Homework 3 Gretel Rajamoney

Q1: posterior of the classifier (leaving out the normalizing constant),

$$P(y=1|x_1,x_2,\dots,x_d) = \frac{P(y=1)\prod_{i=1}^{d} P(x_i|y=1)}{P(x_1,x_2,\dots,x_d)}$$

$$P(y=1|x_1,x_2,\dots x_d) = \theta_1 \prod_{i=1}^{d} \theta_{i1}^{x_i}(1-\theta_{i1})^{1-x_i} \longrightarrow \textcircled{1}$$

$$P(y=0|x_1,x_2,\dots,x_d) = \frac{P(y=0)\prod_{i=1}^{d} P(x_i|y=0)}{P(x_1,x_2,\dots,x_d)}$$

$$P(y=0|x_1,x_2,\dots,x_d) = \theta_0 \prod_{i=1}^{d} \theta_{i0}^{x_i}(1-\theta_{i0})^{1-x_i} \longrightarrow \textcircled{2}$$

therefore, we can show that:

$$\frac{P(y=1|x_1,x_2,\dots,x_d)}{P(y=0|x_1,x_2,\dots,x_d)} > 1 \Rightarrow b + \sum_{i=1}^{d} w_i x_i > 0 \qquad \text{original equation}$$

consider, $\frac{P(y=1|x_1,x_2,\dots,x_d)}{P(y=0|x_1,x_2,\dots,x_d)} > 1 \longrightarrow \textcircled{3}$

putting equations ①, ②, ③ in, we get:

$$\frac{\theta_1 \prod_{i=1}^{d} \theta_{i1}^{x_i}(1-\theta_{i1})^{1-x_i}}{\theta_0 \prod_{i=1}^{d} \theta_{i0}^{x_i}(1-\theta_{i0})^{1-x_i}} > 1 = \frac{\theta_1}{\theta_0}\cdot\frac{\prod_{i=1}^{d} \theta_{i1}^{x_i}(1-\theta_{i1})^{1-x_i}}{\prod_{i=1}^{d} \theta_{i0}^{x_i}(1-\theta_{i0})^{1-x_i}} > 1$$

$$\frac{\theta_1}{\theta_0}\cdot\frac{\prod_{i=1}^{d}\left(\frac{\theta_{i1}}{1-\theta_{i1}}\right)^{x_i}(1-\theta_{i1})}{\prod_{i=1}^{d}\left(\frac{\theta_{i0}}{1-\theta_{i0}}\right)^{x_i}(1-\theta_{i0})} > 1$$ taking the log, we get: $$\log\left(\frac{\theta_1}{\theta_0}\right) + \frac{\sum_{i=1}^{d}\log\left[\left(\frac{\theta_{i1}}{1-\theta_{i1}}\right)^{x_i}(1-\theta_{i1})\right]}{\sum_{i=1}^{d}\log\left[\left(\frac{\theta_{i0}}{1-\theta_{i0}}\right)^{x_i}(1-\theta_{i0})\right]} > 0$$

$$\log\left(\frac{\theta_1}{\theta_0}\right) + \sum_{i=1}^{d}\log\left[\left(\frac{\theta_{i1}}{1-\theta_{i1}}\right)^{x_i}(1-\theta_{i1})\right] - \sum_{i=1}^{d}\log\left[\left(\frac{\theta_{i0}}{1-\theta_{i0}}\right)^{x_i}(1-\theta_{i0})\right] > 0$$

$$\log\left(\frac{\theta_1}{\theta_0}\right) + \sum_{i=1}^{d}\log\left[\left[\frac{\left(\frac{\theta_{i1}}{1-\theta_{i1}}\right)}{\left(\frac{\theta_{i0}}{1-\theta_{i0}}\right)}\right]^{x_i}\left(\frac{1-\theta_{i1}}{1-\theta_{i0}}\right)\right] > 0 = \log\left(\frac{\theta_1}{\theta_0}\right) + \sum_{i=1}^{d}\log\left[\left[\frac{\theta_{i1}(1-\theta_{i0})}{\theta_{i0}(1-\theta_{i1})}\right]^{x_i}\left(\frac{1-\theta_{i1}}{1-\theta_{i0}}\right)\right] > 0$$

