CS615 Assignment1

Himanshu Gupta

April 2020

1 Theory

1.1

The equation of calculating Cross Entropy (H) given below:

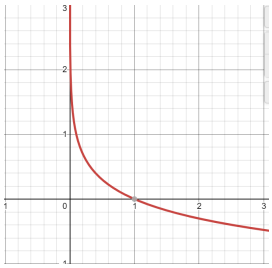
H = $\sum_{k=1}^{K} -a_k log b_k$
where a_k and b_k are two distributions.

Given the correct class is the first class making the a_k distribution [1, 0, 0] and b_k distribution is [0.1, 0.8, 0.1], we can directly compute the Cross Entropy:

H = $1 \times log(0.1) + 0 \times log(0.8) + 0 \times log(0.1) = 1 \times log(0.1) = log(0.1)$
H = -2.3025851

1.2

The Cross Entropy objective function is given below:
J = $\sum_{k=1}^{K} -y_k log \hat{y}_k$
where y_k is our true given output distribution typically one hot encoded and $\hat{y}_k$ is our activation function. Even though J is the summation, there is a negative sign and the graph would look similar to that of $-log(x)$ shown below. We are attempting to either find maxima or minima of our objective functions. In this case, function is in decreasing manner so we attempt to minimize the Cross Entropy objective function. Also as the name suggests, it is a type of loss so we would like to minimize it.



1.3

1.3.1

Class Priors(C) are the baseline accuracy. From the confusion matrix, we can deduct number of observations with true class labels. This can done by simply adding up the column values for respective class labels as columns represent true class labels in this confusion matrix. Following the above mentioned procedure, the class priors are calculated below:

C(1) = $\frac{5+8+0+10}{5+8+0+10+2+12+8+0+3+30+45+5+4+4+4+80} = \frac{23}{220} = 0.1045$

C(2) = $\frac{2+12+8+0}{5+8+0+10+2+12+8+0+3+30+45+5+4+4+4+80} = \frac{22}{220} = 0.1000$

C(3) = $\frac{3+30+45+5}{5+8+0+10+2+12+8+0+3+30+45+5+4+4+4+80} = \frac{83}{220} = 0.3773$

$$C(4) = \frac{4+4+4+80}{5+8+0+10+2+12+8+0+3+30+45+5+4+4+4+80} = \frac{92}{220} = 0.4181$$

1.3.2

To calculate the overall accuracy, we need to find the number of observations classified correctly, and divide that by total number of observations.

