Portfolio assignment 2

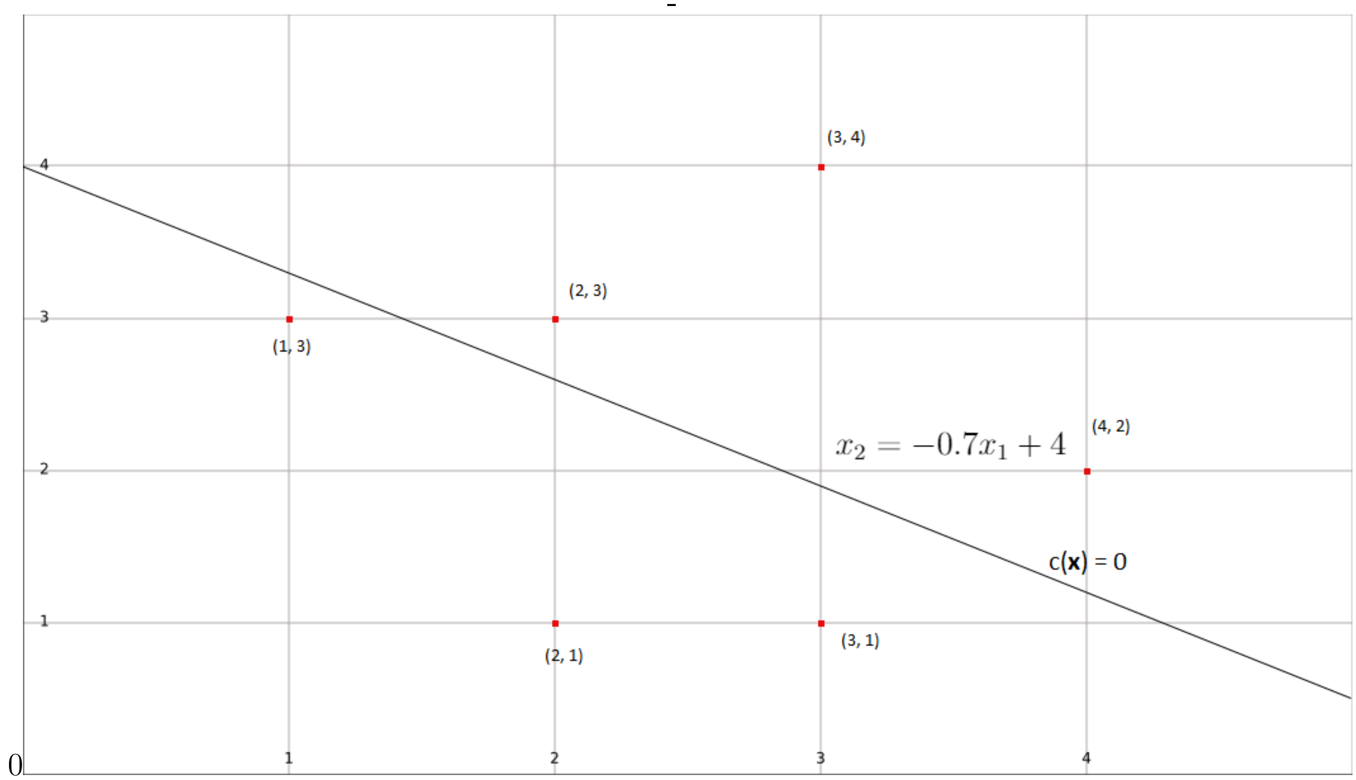
Candidate25

Problem 1

(1a)

Figure 1: Decision boudary of a logistic descrimination classifier

-



In the plot above we see the points from the training set: $\{(\mathbf{x}^i, y^i)\}_{i=1}^{6}$ as red points. $\mathbf{x}^i$ is the trainig data, and y^i is the ground truth. The line going somewhat diagonally across the plot is the decision boundary. This is used to classify the points above the line as $y^i = 0$, and points underneath the decision boundary as $y^i = 1$. This decision boundary is given by the equation:

$$C(\mathbf{x}) = \boldsymbol{w}^T \cdot \mathbf{x} + w_0 \tag{1}$$

Which put simply, assigns a class to an vector $\mathbf{x}^i$. If the vector after being put into $C(\mathbf{x}) > 0 \implies \mathbf{x} \rightarrow C^1$ or the opposite case where $C(\mathbf{x}) < 0 \implies \mathbf{x} \rightarrow C^2$. Which means at

$C(\mathbf{x}) = 0$ we are on the decision boundary. From this we can derive the equation of the decision boundary, $x_2 = -\frac{w_1}{w_2} \cdot x_1 - \frac{w_0}{w_2}$. In my guess for the decision boundary i used $w_1 = -0.7, w_2 = 1$ and $w_0 = 4$ to give the decision boundary of $x_2 = -0.7 \cdot x_1 + 4$. Here we see that the weights, $\boldsymbol{w}$ of the classifier determines the slope of the decision boundary, whereas the bias, w_0 determines the intercept with the second axis, x_2.

