Machine Learning Homework 2

Daniel Castro 87644
João Pinto Aparício 97155

I) Polynomial Regression

a)
$$\Phi = \begin{bmatrix} 1 & 0,8 & 0,64 & 0,512 \\ 1 & 1 & 1 & 1 \\ 1 & 1,2 & 1,44 & 1.728 \\ 1 & 1,4 & 1,96 & 2.744 \\ 1 & 1,6 & 2.56 & 4.096 \end{bmatrix}$$

b) Polynomial regression

$$E(w) = \sum_{i=1}^{m} (t^{(i)} - o^{(i)})^2 = \sum_{i=1}^{m} (t^{(i)} - w^T x^{(i)})^2 = (T - Xw)^T (T - Xw)$$

$$\frac{\partial E(w)}{\partial w} = 0 \iff \frac{\partial (T-Xw)^T (T-Xw)}{\partial w} = 0 \iff$$

$$\iff \left(\frac{\partial}{\partial w}(T - Xw)^T\right) \cdot (T - Xw) + (T - Xw)^T \left(\frac{\partial}{\partial w}(T - Xw)\right) = 0 \iff$$

$$\iff (-X^T)(T - Xw) + (T - Xw)^T(-X) = 0 \iff$$

$$\iff (-X^T)(T - Xw) + (-X)^T(T - Xw) = 0 \iff$$

$$\iff -2X^T(T - Xw) = 0 \iff X^T(T - Xw) = 0 \iff$$

$$\iff X^T T - X^T X w = 0 \iff X^T X W = X^T T \iff$$

$$\iff \underline{w = (X^T X)^{-1} X^T T}$$

$$w = \left(\begin{bmatrix} 1 & 0,8 & 0,64 & 0,512 \\ 1 & 1 & 1 & 1 \\ 1 & 1,2 & 1,44 & 1.728 \\ 1 & 1,4 & 1,96 & 2.744 \\ 1 & 1,6 & 2.56 & 4.096 \end{bmatrix}^T \cdot \begin{bmatrix} 1 & 0,8 & 0,64 & 0,512 \\ 1 & 1 & 1 & 1 \\ 1 & 1,2 & 1,44 & 1.728 \\ 1 & 1,4 & 1,96 & 2.744 \\ 1 & 1,6 & 2.56 & 4.096 \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} 1 & 0,8 & 0,64 & 0,512 \\ 1 & 1 & 1 & 1 \\ 1 & 1,2 & 1,44 & 1.728 \\ 1 & 1,4 & 1,96 & 2.744 \\ 1 & 1,6 & 2.56 & 4.096 \end{bmatrix}^T \cdot \begin{bmatrix} 24 \\ 20 \\ 10 \\ 13 \\ 12 \end{bmatrix} =$$

$$= \begin{bmatrix} 47,942857 \\ -9,7619046 \\ -41,0714288 \\ 20.833333 \end{bmatrix}$$

$$Y(x, w) = 47,942857 - 9,7619046 \cdot x_1 - 41,0714288 \cdot x_2 + 20,833333 \cdot x_3$$

c) Ridge regression (l_2 regularization)

$\log(\lambda/2) = 0 \iff \lambda/2 = e^0 \iff \lambda/2 = 1 \iff \underline{\lambda = 2}$; $\|w\|_2^2 = \sum_{j \in w_2} w_j^2$

$$E(w) = \frac{1}{2} \cdot \sum_{k=1}^{M} (t_k - w^T \cdot \phi_k)^2 \cdot \frac{\lambda}{2}\|w\|_2^2 = (T - Xw)^T (T - Xw) + \lambda w^T w =$$

$$= T^T T - w^T X^T T - T^T X w + w^T X^T X w + \lambda w^T w =$$

$$= T^T T - w^T X^T \cdot T - w^T X^T T + w^T X^T X w + w^T \lambda I w =$$

$$= T^T T - 2 w^T X^T T + w^T (X^T X + \lambda I) w$$

Now, to search for the w that minimizes according to Ridge:

$$\frac{\partial E(w)}{\partial w} = -2X^T T + 2(X^T X + \lambda I) w = 0 \iff$$

$$\iff (X^T X + \lambda I) w = X^T T \implies \underline{w = (X^T X + \lambda I)^{-1} X^T T}$$

