Computational Intelligence - Homework 1

Mohammad Bahrami - 9724133

April 12, 2022

---

1 Question 1

Adding quadratic features to the linear regression problem that we currently have in hand, gives us the predictor function $h(z, q) = \mathrm{s}(\theta q + w^T z + a)$ where $q = x_1^2 + x_2^2$ and θ is the added parameter as the result of adding the new feature. The final expanded equation of our model is shown in equation 1.

$$\begin{aligned} \mathrm{h}(x_1, x_2) &= \theta \cdot (x_1^2 + x_2^2) + w_1 x_1 + w_2 x_2 + a \\ &= \theta x_1^2 + \theta x_2^2 + w_1 x_1 + w_2 x_2 + a \end{aligned} \tag{1}$$

1.1 A

If the introduced quadratic feature does not improve the model in any ways, the model can decide to put $\theta = 0$. hence, creating the same linear regression solution with $h(z) = \mathrm{s}(w^T z + a)$ again.

1.2 B

Independent of the values of w vector, if the θ is not equal to 0 , $h(z)$ will be the equation of a circle.

1.3 C

Because the coefficient of both x_1^2 and x_2^2 is θ, their coefficient is always equal to each other, this causes the equation to always be circle and never be an ellipse.

1.4 D

There is no S shape curve that can be created with a equation of degree two. Therefore, this option can not be created with the added features.

---

2 Question 2

One node of an neural network has the power of a linear classifier, hence the data must be separable with a line for one node to be able to model the data. As depicted in figure 1, this data can be separated by numerous lines, one of them is shown in the sme figure.

Figure 2 shows the neural node of a linear decision boundary for the f function, considering the activation function to be the *sigmoid* function. 1.

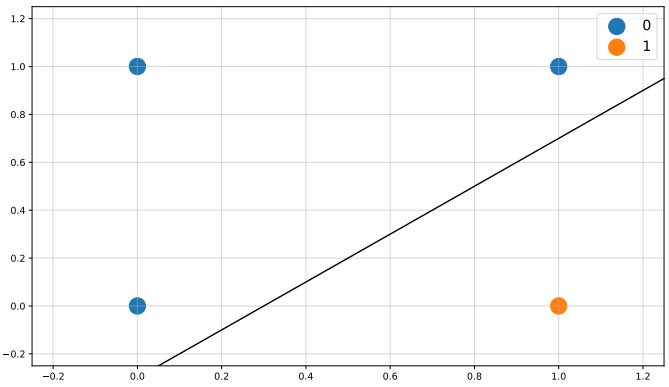

Figure 1: Visualizing the data and one of the linear decision boundaries.

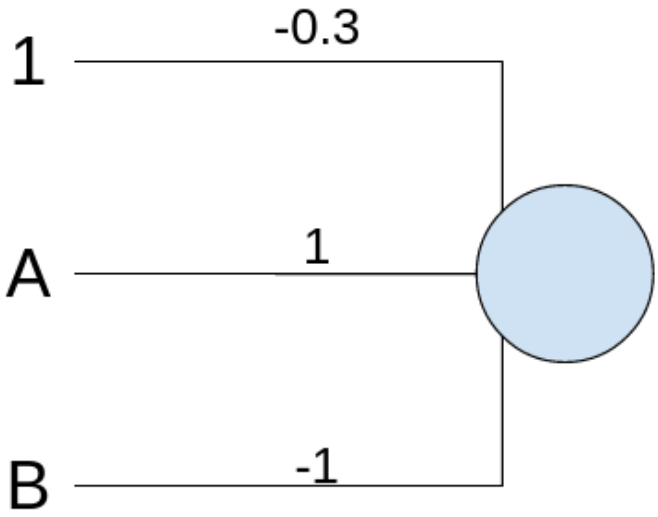

Figure 2: The neural node representing the decision boundary for the f function.

---

3 Question 3

$$X = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{ccc} x_0 & x_1 & x_2 \end{array} \\ \begin{bmatrix} -1 & 1 & 2 \\ -1 & 2 & 1 \\ -1 & 1 & 1 \\ -1 & 1 & 0 \end{bmatrix} \end{array} \qquad Y = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{c} Class \end{array} \\ \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \end{array} \qquad W = \begin{array}{c} \\ w_0 \\ w_1 \\ w_2 \end{array} \begin{array}{c} \\ \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} \end{array}$$

3.1 A

As shown in the figure 3, this data is linearly separable with infinite number of lines.For instance, one of them is drawn in black.

