Justin Jose
Artificial Intelligence

Project 2: Neural Network

How to run the program

Run the command: python3 main.py

Enter 0 or 1 after program begins to select training or testing program.

Dataset Explanation

The data set was created by pulling data from the YouTube api and formatting it according to the project requirements. Over the thanksgiving break I worked project with my friends to classify YouTube videos as trending or non-trending. My friend scraped trending and non-trending video ids from YouTube and stored them in corresponding csv files (trending-yt.csv and nontrending-yt.csv).

Over the break I wrote a python script (retrieve_youtube_data.py) to grab data (view count, like count, dislike count, comment count, etc.) from the YouTube api using those video ids. I repurposed the data I got from that project here by writing another python script (create_train_test_files_youtube.py) to generate training and testing files appropriate for this project. These files can be found in the 'youtube-dataset-misc-files' directory.

I standardized the data (made the data 0 mean with a standard deviation of 1) by subtracting from each value the mean of its column and divided by the standard deviation of the column. I initially tried normalizing by dividing the values by the max along each column but I was not able to get good metrics that way as sometimes there are outlier videos (e.g. a video that may have 200

million views may skew the entire column of view count closer to 0 making it difficult to classify).

Dataset Description

The training set and testing set each have 800 examples consisting of data 400 trending videos and 400 non-trending videos. Each example has 4 inputs and 1 output.

The input attributes consist of:

1. View Count
2. Like Count
3. Dislike Count
4. Comment Count

The outputs consist of:

1. Trending or Non-trending (Indicated by 1 or 0)

Initial Weights of Neural Network

The initial weights of the network were created by generating pseudo random numbers in the range of 0 to 1.

Parameters for Reasonable Learning

After some experimentation a network with 2 hidden nodes seemed to fit the task best. The network had 4 input nodes, 2 hidden nodes, and 1 output node.

Justin Jose
Artificial Intelligence

Trained with a learning rate of 0.1 and 100 epochs, the network performs well. With a learning rate of 0.05 and 200 epochs the network has slightly better metrics. The best metrics seen come from training the network with a learning rate of 0.1 and 500 epochs.

The .init, .train, .trained, .test, and .results files for the YouTube dataset have been placed in the directory ‘youtube-neural-network-files’.