixels) this list of features is clearly not nearly enough; therefore we concluded that deep learning would be the best approach to classifying the data. Below summarizes what Dustin and DeepLearning.net has taught us about Convolutional Neural Networks.

What a convolutional neural network is: A convolutional neural network is a series connected feature maps that are formed by convolution and sub-sampling. To start we will have our original image which is down sampled into a subset of pixels i.e. 10x10 or 28x28. The image will then have a frame of size NxN and a defined stride S. This frame will start at some corner of the input and move with the length of the stride. With each new stride and frame a a series of weights are applied. Each weight set is applied will then go to form an individual feature map of varying weights. Then a subsampling layer is formed by taking all the feature maps in the previous layer and pooling their weights to form the next set of feature maps. This process will be repeated till a series of connected receptive field which can then be used in conjunction with a Multilayered Perceptron (which is a hidden layer logistic regression). A diagram below illustrates this process.



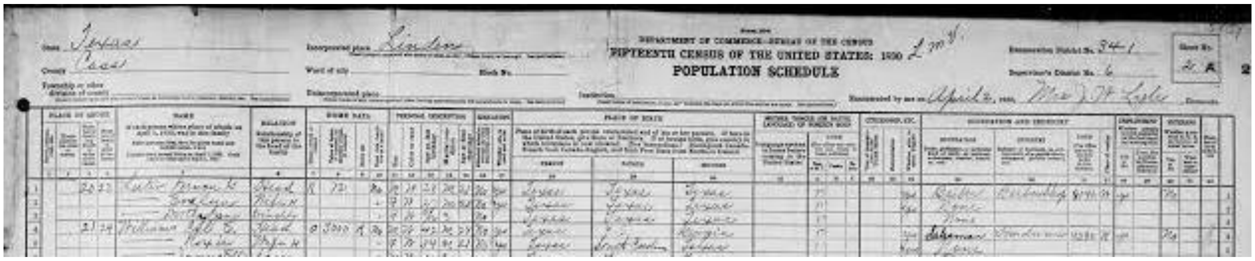
Research with Logistic Regression in identifying MNIST Digits were then studied, and summarized below. Logistic Regression is a linear classifier that takes an input and creates a series of hyperplanes which are a series of classifiers that output a probability that the input is a number. In other words the series of classifiers are probability that is 0, probability that is 1, ….probability that is 9. Logistic Regression also uses Stochastic Gradient Decent to minimize the cost/loss function.

Because we are still waiting for the church to send us the data we have improvised and are currently working on the MNIST Digits. We that if we could create a CNN of the MNIST Digits we would be well prepared for when we get the census data from the church. We also have asked the CADE support groups to install theano on the CADE machines.

Plan for completion:

Our first objective is to understand better how to use Theano to program Multilayered Perceptrons and CNNs. We will use this to create a series of MNIST classifiers. Once we have received data from the church we will built upon the previous code to design a series of alphabetical classifiers to classify gender, ethnicity, yes/no, and marital statuses of individuals.

Other details to be considered: upon receiving the data we expect it to have images like this:



DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS

FIFTEENTH CENSUS OF THE UNITED STATES: 1930

POPULATION SCHEDULE

State Texas

County Cass

Incorporated place Linden

Enumeration District No. 34-1

Supervisor's District No. 6

Sheet No. 2 A

Enumerated by me on April 2, 1930, Mrs J H Lisle, Enumerator.

Image processing will be needed to isolate individual cells or areas of interest that will need to be down sampled/compressed and then sent to our CNN code. This may be done by calculating the derivative of the image and locating edge boxes to locate areas of interest. Furthermore it has also been recommended to expand our project to include not only classifying characters but to also see if we can have the classifier predict the location of the character on the original image.