$$\text{accuracy} = \frac{1}{N}\sum_{k=1}^{K}(Y_k == \hat{Y}_k)$$

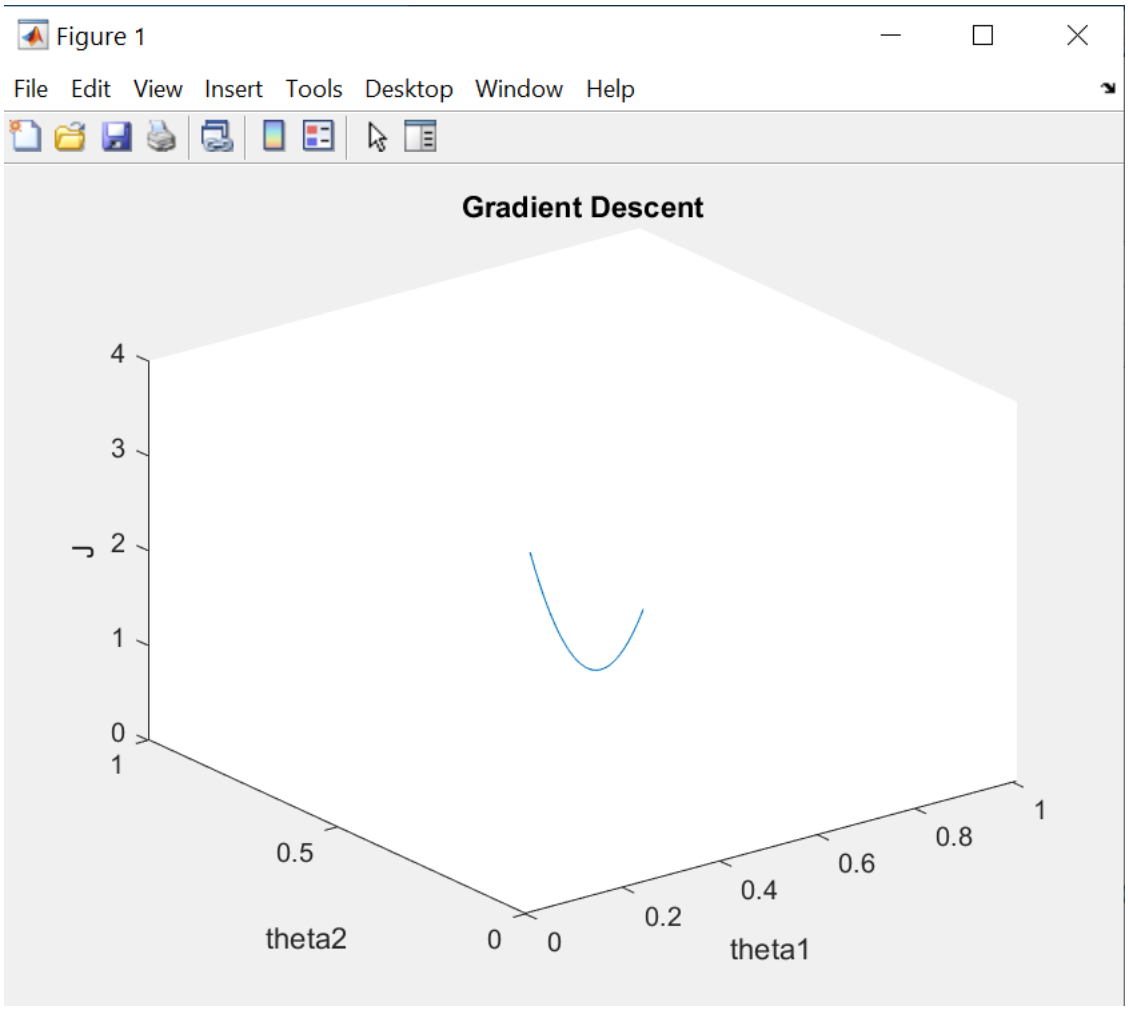
In the confusion matrix, diagonals are the one representing number of observations classified correctly so accuracy would be sum of diagonals (trace(Confusion Matrix)) divided by total number of observations.

$$\text{accuracy} = \frac{trace(ConfusionMatrix)}{N} = \frac{5+12+45+80}{5+8+0+10+2+12+8+0+3+30+45+5+4+4+4+80}$$

$$\text{accuracy} = \frac{142}{220} = 0.6454$$

2 Gradient Descent

The plot is shown below and the source code is in file named "GD1.m":