$$w = \left(\begin{bmatrix} 1 & 0,8 & 0,64 & 0,512 \\ 1 & 1 & 1 & 1 \\ 1 & 1,2 & 1,44 & 1.728 \\ 1 & 1,4 & 1,96 & 2.744 \\ 1 & 1,6 & 2.56 & 4.096 \end{bmatrix}^T \cdot \begin{bmatrix} 1 & 0,8 & 0,64 & 0,512 \\ 1 & 1 & 1 & 1 \\ 1 & 1,2 & 1,44 & 1.728 \\ 1 & 1,4 & 1,96 & 2.744 \\ 1 & 1,6 & 2.56 & 4.096 \end{bmatrix} + 2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} 1 & 0,8 & 0,64 & 0,512 \\ 1 & 1 & 1 & 1 \\ 1 & 1,2 & 1,44 & 1.728 \\ 1 & 1,4 & 1,96 & 2.744 \\ 1 & 1,6 & 2.56 & 4.096 \end{bmatrix}^T \cdot \begin{bmatrix} 24 \\ 20 \\ 10 \\ 13 \\ 12 \end{bmatrix} =$$

$$= \begin{bmatrix} 7,0450759 \\ 4,64092765 \\ 1,96734046 \\ -1,300881417 \end{bmatrix}$$

$$Y(x, w) = 7,0450759 + 4,64092765 \cdot x_1 + 1,96734046 \cdot x_2 - 1,300881417 \cdot x_3$$

d) LASSO (l_1 regularization)

$$E(w) = \frac{1}{2} \cdot \sum_{k=1}^{M} (t_k - w^T \cdot \phi_k)^2 + \lambda \cdot \|w\|_1 = (T - Xw)^T (T - Xw) + \lambda |w| =$$

$$= T^T T - w^T X^T T - T^T X w + w^T X^T X w + \lambda |w| =$$

$$= T^T T - w^T X^T \cdot T - w^T X^T T + w^T X^T X w + \lambda I |w| =$$

$$= T^T T - 2 w^T X^T T + w^T (X^T X) w + \lambda I |w|$$

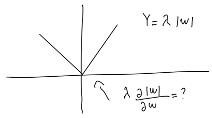
$$\frac{\partial E(w)}{\partial w} = -2X^T T + 2(X^T X) w + \frac{\partial \lambda I |w|}{\partial w} = 0$$

This is not differentiable

So we cannot use differentiation for a closed form of Lasso (l_1 regularization)

Lasso (l_1 regularization) lacks a closed form solution.

This happens because $Y = \lambda|w|$ is a non-differentiable function in $w = 0$



11) Neural Network

a) $W_1 = \begin{bmatrix} 1&1&1&1&1 \\ 0&0&0&0&0 \\ 1&1&1&1&1 \end{bmatrix}$ $b_1 = \begin{bmatrix}1\\1\\1\end{bmatrix}$ $f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$

$W_2 = \begin{bmatrix} 1&1&1 \\ 1&1&1 \end{bmatrix}$ $b_2 = \begin{bmatrix}1\\1\end{bmatrix}$ $\eta = 0.1$

$W_3 = \begin{bmatrix} 0&0 \\ 0&0 \end{bmatrix}$ $b_3 = \begin{bmatrix}0\\0\end{bmatrix}$ $f(x) = \tanh(x)$

$x_0 = \begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}$ $t = \begin{bmatrix}1\\-1\end{bmatrix}$

$$z_1 = W_1 \cdot x_0 + b_1 = \begin{bmatrix} 1&1&1&1&1 \\ 0&0&0&0&0 \\ 1&1&1&1&1 \end{bmatrix} \cdot \begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix} + \begin{bmatrix}1\\1\\1\end{bmatrix} = \begin{bmatrix}6\\1\\6\end{bmatrix}$$

$$x_1 = f(z_1) = \begin{bmatrix} f(6) \\ f(1) \\ f(6) \end{bmatrix} = \begin{bmatrix} \tanh(6) \\ \tanh(1) \\ \tanh(6) \end{bmatrix} = \begin{bmatrix} 0.999(9) \\ 0.76159 \\ 0.999(9) \end{bmatrix}$$

$$z_2 = W_2 \cdot x_1 + b_2 = \begin{bmatrix} 1&1&1 \\ 1&1&1 \end{bmatrix} \cdot \begin{bmatrix} 0.999(9) \\ 0.76159 \\ 0.999(9) \end{bmatrix} + \begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix} 3.76157 \\ 3.76157 \end{bmatrix}$$

$$x_2 = f(z_2) = \begin{bmatrix} \tanh(3{,}76157) \\ \tanh(3{,}76157) \end{bmatrix} = \begin{bmatrix} 0.99892 \\ 0.99892 \end{bmatrix}$$

$$z_3 = W_3 \cdot x_2 + b_3 = \begin{bmatrix}0\\0\end{bmatrix}$$

$$x_3 = \tanh\left(\begin{bmatrix}0\\0\end{bmatrix}\right) = \begin{bmatrix}0\\0\end{bmatrix}$$