$$\log\left(\frac{\theta_1}{\theta_0}\right) + \sum_{i=1}^{d} x_i\left[\log\left(\frac{\theta_{i1}(1-\theta_{i0})}{\theta_{i0}(1-\theta_{i1})}\right) + \log\left(\frac{1-\theta_{i1}}{1-\theta_{i0}}\right)\right] > 0 \longrightarrow \textcircled{4}$$

comparing ④ with the original equation, we get:

$$b = \log\left(\frac{\theta_1}{\theta_0}\right) \quad w_i = \log\left[\frac{\theta_{i1}(1-\theta_{i0})}{\theta_{i0}(1-\theta_{i1})}\right] + \log\left(\frac{1-\theta_{i1}}{1-\theta_{i0}}\right) = \log\left[\frac{\theta_{i1}}{\theta_{i0}}\right]$$

therefore,

bias $(b) = \log\left(\frac{\theta_1}{\theta_0}\right)$ and weight $(w_i) = \log\left(\frac{\theta_{i1}}{\theta_{i0}}\right)$

Q2: $P(y=1|X_1=x_1) > P(y=0|X_1=x_1) \longrightarrow$ ⑤

$$= \frac{P(X_1=x_1|y=1)P(y=1)}{P(X_1=x_1|y=1)P(y=1)+P(X_1=x_1|y=0)P(y=0)} > \frac{P(X_1=x_1|y=0)P(y=0)}{P(X_1=x_1|y=1)P(y=1)+P(X_1=x_1|y=0)P(y=0)}$$

$P(X_1=x_1|y=1) > P(X_1=x_1|y=0) \longrightarrow$ ①

$\frac{P(X_1=x_1|y=0)}{P(X_1=x_1|y=1)} < 1 \longrightarrow$ ②

$P(y=1|X_1=x_1, X_2=x_2) > P(y=1|X_1=x_1) \longrightarrow$ ⑥

$$= \frac{P(X_1=x_1, X_2=x_2|y=1)P(y=1)}{P(X_1=x_1, X_2=x_2|y=1)P(y=1)+P(X_1=x_1, X_2=x_2|y=0)P(y=0)}$$

since $P(y=1)=P(y=0)$,

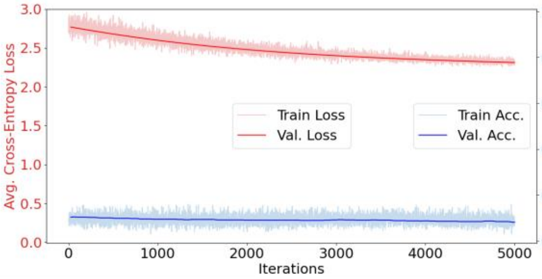
$$P(y=1|X_1=x_1, X_2=x_2) = \frac{P(X_1=x_1|y=1)^2}{P(X_1=x_1|y=1)^2+P(X_1=x_1|y=0)^2}$$

$= \frac{1}{1+\left(\frac{P(X_1=x_1|y=0)}{P(X_1=x_1|y=1)}\right)^2} \longrightarrow$ ③