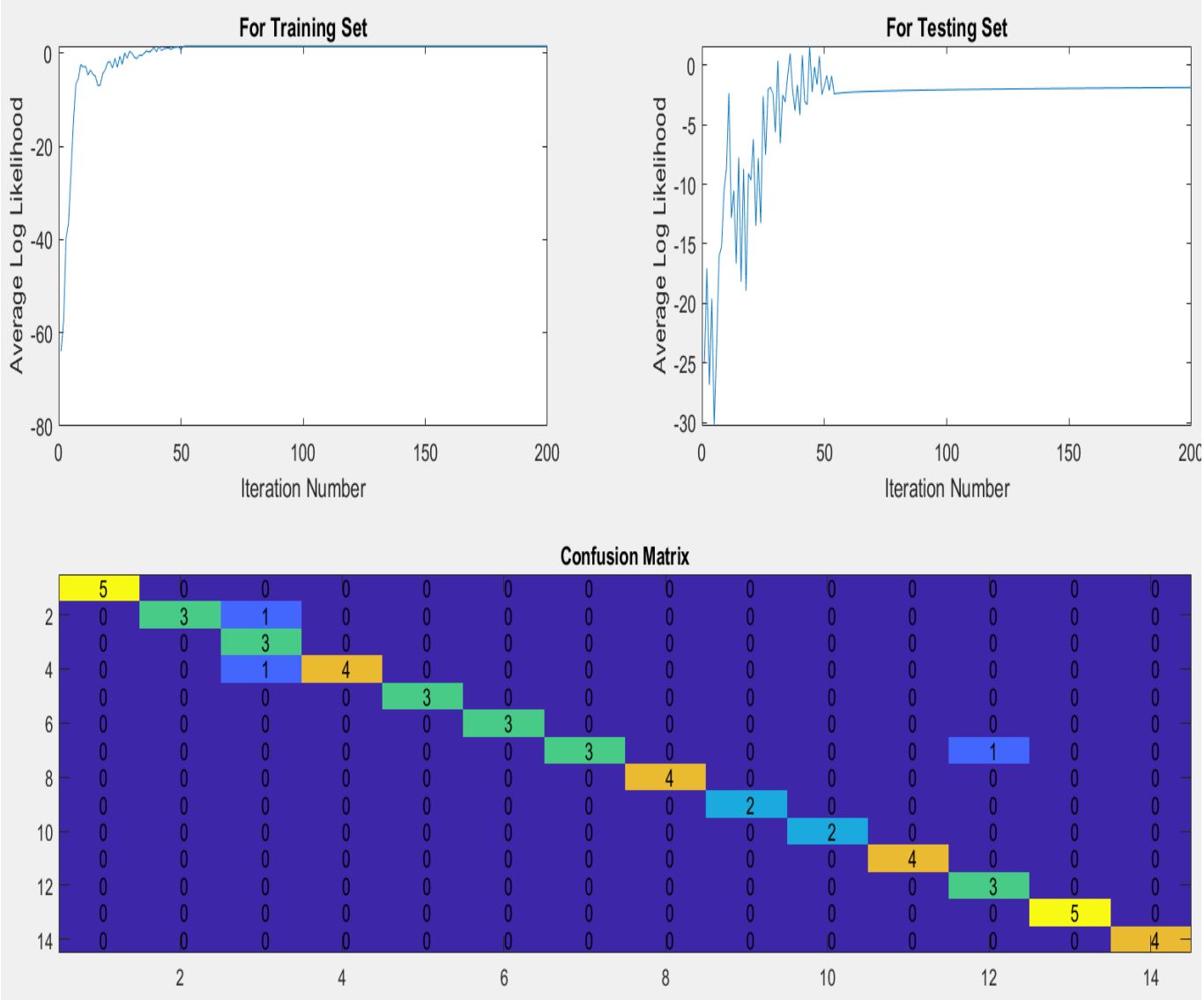


3 Gradient Descent Logistic Regression

Hyper parameters:

1) The initial values of θs are randomly chosen between [0,1].
2) The learning rate chosen after a lot of experimentation is 0.45.
3) The L2 regularization amount after a lot of experimentation is 0.001.
4) The termination criteria is either the change in average log likelihood between the iterations becomes less than 2^{-23} or the number of iterations becomes more than 200.
5) The Final Testing Accuracy is 94.12%.

The Graph for average log likelihood for the training and testing sets as a function of the training iteration number and the confusion matrix is shown below:

