CS 395 – Final Project

Part 2

Date: April 14th, 2019

By: Lauren Simms and Joshua Eli Swick

1. Choose and describe the problem you are trying to solve. Clarity is important in your answer.

   We are attempting to predict the income, over $50,000 per year or under $50,000 per year, for a person based on census data as accurately as possible. This data includes age, race, education, marital status, income, sex, relationship status and country of origin. We are approaching this problem using classification.

2. Describe your approach in detail, taking into account all the feedback you have received (and any additional research you have performed).
   a. What kind of problem are you trying to solve? Regression? Classification? Clustering?
   b. If classification, what classes are your outputs?
   c. Why did you take this approach?

   We are trying to solve a classification problem, with the potential for it to be a clustering problem. Our outputs are above $50,000 and under $50,000.

   We left out the 'fnlwgt' column as this value is related to the sample size in the census and not a feature of an individual person represented in the data set.

   The ' <=50k' and ' >50k' target features were converted to a 0 and 1 respectively in order for them to be represented numerically.

   The model's activation functions, RelU and SoftMax, and loss function Cross Entropy are described in more detail below.

   Classification analysis is appropriate for this problem and dataset as predicting if an individual's income is above or below $50,000 per year can be represented as two categories. Additionally, as identified throughout our work this semester, the SoftMax activation function works well as the activation function layer in classification models.

3. Describe the state-of-the-art solutions by others on the same dataset.
   a. What are the differences with your approach with theirs?
   b. What are the differences according to your metrics?

We found one instance where another data scientist used this dataset for income classification. This model used the Python package ‘sklearn' and had an accuracy of 82.5%.

4. What is the metric(s) you are using? Justify their use. If you are using data from a competition, it may be that others are using that metric, so you want to compare your answer against theirs.

   This model focuses on accuracy, as we want the model to accurately predict an individual’s income category.

5. What is/are the activation functions used in your model? Where are they used? Show a few variations and how this affects your results.

   We are going to use the SoftMax and Relu activation functions. Soft max is useful because we are using classification and being able to draw-the-line between income predictions above $50,000 per year and below $50,000 per year is valuable. We are still running into data formatting issues when trying to normalize and encode the data set, therefore we are unable to provide results at this time.

6. Next, after choosing the best activation functions, what is/are the loss functions used in your model? Where are they used? Show a few variations and how this affects your results.

   The loss function used in this model is, Cross-Entropy. The loss function set during the configuration of the model in preparation for training using the model’s compile method. The loss function is related to the error value that will be minimized by the model. We are unable to show a few variations as our model and dataset have a formatting conflict.