To figure out the distance between the origin, $\boldsymbol{O}$, and the decision boundary we can calulate this with the formula,

$$d = \frac{w_0}{\|\boldsymbol{w}\|} = \frac{4}{\sqrt{(-0.7)^2 + 1^2}} \approx \underline{\underline{3,277}}$$

Now i shall demonstrate how this decision boundary could give us a decision rule, based on the estimated weights and biases. Consider a test point,

$$\mathbf{x}^t = \begin{bmatrix} x_1^t \\ x_2^t \end{bmatrix}$$

this point will from what I stated earlier follow the decision rule,

$$\begin{cases} C(\mathbf{x^t}) > 0, & y^t = 0 \\ C(\mathbf{x^t}) \leq 0, & y^t = 1 \end{cases}$$

Here i have also made the assumption that if $C(\mathbf{x^t}) = 0$ then $\mathbf{x^t}$ is assigned to $y^t = 1$, althoug in reality this is quite rare.

(1b)

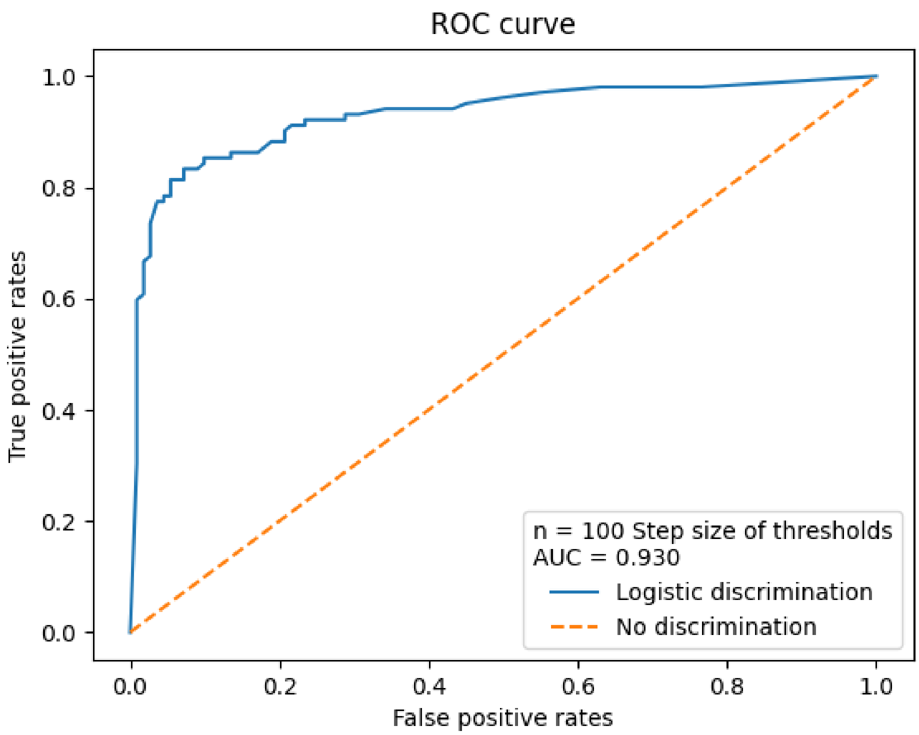
After having trained my logistic discrimination model on the *seals_train.csv* file, and the used the trained weights to classify the test data. I produce with the *seals_test.csv* file the confusion matrix:

$$\text{confusion matrix} = \begin{bmatrix} 87 & 15 \\ 11 & 100 \end{bmatrix}, \text{ and accuracy} \approx \underline{0.878}$$

Which means the model classified 87 class 0's correctly, 11 class 1's as class 0's, 15 class 0's as class 1's and 100 class 1's correctly.

(1c)

Figure 2: The Reciever Operating Characteristic



The ROC curve represents

For the Area Under the Curve score i got AUC $\approx$ 0.930 by using *sklearn.metrics* built-in function for calculating the AUC score.