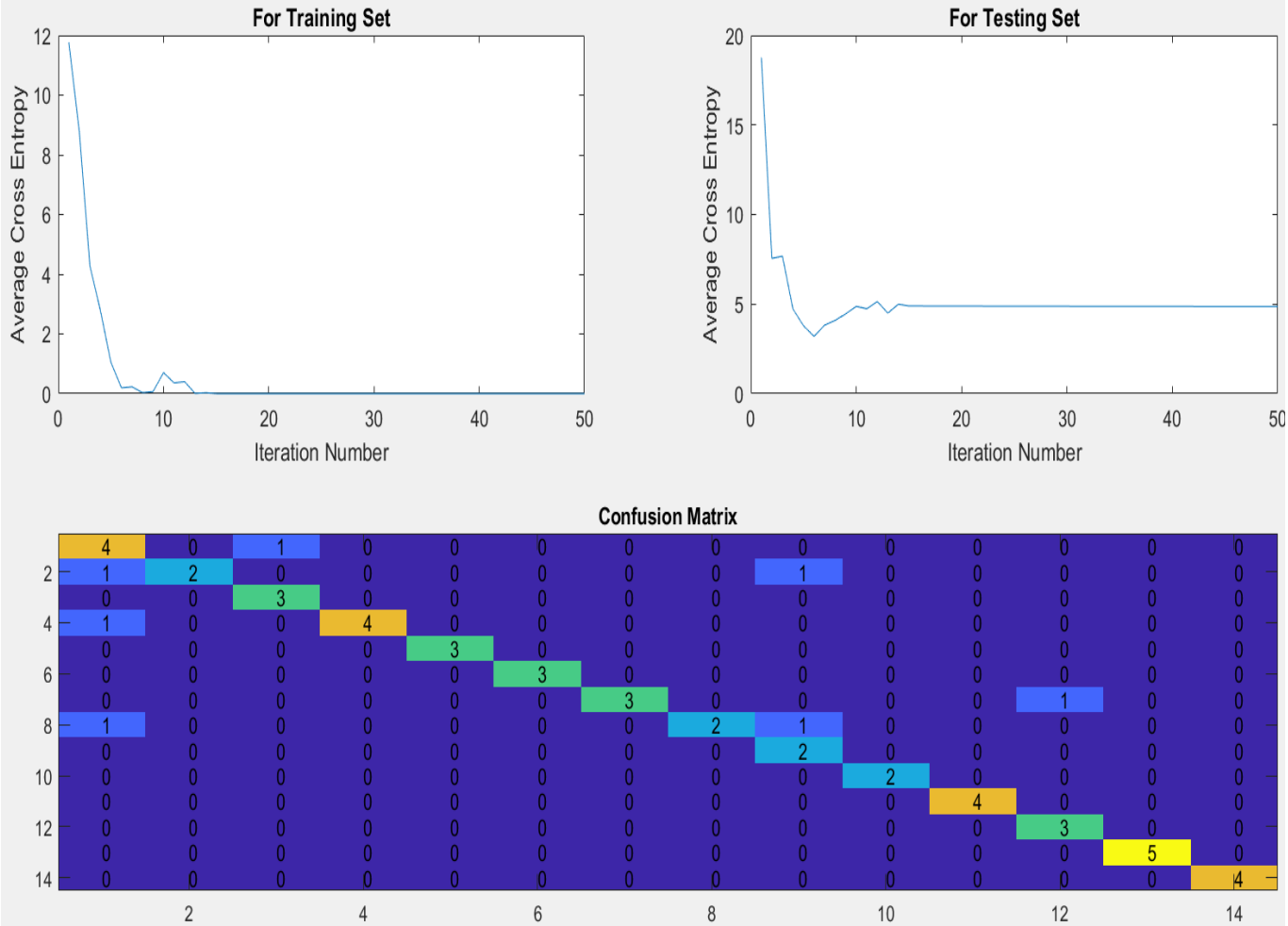


Note: The source code is in the file named "GD2.m". The above graph is also generated by the source code.

4 Gradient Descent w/ Softmax and Cross-Entropy

Hyper parameters:

1) The initial values of θs are randomly chosen between [0,1].
2) The learning rate chosen after a lot of experimentation is 0.95.
3) The L2 regularization amount after a lot of experimentation is 0.001.
4) The termination criteria is either the change in average log likelihood between the iterations becomes less than 2^{-20} or the number of iterations becomes more than 50.
5) The Final Testing Accuracy is 86.27%.
The Graph for average log likelihood for the training and testing sets as a function of the training iteration number and the confusion matrix is shown below:



Note: The source code is in the file named "GD3.m". The above graph is also generated by the source code.