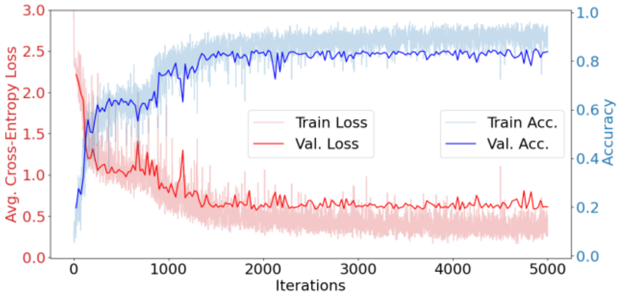
result of ⑥: $P(y=1|X_1=1) = \frac{1}{1+\left(\frac{P(X_1=x_1|y=0)}{P(X_1=x_1|y=1)}\right)} \longrightarrow$ ④

since ③ > ④ as shown in ②, we have proven ⑥

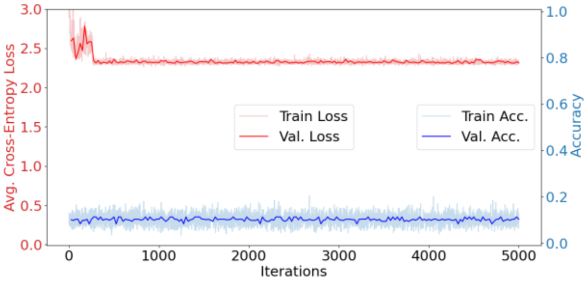
Question 4:



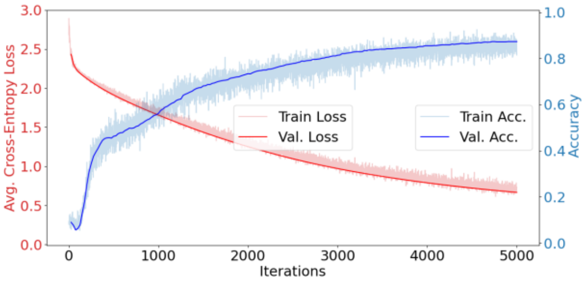
Step Size = 0.0001



Step Size = 5



Step Size = 10



Step Size = 0.01 (default)

a. The graph for the default value of step size 0.01, intersects at approximately 1000 iterations. The line representing the losses has a negative slope causing it to move towards 0 average cross-entropy loss as the number of iterations increases. Meanwhile, the line representing the accuracies has a positive slope, meaning that as the number of iterations increases, the accuracy increases and travels towards 100%.

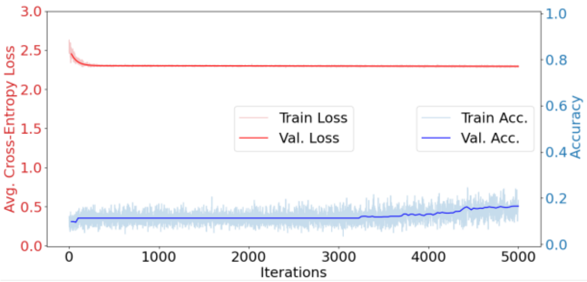
The graph for the step size of 0.0001 essentially is comprised of a line representing the losses and a line representing the accuracies which are practically parallel to each other. The line representing losses begins at a 3.0 average cross-entropy loss when the iterations are 0 and travels down to a 2.5 average cross-entropy loss when the iterations reach 5000. The line representing the accuracies stays relatively consistent at 10% throughout the duration of all of the iterations.

The graph for the step size of 5, intersects at approximately 100 iterations. The line representing the losses has a negative slope causing it to move towards 0 average cross-entropy loss as the number of iterations increases. Meanwhile, the line representing the accuracies has a positive slope, meaning that as the number of iterations increases, the accuracy increases and travels towards 100%. The overall shape of the graph is very similar to the default graph containing the step size of 0.01, the only difference is that the graph containing a step size of 0.01 is much more stable and less varied than the graph containing a step size of 5.

