Philip Blumin
AI Fall 2020
Project #2 Write-up
Professor Carl Sable

How to use my program

- Use the makefile to compile
- Upon running the program the user it prompted to either go into train or test mode
    - The user must enter 0 for train and 1 for test
- In train mode:
    - The user is prompted to enter the name of a valid initial weights file
    - Next, the user is prompted to enter a valid train file
    - After a train file is provided the user must enter an output file in which they wish to have their trained weights
    - Finally, after all this the user is asked to enter the amount of epochs and the learning rate
- In test mode:
    - In test mode the user to prompted to enter the trained weights file
    - Next, the user is prompted to enter a valid test file
    - After this the user is asked to enter an output file in which they wish to have all of their final results

My Dataset

The data set I choose is an Ecoli data set taken from the following link:
http://archive.ics.uci.edu/ml/datasets/Ecoli

About the Data set

The data set classifies ecoli (provided by sequence names) into different locations in a cell
The original data set has a total of 9 columns:

1. **Sequence Name:** Accession number for the SWISS-PROT database
2. **mcg:** McGeoch's method for signal sequence recognition.
3. **gvh:** von Heijne's method for signal sequence recognition.
4. **lip:** von Heijne's Signal Peptidase II consensus sequence score.
5. **chg:** Presence of charge on N-terminus of predicted lipoproteins.
6. **aac:** score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins.
7. alm1: score of the ALOM membrane spanning region prediction program.
8. alm2: score of ALOM program after excluding putative cleavable signal regions from the sequence.
9. **Class Distribution**: The class is the localization site. There are a total 8 different classes:

a. cp (cytoplasm) -143 instances
b. im (inner membrane without signal sequence) - 77 instances
c. pp (perisplasm) - 52 instances
d. imU (inner membrane, uncleavable signal sequence) - 35 instances
e. om (outer membrane) - 20 instances
f. omL (outer membrane lipoprotein) - 5 instances
g. imL (inner membrane lipoprotein) - 2 instances
h. imS (inner membrane, cleavable signal sequence) - 2 instances

Preprocessing and Modification

Since the first column is just sequence names I decided to drop the columns. The last column (what the data set is trying to classify) I one hot encoded each of the locations. Each location site now is classified as the following

1. cp - 0 0 0
2. im - 0 0 1
3. pp - 0 1 0
4. imU - 0 1 1
5. om - 1 0 0
6. omL - 1 0 1
7. imL - 1 1 0
8. imS - 1 1 1

The data set contained a total of 376 data points. I did a random 80 - 20 split on the data, meaning the training data constrained 300 points.

Parameters

After going through several different combinations of hidden nodes, learning rate, and epochs I found the following to work pretty well:

- **15 hidden nodes**
- **100 epochs**
- **Learning rate of 0.1**

Initial Weights (ecoli_init.txt)

For the initial weights I just like in the WDBC and grades data set examples I randomly generated numbers between 0 and 1 that go up to the thousandth place.