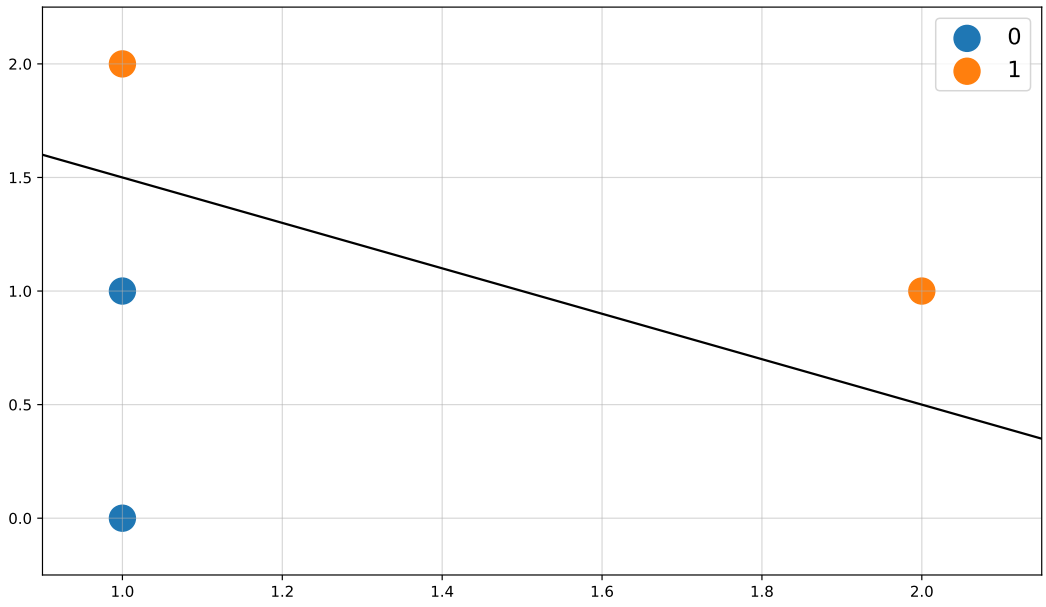

Figure 3: Visualizing the data and one of the linear decision boundaries.

3.2 B

The equation of a single-layer MLP with a sigmoid activator is:

$$h(X) = \text{sigmoid}(XW)$$

where the bias term is included in X as x_0. If we consider the loss function to be the cross entropy loss, we will have:

$$\mathcal{L}_W(y, \hat{y}) = -(y \log \hat{y} + (1 - y) \log (1 - \hat{y})) \tag{2}$$

3.2.1 Point A

$$X_A = \begin{matrix} x_0 \\ x_1 \\ x_2 \end{matrix} \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix} \qquad Y_A = 1 \qquad W = \begin{matrix} w_0 \\ w_1 \\ w_2 \end{matrix} \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$$

$$\begin{aligned} h(X_A) &= \text{sigmoid}(W^T X_A) \\ &= \text{sigmoid}(\begin{bmatrix} 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix}) \\ &= \text{sigmoid}(0 - 1 + 2) \\ &= \text{sigmoid}(1) \approx 0.731 = \hat{y} \end{aligned}$$

considering equation 2 as the loss function, we calculate the loss for point A.

$$\begin{aligned} \mathcal{L}_W(y_A, \hat{y_A}) &= -(1 \cdot \log 0.731 + (1 - 1) log(1 - 0.731)) \\ &= -log 0.731 \approx 0.313 \end{aligned}$$

To perform a gradient descent step, we need to calculate the derivative of the loss function with respect to every parameter in the model.

