Machine Learning - Homework 3

Pedro Curvo (ist1102716) | Salvador Torpes (ist1102474)

1st Term - 23/24

Pen and Paper Exercises

1st Question

Dataset

In this exercise we aim to learn a regression model for the following dataset:

Observation	x_0	x_1	x_2	output - z
$\vec{x}_1$	1	0.7	-0.3	0.8
$\vec{x}_2$	1	0.4	0.5	0.6
$\vec{x}_3$	1	-0.2	0.8	0.3
$\vec{x}_4$	1	-0.4	0.3	0.3

Table 1: Dataset

$$X = \begin{bmatrix} 1 & 0.7 & -0.3 \\ 1 & 0.4 & 0.5 \\ 1 & -0.2 & 0.8 \\ 1 & -0.4 & 0.3 \end{bmatrix} \qquad Z = \begin{bmatrix} 0.8 \\ 0.6 \\ 0.3 \\ 0.3 \end{bmatrix}$$

$$\vec{x}_1 = \begin{bmatrix} 0.7 \\ -0.3 \end{bmatrix} \qquad \vec{x}_2 = \begin{bmatrix} 0.4 \\ 0.5 \end{bmatrix} \qquad \vec{x}_3 = \begin{bmatrix} -0.2 \\ 0.8 \end{bmatrix} \qquad \vec{x}_4 = \begin{bmatrix} -0.4 \\ 0.3 \end{bmatrix}$$

a)

Transforming the data

We are transforming our original data into a new space, according to the radial basis function:

$$\phi_j(\vec{x}) = \exp\left(-\frac{||\vec{x} - c_j||^2}{2}\right)$$

$$c_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad c_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \qquad c_3 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

After applying the transformation, we will have 3 new inputs for each observation. Therefore, the new dataset will look like:

$$\Phi(X) = X_{trans} = \begin{bmatrix} 1 & \phi_1(\vec{x}_1) & \phi_2(\vec{x}_1) & \phi_3(\vec{x}_1) \\ 1 & \phi_1(\vec{x}_2) & \phi_2(\vec{x}_2) & \phi_3(\vec{x}_2) \\ 1 & \phi_1(\vec{x}_3) & \phi_2(\vec{x}_3) & \phi_3(\vec{x}_3) \\ 1 & \phi_1(\vec{x}_4) & \phi_2(\vec{x}_4) & \phi_3(\vec{x}_4) \end{bmatrix}$$

Observation 1 If we apply our transformation to the first observation $\vec{x}_1$, we get:

$$\phi_1(\vec{x}_1) = \exp\left(-\frac{||\vec{x}_1 - c_1||^2}{2}\right) = \exp\left(-\frac{0.58}{2}\right) = 0.74826$$

$$\phi_2(\vec{x}_1) = \exp\left(-\frac{||\vec{x}_1 - c_2||^2}{2}\right) = \exp\left(-\frac{0.58}{2}\right) = 0.74826$$

$$\phi_3(\vec{x}_1) = \exp\left(-\frac{||\vec{x}_1 - c_3||^2}{2}\right) = \exp\left(-\frac{4.58}{2}\right) = 0.10127$$

Observation 2 If we apply our transformation to the second observation $\vec{x}_2$, we get:

$$\phi_1(\vec{x}_2) = \exp\left(-\frac{||\vec{x}_2 - c_1||^2}{2}\right) = \exp\left(-\frac{0.41}{2}\right) = 0.81465$$

$$\phi_2(\vec{x}_2) = \exp\left(-\frac{||\vec{x}_2 - c_2||^2}{2}\right) = \exp\left(-\frac{2.61}{2}\right) = 0.27117$$

$$\phi_3(\vec{x}_2) = \exp\left(-\frac{||\vec{x}_2 - c_3||^2}{2}\right) = \exp\left(-\frac{2.21}{2}\right) = 0.33121$$

Observation 3 If we apply our transformation to the third observation $\vec{x}_3$, we get:

$$\phi_1(\vec{x}_3) = \exp\left(-\frac{||\vec{x}_3 - c_1||^2}{2}\right) = \exp\left(-\frac{0.68}{2}\right) = 0.71177$$

$$\phi_2(\vec{x}_3) = \exp\left(-\frac{||\vec{x}_3 - c_2||^2}{2}\right) = \exp\left(-\frac{4.68}{2}\right) = 0.09633$$

$$\phi_3(\vec{x}_3) = \exp\left(-\frac{||\vec{x}_3 - c_3||^2}{2}\right) = \exp\left(-\frac{0.68}{2}\right) = 0.71177$$