Now, we want to do the backward phase

$E(t, x_l) = \frac{1}{2}(x_l - t)^2$ Squared error lost

In general we will need to know how to derive all functions in our network.
Lets compute them beforehand:

$\frac{\partial E}{\partial x_l}(t, x_l) = x_l - t$

$\frac{\partial x_l}{\partial z_l}(z_l) = 1 - \tanh(z_l)^2$

$\frac{\partial z_l}{\partial w_l}(w_l, b_l, x_{l-1}) = x_{l-1}$

$\frac{\partial z_l}{\partial b_l}(w_l, b_l, x_{l-1}) = 1$

$\frac{\partial z_l}{\partial x_{l-1}}(w_l, b_l, x_{l-1}) = w_l$

To start the recursion, we need the delta from the past layer.

$$\delta_3 = \frac{\partial E}{\partial x_3} \circ \frac{\partial x_3}{\partial z_3} = (x_3 - t) \circ (1 - \tanh(z_3)^2) =$$

$$= \left(\begin{bmatrix}0\\0\end{bmatrix} - \begin{bmatrix}1\\-1\end{bmatrix}\right) \circ \left(\begin{bmatrix}1\\1\end{bmatrix} - \begin{bmatrix}0\\0\end{bmatrix}\right) = \begin{bmatrix}-1\\1\end{bmatrix}$$

We can use the recursion to compute the delta from the hidden layers.

$$\delta_2 = \frac{\partial z_2}{\partial x_2}^T \cdot \delta_3 \circ \frac{\partial x_2}{\partial z_2} = (W_3)^T \cdot \delta_3 \circ (1 - \tanh(z_2)^2) =$$

$$= \begin{bmatrix} 0&0 \\ 0&0 \end{bmatrix} \cdot \begin{bmatrix}-1\\1\end{bmatrix} \circ \left(\begin{bmatrix}1\\1\end{bmatrix} - \begin{bmatrix} 0.99892^2 \\ 0.99892^2 \end{bmatrix}\right) = \begin{bmatrix}0\\0\end{bmatrix}$$

$$\delta_1 = \frac{\partial z_2}{\partial x_1}^T \cdot \delta_2 \circ \frac{\partial x_1}{\partial z_1} = (W_2)^T \cdot \delta_2 \circ (1 - \tanh(z_1)^2) =$$

$$= \begin{bmatrix} 1&1 \\ 1&1 \\ 1&1 \end{bmatrix} \cdot \begin{bmatrix}0\\0\end{bmatrix} \circ \left(\begin{bmatrix}1\\1\\1\end{bmatrix} - \begin{bmatrix} 0{,}999(9)^2 \\ 0{,}76159^2 \\ 0{,}999(9)^2 \end{bmatrix}\right) = \begin{bmatrix}0\\0\\0\end{bmatrix}$$

Finaly, we can go to the last phase and preform the updates. We start with the first layer.

$$\frac{\partial E}{\partial w_1} = \delta_1 \cdot \frac{\partial z_1}{\partial w_1}^T = \delta_1 \cdot (x_0)^T = \begin{bmatrix}0\\0\\0\end{bmatrix} \cdot \begin{bmatrix}1&1&1&1&1\end{bmatrix} = \begin{bmatrix} 0&0&0&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \end{bmatrix}$$

$$w_1 = w_1 - \eta \frac{\partial E}{\partial w_1} = \begin{bmatrix} 1&1&1&1&1 \\ 0&0&0&0&0 \\ 1&1&1&1&1 \end{bmatrix} - 0.1 \begin{bmatrix} 0&0&0&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \end{bmatrix} = \begin{bmatrix} 1&1&1&1&1 \\ 0&0&0&0&0 \\ 1&1&1&1&1 \end{bmatrix}$$

$$\frac{\partial E}{\partial b_1} = \delta_1 \cdot \frac{\partial z_1}{\partial b_1}^T = \delta_1 = \begin{bmatrix}0\\0\\0\end{bmatrix}$$

$$b_1 = b_1 - \eta \frac{\partial E}{\partial b_1} = \begin{bmatrix}1\\1\\1\end{bmatrix} - 0.1 \begin{bmatrix}0\\0\\0\end{bmatrix} = \begin{bmatrix}1\\1\\1\end{bmatrix}$$