$$w_j = w_j - \alpha \frac{\partial \mathcal{L}_W(y, \hat{y})}{\partial w_j}$$

For our case, using cross-entropy loss defined in equation 2, the derivatives are calculated using the following equation:

$$\begin{aligned} \frac{\partial \mathcal{L}_W(y, \hat{y})}{\partial w_j} &= \frac{\partial \mathcal{L}_W(y, \hat{y})}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial W^T X} \times \frac{\partial W^T X}{\partial w_j} \\ &= \frac{\hat{y} - y}{\hat{y}(1 - \hat{y})} \times \hat{y}(1 - \hat{y}) \times x_j \end{aligned}$$

Hence, the derivative vector will be:

$$\begin{aligned} \frac{\partial \mathcal{L}_W(y, \hat{y})}{\partial W} &= (\hat{y} - y) \cdot \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} \\ &= (0.731 - 1) \cdot \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix} \\ &= \begin{bmatrix} 0.269 \\ -0.269 \\ -0.538 \end{bmatrix} \end{aligned}$$

Performing the gradient descent step gives us:

$$W = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} 0.269 \\ -0.269 \\ -0.538 \end{bmatrix} = \begin{bmatrix} -0.026 \\ -0.974 \\ 1.053 \end{bmatrix}$$

3.2.2 Point B

$$X_B = \begin{matrix} x_0 \\ x_1 \\ x_2 \end{matrix} \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix} \qquad Y_B = 1 \qquad W = \begin{matrix} w_0 \\ w_1 \\ w_2 \end{matrix} \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$$

$$\begin{aligned} h(X_B) &= \text{sigmoid}(w^T X_B) \\ &= \text{sigmoid}(\begin{bmatrix} 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}) \\ &= \text{sigmoid}(0 - 2 + 1) \\ &= \text{sigmoid}(-1) \approx 0.268 = \hat{y} \end{aligned}$$

Using the obtained derivatives in the previous part, we can write:

$$\begin{aligned} \frac{\partial \mathcal{L}_W(y, \hat{y})}{\partial W} &= (\hat{y} - y) \cdot \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} \\ &= (0.268 - 1) \cdot \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 0.732 \\ -1.464 \\ -0.732 \end{bmatrix} \end{aligned}$$

Performing the gradient descent step gives us:

$$W = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} 0.732 \\ -1.464 \\ -0.732 \end{bmatrix} = \begin{bmatrix} -0.073 \\ -0.854 \\ 1.073 \end{bmatrix}$$

3.3 C

The equation of the linear regression model is:

$$h(X) = XW$$

where the bias term is included in X as x_0. If we consider the loss function to be the mean square error, we will have:

$$\mathcal{L}_W(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2 \tag{3}$$

3.3.1 Point A

$$X_A = \begin{matrix} x_0 \\ x_1 \\ x_2 \end{matrix} \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix} \qquad Y_A = 1 \qquad W = \begin{matrix} w_0 \\ w_1 \\ w_2 \end{matrix} \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$$

$$\begin{aligned} h(X_A) &= W^T X_A \\ &= \begin{bmatrix} 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix} \\ &= 0 - 1 + 2 \\ &= 1 = \hat{y} \end{aligned}$$

using the MSE loss function defined in the equation 3, we calculate the partial derivatives of the loss function with respect to each parameter.

$$\begin{aligned} \frac{\partial \mathcal{L}_W(y, \hat{y})}{\partial w_j} &= \frac{\partial \mathcal{L}_W(y, \hat{y})}{W^T X_B} \times \frac{W^T X_B}{w_j} \\ &= (\hat{y} - y) \times x_j \end{aligned}$$

Hence, the derivative vector will be:

$$\begin{aligned} \frac{\partial \mathcal{L}_W(y, \hat{y})}{\partial w_j} &= (\hat{y} - y) \cdot \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix} \\ &= (1 - 1) \cdot \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix} \\ &= 0 \end{aligned}$$

Therefore, this data point will not change the parameters as the loss and its derivative with respect to all of the parameters is zero.