Observation 4 If we apply our transformation to the fourth observation $\vec{x}_4$, we get:

$$\phi_1(\vec{x}_4) = \exp\left(-\frac{||\vec{x}_4 - c_1||^2}{2}\right) = \exp\left(-\frac{0.25}{2}\right) = 0.88250$$

$$\phi_2(\vec{x}_4) = \exp\left(-\frac{||\vec{x}_4 - c_2||^2}{2}\right) = \exp\left(-\frac{3.65}{2}\right) = 0.16122$$

$$\phi_3(\vec{x}_4) = \exp\left(-\frac{||\vec{x}_4 - c_3||^2}{2}\right) = \exp\left(-\frac{0.85}{2}\right) = 0.65377$$

Transformed Dataset

After applying the transformation, we get the following dataset:

$$\Phi(X) = X_{trans} = \begin{bmatrix} 1 & 0.74826 & 0.74826 & 0.10127 \\ 1 & 0.81465 & 0.27117 & 0.33121 \\ 1 & 0.71177 & 0.09633 & 0.71177 \\ 1 & 0.88250 & 0.16122 & 0.65377 \end{bmatrix}$$

Observation	ϕ_0	ϕ_1	ϕ_2	ϕ_3	output - z
$\vec{x}_1$	1	0.74826	0.74826	0.10127	0.8
$\vec{x}_2$	1	0.81465	0.27117	0.33121	0.6
$\vec{x}_3$	1	0.71177	0.09633	0.71177	0.3
$\vec{x}_4$	1	0.88250	0.16122	0.65377	0.3

Table 2: Transformed Dataset

Ridge Regression

A regression model is characterized by a column matrix of weights W - if we multiply W by a new observation, we get the estimated output for that observation.

$$\hat{z} = w_0 + \sum_{j=1}^{M} w_j x_j = X \cdot W$$

X is the matrix of observations, and W is the matrix of weights:

$$X = \begin{bmatrix} 1 & \vec{x}_1^T \\ 1 & \vec{x}_2^T \\ 1 & \vec{x}_3^T \\ 1 & \vec{x}_4^T \end{bmatrix} \qquad W = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix}$$

When considering the case where we **transform** our data according to a function ϕ, the regression formula is:

$$\hat{z} = w_0 + \sum_{j=1}^{M} w_j \phi_j(x) = \Phi(X) \cdot W$$

The Ridge Regression (l_2 regularization) is a method that penalizes the weights of the model, in order to avoid overfitting. The formula for W matrix in the Ridge Regression is:

$$W = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T Z$$

Where λ is the regularization parameter ($\lambda = 0.1$), I is the identity matrix and Φ is the matrix of transformed observations.

Computing the weights

Using the formula for W, we get:

$$\Phi = \begin{bmatrix} 1 & 0.74826 & 0.74826 & 0.10127 \\ 1 & 0.81465 & 0.27117 & 0.33121 \\ 1 & 0.71177 & 0.09633 & 0.71177 \\ 1 & 0.88250 & 0.16122 & 0.65377 \end{bmatrix} \quad Z = \begin{bmatrix} 0.8 \\ 0.6 \\ 0.3 \\ 0.3 \end{bmatrix} \quad \Phi^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0.74826 & 0.81465 & 0.71177 & 0.88250 \\ 0.74826 & 0.27117 & 0.09633 & 0.16122 \\ 0.10127 & 0.33121 & 0.71177 & 0.65377 \end{bmatrix}$$

$$(\Phi^T \Phi - \lambda I)^{-1} = \begin{bmatrix} 4.54826 & -3.77682 & -1.86117 & -1.86155 \\ -3.77682 & 5.98285 & -0.88543 & -1.26432 \\ -1.86117 & -0.88543 & 4.33276 & 2.72156 \\ -1.86155 & -1.26432 & 2.72156 & 4.53204 \end{bmatrix}$$

$$W = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T Z = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} 0.33914 \\ 0.19945 \\ 0.40096 \\ -0.29600 \end{bmatrix}$$

Final form of the prediction function

In order to compute $\hat{z}$, we need to multiply the weights by the transformed observation:

$$\hat{z} = \sum_{j=0}^{3} w_j \phi_j(x) = \Phi(X) \cdot W \Leftrightarrow$$

$$\Leftrightarrow \hat{z} = w_0 + w_1 \cdot \phi_1 + w_2 \cdot \phi_2 + w_3 \cdot \phi_3 = 0.33914 + 0.19945 \cdot \phi_1 + 0.40096 \cdot \phi_2 - 0.29600 \cdot \phi_3$$

Using our dataset, the predicted values are:

$$\hat{z} = \begin{bmatrix} 0.75844 \\ 0.51232 \\ 0.30905 \\ 0.38629 \end{bmatrix}$$

b)

The RMSE (Root Mean Squared Error) is a metric that measures the difference between the predicted values and the actual values. It is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(z_i - \hat{z}_i)^2}$$

