Ayden Shankman
ECE - 469
Prof Sable
Fall 2022

Heart Disease Dataset

The custom dataset I used is a Heart Disease dataset found on Kaggle and can be found using the link: https://www.kaggle.com/datasets/zeeshanmulla/heart-disease-dataset. The data consists of 304 rows each representing a patient. Each row has 13 input parameters that represent patient attributes when admitted to the hospital and one label that is “0” if the patient doesn’t have heart disease or “1” if they do. The 13 input attributes are the following:

1. Age in years
2. Gender (“1” for male; “0” for female)
3. Chest pain type (“0” for typical angina; “1” for atypical angina; “2” for non-anginal pain; “3” for asymptomatic)
4. Resting blood pressure (in mm Hg on admission to the hospital)
5. Cholesterol level (in mg/dl)
6. If Fasting Blood Sugar level is more than 120 mg/dl (“0” for false; “1” for true)
7. Resting Electrocardiographic results (“0” for normal; “1” for having ST-T wave abnormality; “2” for showing probable or definite left ventricular hypertrophy by Estes' criteria)
8. Maximum Heart Rate achieved (in bpm)
9. Exercise Induced Angina (“0” for no; “1” for yes)
10. ST depression induced by exercise relative to rest
11. The slope of the peak exercise ST segment (“1” for upsloping; “2” for flat; “3” for downsloping)
12. Number of major vessels colored by Flourosopy (0,1,2, or 3)
13. A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversible defect)

The dataset was downloaded as a CSV file and converted to a pandas data frame object in python. The data was then shuffled and then split so that 75% of the data was for training and 25% for testing. The columns of each data set were then normalized using standardization (also known as Z-score) which was done by finding the mean and mean and standard deviation of each column of the training data and using the following formulas:

$$Training\ Data\ = \frac{Training\ Data - Training\ Data\ Mean}{Training\ Data\ Standard\ Deviation}$$

$$Testing\ Data\ = \frac{Testing\ Data - Training\ Data\ Mean}{Training\ Data\ Standard\ Deviation}$$

To create the initial neural network layout, I tested hidden layers of various lengths from 5 to 20 nodes. Each weight in the initial network was a random uniform number between 0 and 1. After some testing, it was found that a network with **8 hidden** nodes performed the best when used with a **learning rate of 0.1** and trained it for **100 epochs**. Using these parameters, the network achieved a **test accuracy of 93%**. All file names can be found below.

File Names

- **“Heart Disease Dataset”** - Unprocessed Heart Disease Data CSV file
- **“heart_train_data.txt”** - Processed Heart Disease Training Data
- **“heart_test_data.txt”** - Processed Heart Disease Testing Data
- **“heart_init_NN.txt”** - Initial, untrained network with 8 nodes in hidden layer
- **“heart_trained_NN.txt”** - Trained network
- **“heart_results.txt”** - Heart Disease Test Results