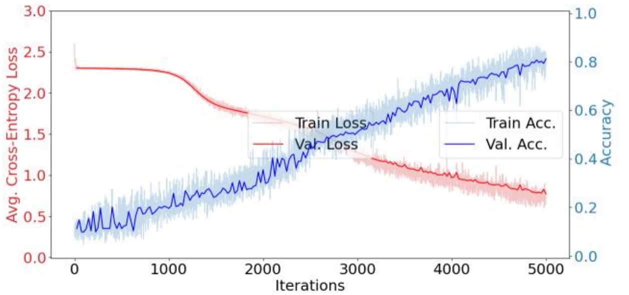
The graph containing a step size of 10 is incredibly similar to the graph containing a step size of 0.0001, this is because both the losses as well as the accuracies lines remain stagnant and parallel throughout the duration of the iterations. The losses line begins with slight fluctuations between the first 500 iterations, and then levels out to a practically flat line stationed at approximately 2.25 average cross-entropy loss. The accuracies line remains stagnant at 10% throughout the duration of all of the iterations.

b. As the number of epochs increases, the more times the weight is changed. This means that the curve transforms from underfitting the data, to best fitting the data, and then trends to overfitting the data. This means that having too many epochs will cause the model to overfit the training data, meaning that it does not learn the data, and essentially performs incredibly poorly on the test data. On the contrary, having too small of a number of epochs causes the model to poorly fit the data, resulting in the model underfitting and performing poorly on both the train data as well as the test data. Finding the perfect number of epochs results in the most optimal fit for both the test as well as the train data, since it learns the model without overfitting the training data.

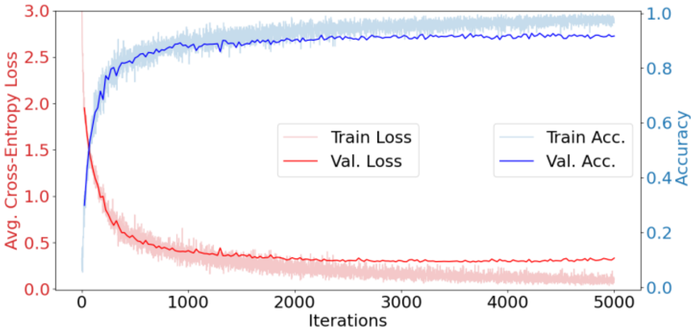
Question 5:



5 -Layers with Sigmoid Activation



5-Layers with Sigmoid Activation and Step Size 0.1



5 -Layers with ReLU Activation

a. The graph containing 5-Layers with Sigmoid Activation contains both a losses line as well as an accuracies line that are practically parallel to each other. The losses line begins at its average cross-entropy loss y-intercept at 0 iterations, at approximately 100 iterations, the line levels out to a horizontal line at approximately 2.35 average cross-entropy loss. The accuracies line begins at its origin of 0% at 0 iterations, it then gradually increases with a small positive slope as the number of iterations increases, at 5000 iterations, the accuracy falls just short of 20%.

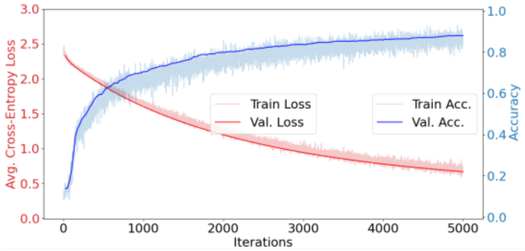
   The graph containing 5-Layers with Sigmoid Activation and Step Size 0.1 contains both a losses line as well as an accuracies line that intersect at approximately 2500 iterations. The losses line begins at its average cross-entropy loss y-intercept at 0 iterations, it steadily decreases linearly and arrives at approximately 0.7 average cross-entropy loss when the iterations reach 5000. The accuracies line begins at its origin of 0% at 0 iterations, it then increases with a positive linear slope as the number of iterations increases, at 5000 iterations the accuracy is approximately 80%.

   The graph containing 5-Layers with ReLU Activation contains both a losses line as well as an accuracies line that intersect after a few iterations. The losses line begins at a high average cross-entropy loss and then exponentially decreases until it reaches essentially a horizontal asymptote at just under 0.5 average cross-entropy loss. The accuracies line begins at its origin of 0% at 0 iterations, it then increases exponentially and then essentially arrives at a horizontal asymptote at an accuracy slightly over 90% once 5000 iterations have been completed. The graph appears to have symmetry between both the losses and the accuracies line, if a horizontal line were to be placed where the two lines intersect, both sides of the line would resemble symmetry.