Where z_i is the actual value and $\hat{z}_i$ is the predicted value. In our case, we have the following data:

z_i	$\hat{z}_i$
0.8	0.75844
0.6	0.51232
0.3	0.30905
0.3	0.38629

Table 3: Actual and Predicted Values

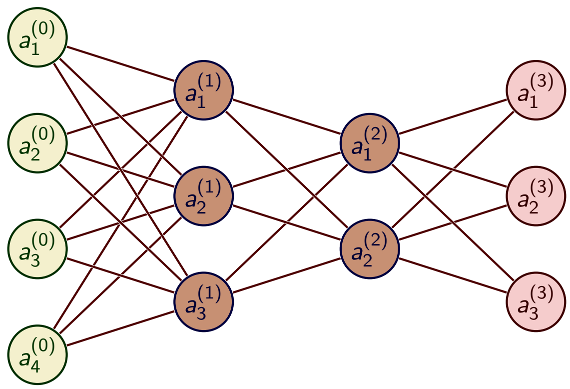
The RMSE is:

$$RMSE = \sqrt{\frac{1}{4}\sum_{i=1}^{4}(z_i - \hat{z}_i)^2} = \sqrt{\frac{1}{4} \cdot 0.01694} = 0.06508$$

2nd Question

Structure of the Network

We are considering a MLP (Multi-Layer Perceptron) with 2 hidden layers. The input and output layers each have 3 neurons. Our structure is the following:



Activation Function

The activation function is the hyperbolic tangent function and it is the same for all layers:

$$\Phi(x) = f(x) = \tanh(0.5x - 2)$$

$$\Phi'(x) = f'(x) = \frac{0.5}{\cosh^2(0.5x - 2)} = 0.5 \cdot \left(1 - \tanh^2(0.5x - 2)\right) = 0.5 \cdot \left(1 - \Phi^2(x)\right)$$

Loss Function

The loss function is the mean square error:

$$E(W) = \frac{1}{2}\sum_{i=1}^{N} ||z_i - \hat{z}_i||^2$$

Initial Weights

We are told the initial weights are:

$$w^{[1]} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \qquad w^{[2]} = \begin{bmatrix} 1 & 4 & 1 \\ 1 & 1 & 1 \end{bmatrix} \qquad w^{[3]} = \begin{bmatrix} 1 & 1 \\ 3 & 1 \\ 1 & 1 \end{bmatrix}$$

$$b^{[1]} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \qquad b^{[2]} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad b^{[3]} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

According to these weights, we can compute the initial values for $X^{[1]}$, $X^{[2]}$ and $X^{[3]}$. We are considering two training observations and therefore have two different $X^{[0]}$ vectors:

$$X_1^{[0]} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \qquad X_2^{[0]} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix}$$

With $X^{[0]}$ we can compute $X^{[1]}$, $X^{[2]}$ and $X^{[3]}$ - Propagation of both inputs through the network:

$$X_1^{[1]} = \Phi(W^{[1]} \cdot X_1^{[0]} + b^{[1]}) = \tanh\left(\left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}\right) \cdot 0.5 - 2I\right) = \tanh\left(\begin{bmatrix} 0.5 \\ 1 \\ 0.5 \end{bmatrix}\right) = \begin{bmatrix} 0.46212 \\ 0.76159 \\ 0.46212 \end{bmatrix}$$

$$Z_1^{[1]} = W^{[1]} \cdot X_1^{[0]} + b^{[1]} = \begin{bmatrix} 5 \\ 6 \\ 5 \end{bmatrix}$$

$$X_1^{[2]} = \Phi(W^{[2]} \cdot X_1^{[1]} + b^{[2]}) = \tanh\left(\left(\begin{bmatrix} 1 & 4 & 1 \\ 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0.46212 \\ 0.76159 \\ 0.46212 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) \cdot 0.5 - 2I\right) = \tanh\left(\begin{bmatrix} 0.45048 \\ -0.57642 \end{bmatrix}\right) = \begin{bmatrix} 0.45048 \\ -0.57642 \end{bmatrix}$$

$$Z_1^{[2]} = W^{[2]} \cdot X_1^{[1]} + b^{[2]} = \begin{bmatrix} 4.97061 \\ 2.68583 \end{bmatrix}$$

$$X_1^{[3]} = \Phi(W^{[3]} \cdot X_1^{[2]} + b^{[3]}) = \tanh\left(\left(\begin{bmatrix} 1 & 1 \\ 3 & 1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} -0.9159 \\ -0.80494 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}\right) \cdot 0.5 - 2I\right) = \tanh\left(\begin{bmatrix} -1.56297 \\ -1.11249 \\ -1.56297 \end{bmatrix}\right) = \begin{bmatrix} -0.9159 \\ -0.80494 \\ -0.9159 \end{bmatrix}$$

