Homework 2

Yunian Pan

October 4, 2018

1 Problem 1: Perceptron

Using GD, set the max iteration to be 1000, the the evolution of binary classification error and the perceptron error with iterations from random initialization until convergence on a successful run are shown as 1.1:

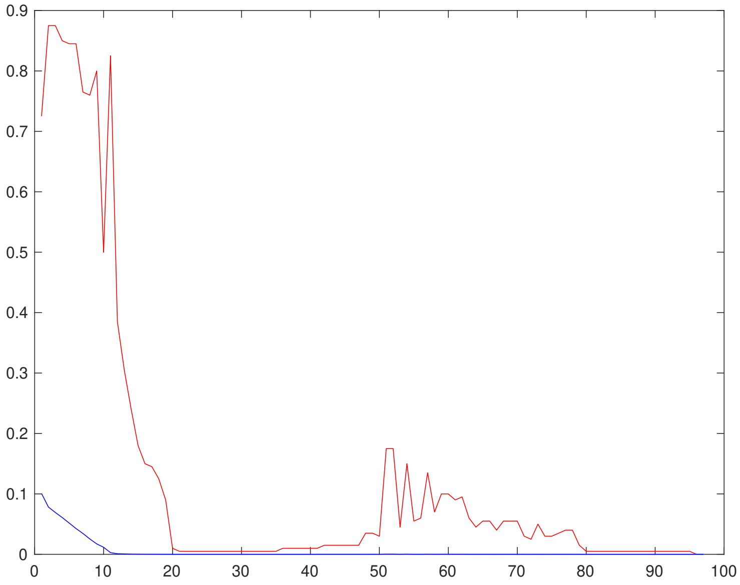

Figure 1.1: evolution of errors with iteration

In which the red represents the classification error given by $R(\theta) = \dfrac{1}{N}\sum_{i=1}^{N} step(-y_i\theta^\mathsf{T} x_i)$ and the blue represents the perceptron error given by $R^{per}(\theta) = -\dfrac{1}{N}\sum_{i\in misclassified} y_i(\theta^\mathsf{T} x_i)$. And the 2D linear boundary is shown in 1.2

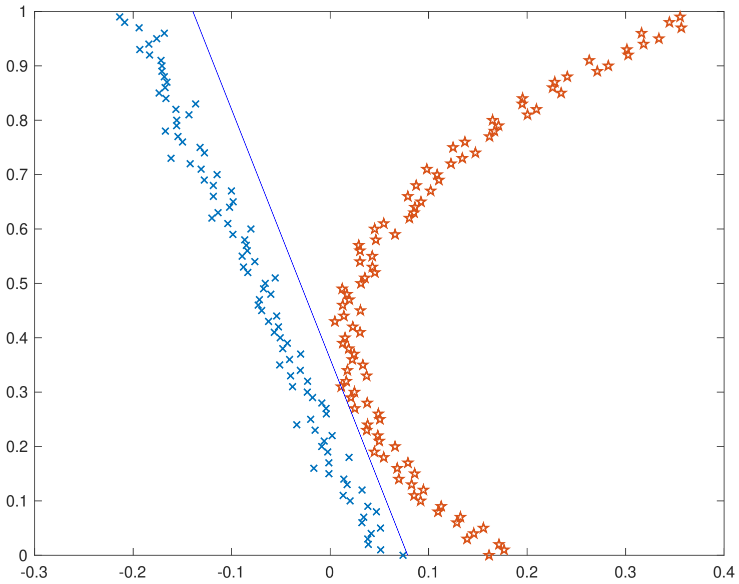

Figure 1.2: boundary

Using SGD, the figures are shown in 1.3 and 1.4, the error oscillates with time.

Discussion: when the learning rate η varies, the iteration that algorithm runs also changes, when $\eta = 1$, the max iteration varies from 60 to 140, when $\eta = 0.1$, the max iteration varies from 100 to 200, when $\eta = 0.01$, the max iteration varies from 400 to 1000 or even more, and when η increases to 10 or 100, the max iteration falls down within 100.