3.3.2 Point B

$$X_B = \begin{matrix} x_0 \\ x_1 \\ x_2 \end{matrix} \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix} \qquad Y_B = 1 \qquad W = \begin{matrix} w_0 \\ w_1 \\ w_2 \end{matrix} \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$$

$$h(X_A) = W^T X_A$$

$$
\begin{aligned}
&= \begin{bmatrix} 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix} \\
&= 0 - 2 + 1 \\
&= -1 = \hat{y}
\end{aligned}
$$

Using the obtained derivative formula in the previous part, we calculate the derivative vector with respect to each parameter.

$$
\begin{aligned}
\frac{\partial \mathcal{L}_W(y, \hat{y})}{\partial w_j} &= (\hat{y} - y) \cdot \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix} \\
&= (-1 - 1) \cdot \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} 2 \\ -4 \\ -2 \end{bmatrix}
\end{aligned}
$$

Now, we can perform a gradient descent step.

$$
\begin{aligned}
W &= \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} 2 \\ -4 \\ -2 \end{bmatrix} \\
&= \begin{bmatrix} -0.2 \\ -0.6 \\ 1.2 \end{bmatrix}
\end{aligned}
$$

---

4 Question 4

4.1 A

- W_1: $(D_{a_1} \times D_x)$
- b_1: $(D_{a_1} \times 1)$
- W_2: $(1 \times D_{a_1})$
- b_2: (1×1)
- X: $(D_x \times m)$
- Y: $(1 \times m)$

4.2 B

$$
\begin{aligned}
\frac{\partial J}{\partial z_3} &= \frac{\partial J}{\partial L} \times \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_3} \\
&= -\frac{1}{m} \times \frac{y - \hat{y}}{\hat{y}(1 - \hat{y})} \times \hat{y}(1 - \hat{y})
\end{aligned}
$$

$$= \frac{1}{m}(\hat{y} - y)$$

$$\begin{aligned}\frac{\partial a}{\partial z_2} &= \frac{\partial a}{\partial a_2} \times \frac{\partial a_2}{\partial z_2} \\ &= -1 \times \begin{cases} 1 & z_2 > 0 \\ 0 & z_2 < 0 \end{cases}\end{aligned}$$

$$\begin{aligned}\frac{\partial J}{\partial W_1} &= \frac{\partial J}{\partial L} \times \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_3} \times \frac{\partial z_3}{\partial a} \times (\frac{\partial a}{\partial a_1} \times \frac{\partial a_1}{\partial z_1} \times \frac{\partial z_1}{\partial W_1} + \frac{\partial a}{\partial a_2} \times \frac{\partial a_2}{\partial z_2} \times \frac{\partial z_2}{\partial W_1}) \\ &= -\frac{1}{m} \times \frac{y - \hat{y}}{\hat{y}(1 - \hat{y})} \times \hat{y}(1 - \hat{y}) \times W_2 \times (1 \times R'(z_1) \times W_1 + (-1) \times R'(z_2) \times W_1) \\ &= -\frac{1}{m} \times (y - \hat{y}) \times W_2 \times (R'(z_1) \times W_1 - R'(z_2) \times W_1)\end{aligned}$$

Where $R'(z) = \begin{cases} 1 & z > 0 \\ 0 & z < 0 \end{cases}$

4.3 C

4.3.1 W_1

$$\begin{aligned}W_1 &= W_1 - \alpha \cdot \frac{\partial J}{\partial W_1} \\ &= W_1 - \alpha \cdot \left(-\frac{1}{m} \times (y - \hat{y}) \times W_2 \times (R'(z_1) \times W_1 - R'(z_2) \times W_1) \right)\end{aligned}$$

4.3.2 b_1

$$\begin{aligned}b_1 &= b_1 - \alpha \cdot \frac{\partial J}{\partial b_1} \\ &= b_1 - \alpha \cdot \left(-\frac{1}{m} \times (y - \hat{y}) \times W_2 \times (R'(z_1) - R'(z_2)) \right)\end{aligned}$$

4.3.3 W_2

$$\begin{aligned}W_2 &= W_2 - \alpha \cdot \frac{\partial J}{\partial W_2} \\ &= W_2 - \alpha \cdot \left(-\frac{1}{m} \times (y - \hat{y}) \times W_2 \right)\end{aligned}$$

4.3.4 b_2

$$
\begin{aligned}
b_2 &= b_2 - \alpha \cdot \frac{\partial J}{\partial b_2} \\
&= b_2 - \alpha \cdot \left(-\frac{1}{m} \times (y - \hat{y}) \right)
\end{aligned}
$$