$$Z_1^{[3]} = W^{[3]} \cdot X_1^{[2]} + b^{[3]} = \begin{bmatrix} 0.87406 \\ 1.77503 \\ 0.87406 \end{bmatrix}$$

$$X_2^{[1]} = \Phi(W^{[1]} \cdot X_2^{[0]} + b^{[1]}) = \tanh\left(\left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}\right) \cdot 0.5 - 2I\right) = \tanh\left(\begin{bmatrix} -1.5 \\ -1.5 \\ -1.5 \end{bmatrix}\right) = \begin{bmatrix} -0.90515 \\ -0.90515 \\ -0.90515 \end{bmatrix}$$

$$Z_2^{[1]} = W^{[1]} \cdot X_2^{[0]} + b^{[1]} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$X_2^{[2]} = \Phi(W^{[2]} \cdot X_2^{[1]} + b^{[2]}) = \tanh\left(\left(\begin{bmatrix} 1 & 4 & 1 \\ 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} -0.90515 \\ -0.90515 \\ -0.90515 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) \cdot 0.5 - 2I\right) = \tanh\left(\begin{bmatrix} -4.21544 \\ -2.85772 \end{bmatrix}\right) = \begin{bmatrix} -0.99956 \\ -0.99343 \end{bmatrix}$$

$$Z_2^{[2]} = W^{[2]} \cdot X_2^{[1]} + b^{[2]} = \begin{bmatrix} -4.43089 \\ -1.71544 \end{bmatrix}$$

$$X_2^{[3]} = \Phi(W^{[3]} \cdot X_2^{[2]} + b^{[3]}) = \tanh\left(\left(\begin{bmatrix} 1 & 1 \\ 3 & 1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} -0.99956 \\ -0.99343 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}\right) \cdot 0.5 - 2I\right) = \tanh\left(\begin{bmatrix} -2.4965 \\ -3.49606 \\ -2.4965 \end{bmatrix}\right) = \begin{bmatrix} -0.98652 \\ -0.99816 \\ -0.98652 \end{bmatrix}$$

$$Z_2^{[3]} = W^{[3]} \cdot X_2^{[2]} + b^{[3]} = \begin{bmatrix} -0.993 \\ -2.99212 \\ -0.993 \end{bmatrix}$$

Gradient Descent

According to the gradient descent formula, in order to update the weights we need to compute the gradient of the loss function with respect to the weights. We are considering the following loss function:

$$E(W) = \frac{1}{2}||\hat{z} - z||^2 = \frac{1}{2}(\hat{z} - z)^2 = \frac{1}{2}\sum_{k=1}^{N}(\hat{z}_k - z_k)^2 = \frac{1}{2}(t - X^{[P]})^2 = \frac{1}{2}(t - X^{[P]})^T(t - X^{[P]})$$

Where $z = t$ is vector of actual output values and $\hat{z} = X^{[P]}$ (P is the index of the last layer) is the vector of predicted output values. When doing the gradient descent, we need to compute the updated weights for each layer of the network. The updated weight is equal to:

$$W_{\text{new}}^{[p]} = W^{[p]} - \eta \frac{\partial E(W)}{\partial W^{[p]}}$$

$$\frac{\partial E(W)}{\partial W^{[p]}} = \delta^{[p]} \cdot \frac{\partial Z^{[p]}}{\partial W^{[p]}} = \delta^{[P]} \cdot (X^{[P-1]})^T = \frac{\partial E}{\partial X^{[P]}} \cdot \frac{\partial X^{[P]}}{\partial Z^{[P]}} \cdot (X^{[p-1]})^T = (X^{[P]} - t) \cdot \Phi'^{[P]}(Z^{[P]}) \cdot (X^{[p-1]})^T \text{ if } p = P \text{ (output layer)}$$

$$\frac{\partial E(W)}{\partial W^{[p]}} = \delta^{[p]} \cdot \frac{\partial Z^{[p]}}{\partial W^{[p]}} = \delta^{[p]} \cdot (X^{[p-1]})^T = \Phi'^{[p]}(Z^{[p]}) \cdot (W^{[p+1]})^T \cdot \delta^{[p+1]} \cdot (X^{[p-1]})^T \text{ if } p < P \text{ (hidden layers)}$$

Computing the updated weights

Updated weights for X_1

We have $P = 3$:

$$\Delta W^{[3]} = -\eta \frac{\partial E(W)}{\partial W^{[3]}} = -\eta \cdot (X_1^{[3]} - t) \cdot \Phi'^{[3]}(Z_1^{[3]}) \cdot (X_1^{[2]})^T = -\eta \cdot (X_1^{[3]} - t) \cdot 0.5 \cdot \left(1 - \tanh^2(Z^{[3]})\right) \cdot (X_1^{[2]})^T$$