Lab 5

* 1. With a variance of 3 the distributions have a lot more overlaps making it harder to separate them.
  2. Distributions with greater differences in means and smaller variances will be more separable.
  3. T
  4. The error in terms of the log-likelihood is 11.080690120936744. The confusion matrix is

|  |  |  |
| --- | --- | --- |
|  | Actual Class | |
| Predicted class | Distribution 1 | Distribution 2 |
| 5 | 3 |
| 15 | 17 |

A black screen with white text

Description automatically generated

* 1. θ0 = θ0 –

θ1 = θ1 –

θ2 = θ2 –

* 1. Using the data point (0.04081759, -2. 27909233) from distribution 1 and point (-1.39721006,0.62089447) from distribution 2 and the theta above.

θ0 = 0.09390472 –

θ1 = θ1 –

θ2 = θ2 –

* 1. After one cycle of the datapoints the log-likelihood error decreased from 7.6506380294270775 to 7.6506380294270775.

The confusion matrix stayed the same with the form of

|  |  |  |
| --- | --- | --- |
|  | Actual Class | |
| Predicted class | Distribution 1 | Distribution 2 |
| 15 | 0 |
| 5 | 20 |

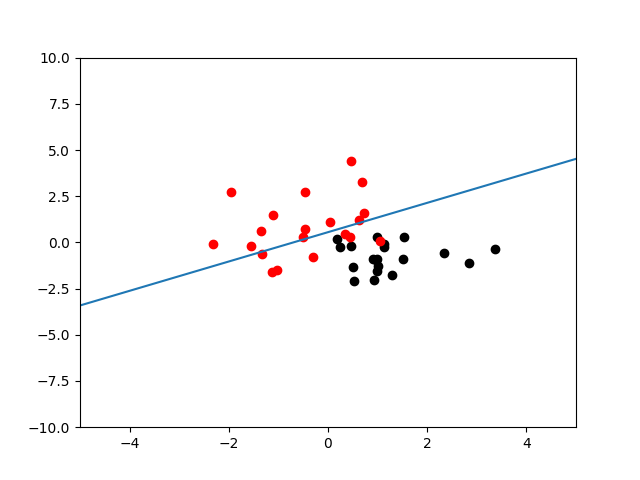
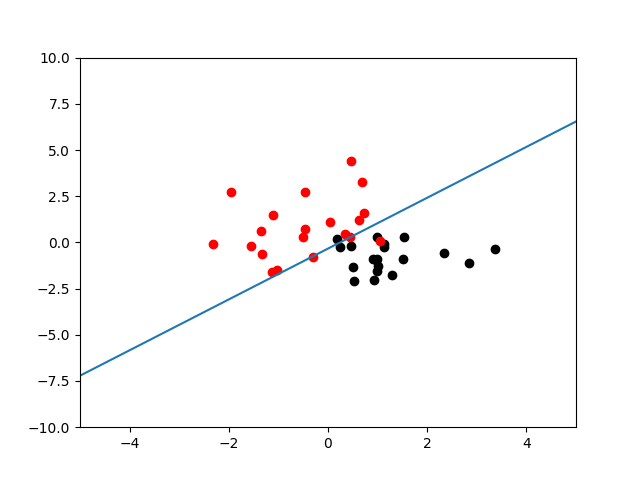
A graph with red and black dots

Description automatically generatedA graph with red and black dots

Description automatically generated

* 1. After plotting the decision boundary I see that the boundary classifies the distribution points better. More of distribution 1 point are on one side of the boundary and the other side has more of distribution 2’s points.

The error on the training data dropped from 10.048965507943208 to 4.426066011300152.



* 1. The error on the training data after training was 2.3427473431252577 with a confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | Actual Class | |
| Predicted class | Distribution 1 | Distribution 2 |
| 20 | 0 |
| 0 | 20 |

The error on the validation data after training was 3.553943501004933 with a confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | Actual Class | |
| Predicted class | Distribution 1 | Distribution 2 |
| 18 | 0 |
| 2 | 19 |

* 1. The hyperparameter ε determines the point we consider the gradient descent algorithm to have converged to a set of parameters. Increasing this value will reduce the iterations of the algorithm and a too small value may never converge or overfit.

The hyperparameter α is the learning rate of the algorithm. This determines the size of the step towards the best or most optimal set of parameters.