|  |  |
| --- | --- |
| Name | Oliver Witrzens |
| Student ID | U3224776 |
| Data set picked | Cancer.csv, 570 records |
| Training Method | Hidden Nodes tried: 1, 5, 20, 50, 100, 200  Eta (training constants) tried: 0.1, 0.01, 0.001 |
| Details of MLP | Hidden Nodes: 5  Epochs: 9  Eta (training constant): 0.1  Testing Accuracy: 92.55  Training Accuracy: 95.26  Validation Accuracy: 93.68  Random State Seed: 42 |
| Best estimate of validation accuracy for a generalized solution | 93.68 |
| Comments | * Data is short, clean, and relatively good. * Data requires standardization due to the inputs being quite spread in terms of scale. I used standardization to smooth outliers. * Due to the short dataset the learning rate affected the results a lot. Low learning rate (0.001) caused a more stable graph with less chance to move the perceptron weight due to the small amount of data. Middle range learning rate (0.01) showed stability in accuracy but an overall lower accuracy. * High learning rate (0.1) caused overtraining to occur quickly, however showed the highest accuracy consistently with many different random states. See the output confusion matrix below:   Confusion Matrix:  [[ 66 6]  [ 8 108]]   * Hidden layers were relatively accurate at 5, 20, 100 however given the low complexity of the data it is preferred to have less hidden layers to improve performance. |

|  |  |
| --- | --- |
| Name | Oliver Witrzens |
| Student ID | U3224776 |
| Data set picked | Card.csv, 280k records |
| Training Method | Hidden Nodes tried: 1, 20, 50, 100, 200  Eta (training constants) tried: 0.1, 0.01, 0.001 |
| Details of MLP | Hidden Nodes: 5  Epochs: 14  Eta (training constant): 0.1  Testing Accuracy: 99.94  Training Accuracy: 99.93  Validation Accuracy: 99.94  Random State Seed: 42 |
| Best estimate of validation accuracy for a generalized solution | 99.94 |
| Comments | * Data is long, very unbalanced. Relatively clean. Length of the dataset causes poor performance when training neural networks. Data requires standardization due to the inputs being quite spread in terms of scale. I used standardization to reduce the smooth outliers. * Due to the length and unbalanced true outputs in the dataset, accuracy is always high with training parameters, therefore the best parameters are those that improve performance. * Learning rate and hidden nodes cause overtraining with large epoch value, hence relatively low epochs have been selected for the final MLP. |

|  |  |
| --- | --- |
| Name | Oliver Witrzens |
| Student ID | U3224776 |
| Data set picked | Abalone.csv |
| Training Method | Hidden Nodes tried: 1,5, 20, 50, 100, 200  Eta (training constants) tried: 0.1, 0.01, 0.001 |
| Details of MLP | Hidden Nodes: 5  Epochs: 250  Eta (training constant): 0.01  Training set accuracy: 66.02  Validation set accuracy: 64.83  Test set accuracy: 66.42  Random State Seed: 42 |
| Best estimate of validation accuracy for a generalized solution | 64.83 |
| Comments about this data set  what did you learn about it if  anything | With this dataset, hidden nodes above 100 caused very quick overtraining of the neural network. Therefore, a lower value for hidden nodes causes more stable and accurate training.  More data allows for better overall training of the neural network, with support for slightly higher hidden nodes and learning rate compared to the abalone baby dataset. The accuracy is also slightly better and more stable. See below for the accuracy/epoch graph of the chosen MLP details (as listed above): |

|  |  |
| --- | --- |
| Name | Oliver Witrzens |
| Student ID | U3224776 |
| Data set picked | Abalonebaby.csv |
| Training Method | Hidden Nodes tried: 1, 20, 50, 100, 200  Eta (training constants) tried: 0.1, 0.01, 0.001 |
| Details of MLP | Hidden Nodes: 1  Epochs: 100  Eta (training constant): 0.1  Training set accuracy: 64.73  Validation set accuracy: 63.30  Test set accuracy: 62.17  Random State Seed: 42 |
| Best estimate of validation accuracy for a generalized solution | 63.30 |
| Comments about this data set  what did you learn about it if  anything | The smaller dataset causes it to become overtrained much faster, as the neural network trains it attempts to classify the training data with 100% accuracy. Compared to the full abalone dataset, the baby dataset overtrains faster as it has less data to train with each epoch. It tries to fit each of the training datasets to be exact which causes larger error, and lower validation accuracy.  Lower learning rate also causes more overfitting, with a LR of 0.1 causing the most stable accuracy/epoch graph. See below (all other variables are equal to those listed above):    Compared to 0.001 learning rate: |