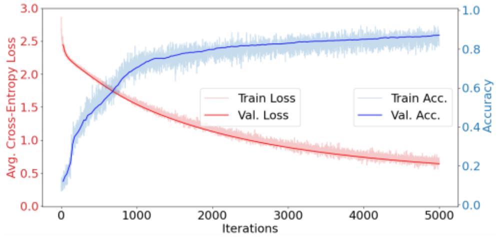
b. The learning rate improves as the step size increases from its default value of 0.01 to a new step size of 0.1. This is due to the fact that the step size controls how quickly our model is able to learn the data. Utilizing a smaller learning rate or step size requires more training epochs because the changes made to the model are very small. Meaning that it takes many more iterations in order to successfully learn the model. Utilizing a larger learning rate or step size requires less training epochs because the changes made to the model are significantly larger in comparison. Since the number of epochs remains constant between both of the experiments, the experiment containing the larger step size of 0.1 performs better in comparison to the experiment containing the smaller step size of 0.01.

c. In comparison to Sigmoid Activation, ReLU Activation computes significantly faster because of its derivative. This fast computation translates to a much simpler model. Utilizing ReLU Activation reduces the likelihood of the gradient to vanish because it utilizes a constant value as its gradient. Meanwhile, the Sigmoid Activation gradient becomes very small as the value of x increases since it isn’t a constant.

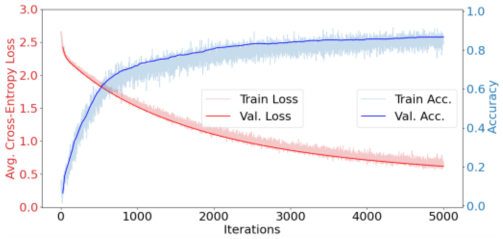
Question 6:



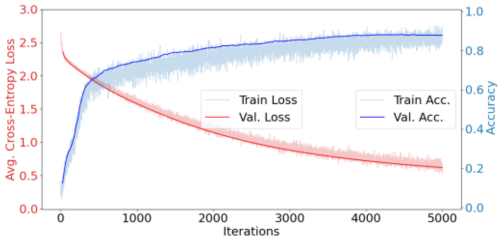
Seed = 50



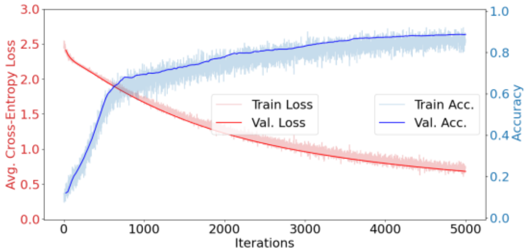
Seed = 200



Seed = 20



Seed = 300



Seed = 1

I tested my model using the default hyperparameters, while modifying the value of the seed. I utilized 5 different seed values: 1, 20, 50, 200, 300. I did not notice any trends or patterns within the graphs, as the values of the seed increases or decreases, I did not notice any consistent changes within the accuracies or other variables. Essentially one can conclude that changing the values of the seed results in small changes within the graph, but nothing of significance since it is random and occurs without any noticeable trends or patterns.

Question 7:

When creating my final Kaggle submission, I kept most hyperparameters constant. I utilized ReLU Activation as opposed to Sigmoid Activation because it was significantly faster, simpler, as well as performed much better on the data as witnessed in Question 5 of this report. I next set the value of the seed to 10 as opposed to 102. Besides modifying these two hyperparameters, all other values remained the same as the original default values.

Debriefing:

1. 12 hours
2. Easy
3. I worked on this assignment mainly alone, I did discuss my findings of Question 6 with a peer in order to ensure that the graphs my program was outputting were similar.
4. 80%
5. None