Now the second:

$$\frac{\partial E}{\partial w_2} = \begin{bmatrix} 0&0&0 \\ 0&0&0 \end{bmatrix}$$

$$w_2 = \begin{bmatrix} 1&1&1 \\ 1&1&1 \end{bmatrix}$$

$$\frac{\partial E}{\partial b_2} = \begin{bmatrix}0\\0\end{bmatrix}$$

$$b_2 = \begin{bmatrix}1\\1\end{bmatrix}$$

$$\frac{\partial E}{\partial w_3} = \begin{bmatrix}-1\\1\end{bmatrix} \cdot \begin{bmatrix} 0.99892 & 0.998892 \end{bmatrix} =$$

$$= \begin{bmatrix} -0{,}99892 & -0{,}99892 \\ 0{,}99892 & 0{,}99892 \end{bmatrix}$$

$$w_3 = \begin{bmatrix} 0&0 \\ 0&0 \end{bmatrix} - 0{,}1 \cdot \begin{bmatrix} -0{,}99892 & -0{,}99892 \\ 0{,}99892 & 0{,}99892 \end{bmatrix} =$$

$$= \begin{bmatrix} -0{,}099892 & -0{,}099892 \\ 0{,}099892 & 0{,}099892 \end{bmatrix}$$

$$\frac{\partial E}{\partial b_3} = \delta_2 = \begin{bmatrix}0\\0\end{bmatrix} \qquad b_3 = b_3 - 0{,}1 \begin{bmatrix}0\\0\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}$$

b) We cannot use Cross Entropy Error in the previous exercise because the $E_m \in [0,1]$. The formula gives us a probability distribution and probabilities $\in [0,1]$. In the previous exercise our target is $t = (1\ -1)^T$ and $-1 \notin [0,1]$, so, we cannot apply Cross Entropy Error.

c) $t = \begin{bmatrix}1\\0\end{bmatrix}$ $x_0 = \begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}$

Because we only apply Softmax in the output unit, we can reuse the computed results from exercise a) as follows:

$x_0 = \begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}$ $z_1 = \begin{bmatrix}6\\1\\6\end{bmatrix}$ $x_1 = \begin{bmatrix} 0{,}999(9) \\ 0.76159 \\ 0.999(9) \end{bmatrix}$ $z_2 = \begin{bmatrix} 3.76157 \\ 3.76157 \end{bmatrix}$

$x_2 = \begin{bmatrix} 0{,}99892 \\ 0{,}99892 \end{bmatrix}$ $z_3 = \begin{bmatrix}0\\0\end{bmatrix}$ we only use sofmax on the x_3: $x_3 = \text{Softmax}\left(\begin{bmatrix}0\\0\end{bmatrix}\right) = \begin{bmatrix}0{,}5\\0{,}5\end{bmatrix}$

Because we applied Softmax on the output unit, we now can use Cross Entropy Loss because we no longer have 0 on $x_3 = \begin{bmatrix}0{,}5\\0{,}5\end{bmatrix}$

$$E(t, x_3) = -\sum_{i=1}^{d} t_i \log x_{3i}$$

$$\delta_{3i} = x_{3i} - t_i$$

The remaining derivatives, can be computed as before like exercise a):

$\frac{\partial x_l}{\partial z_l}(z_l) = 1 - \tanh(z_l)^2$

$\frac{\partial z_l}{\partial w_l}(w_l, b_l, x_{l-1}) = x_{l-1}$

$\frac{\partial z_l}{\partial b_l}(w_l, b_l, x_{l-1}) = 1$

$\frac{\partial z_l}{\partial x_{l-1}}(w_l, b_l, x_{l-1}) = w_l$

To start the recursion, we need the delta from the last layer

$$\delta_3 = (x_3 - t) = \begin{bmatrix}0{,}5\\0{,}5\end{bmatrix} - \begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}-0{,}5\\0{,}5\end{bmatrix}$$

$$\delta_2 = \frac{\partial z_3}{\partial x_2}^T \cdot \delta_3 \circ \frac{\partial x_2}{\partial z_2} = W_3^T \cdot \delta_3 \circ (1 - \tanh(z_2)^2) = \begin{bmatrix} 0&0 \\ 0&0 \end{bmatrix} \begin{bmatrix}-0{,}5\\0{,}5\end{bmatrix} \left(\begin{bmatrix}1\\1\end{bmatrix} - \begin{bmatrix} 0{,}99892^2 \\ 0{,}99892^2 \end{bmatrix}\right) = \begin{bmatrix}0\\0\end{bmatrix}$$