2 Problem 2: Neural network

2.1 a)

The three units can be converted to one, through simplification, $CW^\intercal A$ can be represented by $C[(w_2w_6 + w_1w_5)x_1 + (w_4w_6 + w_3w_5)x_2]$ that's the same as converting the network to one unit with weights vector $[w_2w_6 + w_1w_5, w_4w_6 + w_3w_5]^\intercal$, computing the same function without using any hidden units.

2.2 b)

It is possible. Since it's linear operation that is represented by every layer, take the a) for example, the process is no more than

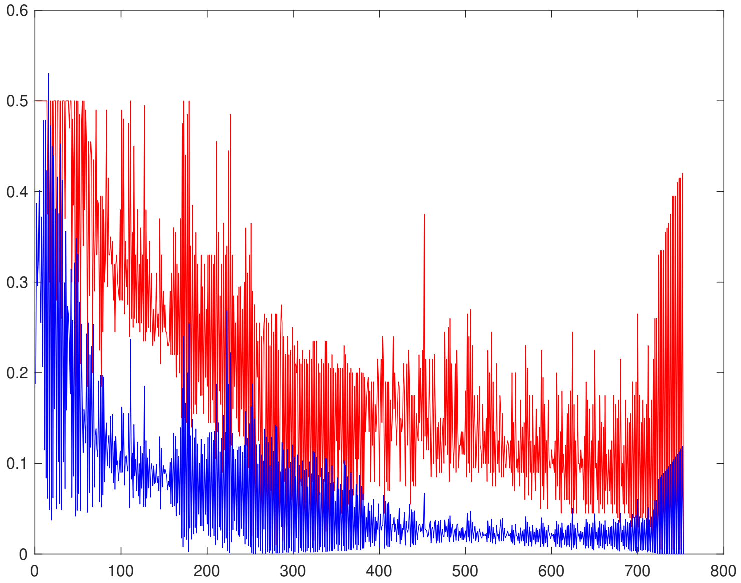

Figure 1.3: evolution of errors with time

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} w_2 & w_3 \\ w_4 & w_6 \end{bmatrix} \begin{bmatrix} w_6 \\ w_5 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} w_2 w_6 + w_1 w_5 \\ w_4 w_6 + w_3 w_5 \end{bmatrix}$$

We can compute the matrix multiplication regardless of how many layers there are, and the final results can be represented by one unit.

2.3 c)

The final form is

$$Y = t\{w_5 \left[1 + \exp(w_1 X_1 + w_3 X_2)\right]^{-1} + w_6 \left[1 + \exp(w_2 X_1 + w_4 X_2)\right]^{-1}\}.$$

It handles X_1 XOR X_2 when $W^\intercal$ satisfies

$$w_5 + w_6 < 0 \tag{1}$$
$$w_5/(1 + \exp(w_1 + w_3)) + w_6/(1 + \exp(w_2 + w_4)) < 0 \tag{2}$$
$$w_5/(1 + \exp(w_1)) + w_6/(1 + \exp(w_2)) > 0 \tag{3}$$
$$w_5/(1 + \exp(w_3)) + w_6/(1 + \exp(w_4)) > 0 \tag{4}$$

Let $w_5 = 1$, $w_6 = -2$, $w_3 = w_1$ and $w_4 = w_2$, we can easily find a couple of values that satisfy the inequality above, e.g. $w_1 = w_3 = 4$, $w_2 = w_4 = 8$.

