*ECE521 Assignment 1*

*Bradley Kotsopoulos – 999632982*

**Task 1:**

|  |  |
| --- | --- |
| **Training Size** | **Number of Validation Errors** |
| 5 | 107 |
| 50 | 40 |
| 100 | 27 |
| 200 | 17 |
| 400 | 9 |
| 800 | 12 |

Increasing the training set size is best, and corresponds to a decreasing validation error, meaning that the model generalizes better and performs better on unseen data, as the training set gives a better representation of the underlying distribution of the data.

**Task 2:**

|  |  |
| --- | --- |
| **K Neighbors** | **Number of Validation Errors** |
| 1 | 12 |
| 3 | 8 |
| 5 | 10 |
| 7 | 9 |
| 21 | 11 |
| 101 | 24 |
| 401 | 51 |

The optimal performance is when K equals three, and decreases as K grows large, because the set of K neighbors starts to include a significant fraction of the entire training set, and therefore many points from both classes.

**Task 3:**

****

**Task 4:**

****

Non-linearity in the model gives it the capacity to better fit a training set that is not from a linear distribution, reducing the training squared error.

**Task 5:**

The algorithm labels predictions that are >= 0.5 as 1, and predictions that are < 0.5 as 0.

Hyper-parameters:

* Epochs: 100
* Learning Rate: 0.001
* Initial weights: 0
* Mini-batch size: 1

|  |  |
| --- | --- |
| **Training Size** | **Number of Validation Errors** |
| 100 | 35 |
| 200 | 24 |
| 400 | 24 |
| 800 | 23 |

A large training size is best, as it decreases the variance, and improves the generalization of the model, as a larger training set better represents the underlying distribution of the data.

**Task 6:**

The algorithm labels predictions that are >= 0.5 as 1, and predictions that are < 0.5 as 0.

Hyper-parameters:

* Epochs: 100
* Learning Rate: 0.002
* Initial weights: 0
* Mini-batch size: 1

****

**Task 7:**

The same binary thresholding and hyper-parameters from Task 6 apply.

|  |  |
| --- | --- |
| **L2 Regularization** | **Number of Validation Errors** |
| 0 | 45 |
| 0.0001 | 45 |
| 0.001 | 45 |
| 0.01 | 45 |
| 0.1 | 52 |
| 0.5 | 82 |

In this case a higher weight decay hurts performance, and the best regularization is none at all, but this is because the weights are already initialized to zero, and forcing them towards zero actually underfits the data because the weights aren’t allowed to grow large enough to model the training distribution.