$$\delta_1 = \frac{\partial z_1}{\partial x_1}^T \cdot \delta_2 \circ \frac{\delta x_1}{\partial z_1} = W_2^T \cdot \delta_2 \circ (1 - \tanh(z_1)^2) = \begin{bmatrix} 1&1 \\ 1&1 \\ 1&1 \end{bmatrix} \begin{bmatrix}0\\0\end{bmatrix} \circ \left(\begin{bmatrix}1\\1\\1\end{bmatrix} - \begin{bmatrix} 0.999(9)^2 \\ 0.76159^2 \\ 0{,}999(9)^2 \end{bmatrix}\right) = \begin{bmatrix}0\\0\\0\end{bmatrix}$$

finaly, we can preform the updates (the last phase). We will start with the first layer:

$$\frac{\partial E}{\partial w_1} = \delta_1 \cdot \frac{\partial z_1}{\partial w_1}^T = \delta_1 \cdot x_0^T = \begin{bmatrix}0\\0\\0\end{bmatrix} \cdot \begin{bmatrix}1&1&1&1&1\end{bmatrix} = \begin{bmatrix} 0&0&0&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \end{bmatrix}$$

$$w_1 = w_1 - \eta \cdot \frac{\partial E}{\partial w_1} = \begin{bmatrix} 1&1&1&1&1 \\ 0&0&0&0&0 \\ 1&1&1&1&1 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} 0&0&0&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \end{bmatrix} = \begin{bmatrix} 1&1&1&1&1 \\ 0&0&0&0&0 \\ 1&1&1&1&1 \end{bmatrix}$$

$$\frac{\partial E}{\partial b_1} = \delta_1 \cdot \frac{\partial z_1}{\partial b_1}^T = \delta_1 = \begin{bmatrix}0\\0\\0\end{bmatrix} \qquad b_1 = b_1 - \eta \cdot \frac{\partial E}{\partial b_1} = \begin{bmatrix}1\\1\\1\end{bmatrix} - 0{,}1 \begin{bmatrix}0\\0\\0\end{bmatrix} = \begin{bmatrix}1\\1\\1\end{bmatrix}$$

Now the second:

$$\frac{\partial E}{\partial w_2} = \delta_2 \cdot \frac{\partial z_2}{\partial w_2}^T = \delta_2 \cdot x_1^T = \begin{bmatrix}0\\0\end{bmatrix} \cdot \begin{bmatrix} 0{,}999(9) & 0.76159 & 0.999(9) \end{bmatrix} = \begin{bmatrix} 0&0&0 \\ 0&0&0 \end{bmatrix}$$

$$w_2 = w_2 - \eta \frac{\partial E}{\partial w_2} = \begin{bmatrix} 1&1&1 \\ 1&1&1 \end{bmatrix} - 0.1 \begin{bmatrix} 0&0&0 \\ 0&0&0 \end{bmatrix} = \begin{bmatrix} 1&1&1 \\ 1&1&1 \end{bmatrix}$$

$$\frac{\delta E}{\partial b_2} = \begin{bmatrix}0\\0\end{bmatrix} \qquad b_2 = b_2 - \eta \frac{\partial E}{\partial b_2} = \begin{bmatrix}1\\1\end{bmatrix}$$

All that is left is to update the parameters for the output layer:

$$\frac{\partial E}{\partial w_3} = \delta_3 \cdot \frac{\partial z_3}{\partial w_3}^T = \delta_3 \cdot x_2^T = \begin{bmatrix}-0{,}5\\0{,}5\end{bmatrix} \begin{bmatrix} 0.99892 & 0.99892 \end{bmatrix} = \begin{bmatrix} -0{,}49946 & -0{,}49946 \\ 0{,}49946 & 0{,}49946 \end{bmatrix}$$

$$w_3 = w_3 - \eta \frac{\partial E}{\partial w_3} = \begin{bmatrix} 0&0 \\ 0&0 \end{bmatrix} - 0.1 \begin{bmatrix} -0{,}49946 & -0{,}49946 \\ 0{,}49946 & 0{,}49946 \end{bmatrix} = \begin{bmatrix} 0{,}49946 & 0{,}49946 \\ -0{,}49946 & -0{,}49946 \end{bmatrix}$$

$$\frac{\partial E}{\partial b_3} = \delta_3 = \begin{bmatrix}-0{,}5\\0{,}5\end{bmatrix}$$

$$b_3 = b_3 - \eta \frac{\partial E}{\partial b_3} = \begin{bmatrix}0\\0\end{bmatrix} - 0{,}1 \begin{bmatrix}-0{,}5\\0{,}5\end{bmatrix} = \begin{bmatrix}0{,}05\\-0{,}05\end{bmatrix}$$