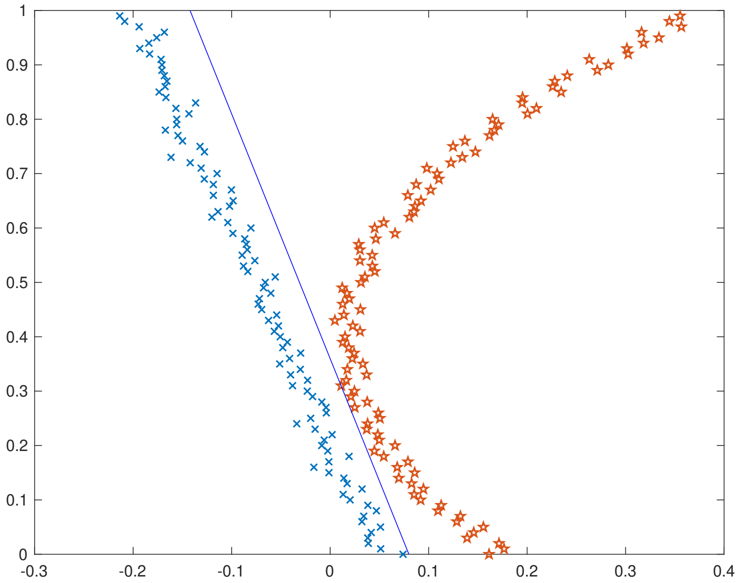

Figure 1.4: boundary

3 Problem 3: Backpropagation

3.1 case 1

Sol:

Given $x_i = g(y_j) = \dfrac{1}{1+e^{-\sum_j w_{ji}y_j}}$, and $\sum_i \dfrac{\partial E}{\partial x_i} = -\sum_i (\dfrac{t_i}{x_i} - \dfrac{1-t_i}{1-x_i})$ apply the chain rule, first take x_i and y_j as fixed to compute the gradient for w_{ji}, then take z_k and y_j as fixed to compute the gradient for w_{kj} we have

$$\begin{aligned}
\frac{\partial E}{\partial w_{ji}} &= \frac{\partial E}{\partial x_i}\frac{\partial x_i}{\partial w_{ji}} \\
&= (\frac{1-t_i}{1-x_i} - \frac{t_i}{x_i})\frac{\partial g(w_{ji}, y_j)}{\partial w_{ji}} \\
&= (\frac{1-t_i}{1-x_i} - \frac{t_i}{x_i})x_i^2 y_j e^{-\sum_j w_{ji}y_j} \\
&= (\frac{1-t_i}{1-x_i} - \frac{t_i}{x_i})x_i y_j (1-x_i) \\
&= (x_i - t_i)y_j
\end{aligned}$$

$$\begin{aligned}\frac{\partial E}{\partial w_{kj}} &= \sum_i (\frac{\partial E}{\partial x_i}\frac{\partial x_i}{\partial y_j})\frac{\partial y_j}{\partial w_{kj}} \\ &= \sum_i \frac{\partial E}{\partial x_i}(-w_{ji})x_i(1-x_i)(-z_k)y_j(1-y_j) \\ &= \sum_i \frac{x_i - x_i t_i - t_i + x_i t_i}{(1-x_i)x_i}(-(-w_{ji})x_i(1-x_i))(-(-z_k)y_j(1-y_j)) \\ &= \sum_i (x_i - t_i)w_{ji}y_j(1-y_j)z_k\end{aligned}$$

Thus, we have $\quad \sum_i \frac{\partial E}{\partial w_{ji}} = \sum_i \delta_j^i y_j, \qquad \sum_j \frac{\partial E}{\partial w_{kj}} = \sum_j \delta_k^j z_k.$

3.2 case 2

Sol:

Given the cross-entropy $E = -\sum_i t_i \log(x_i)$ and softmax activation function $x_i = \frac{e^{\sum_j w_{ji}y_j}}{\sum_i e^{\sum_j w_{ji}y_j}} = f(w_{11}, \ldots, w_{j1}, \ldots, w_{ji}, \ldots, w_{jm}, \ldots, y_j, \ldots)$, we have

$$\frac{\partial E}{\partial x_i} = -\frac{t_i}{x_i}$$

$$\begin{aligned}\frac{\partial x_m}{\partial w_{ji}} &= \frac{y_j(\delta(i-m)e^{\sum_j w_{jm}y_j}(\sum_i e^{\sum_j w_{ji}y_j}) - e^{\sum_j w_{jm}y_j}e^{\sum_j w_{ji}y_j})}{(\sum_i e^{\sum_j w_{ji}y_j})^2} \\ &= y_j x_m(\delta(i-m) - x_i)\end{aligned}$$

$$\begin{aligned}\frac{\partial E}{\partial w_{ji}} &= \sum_m \frac{\partial E}{\partial x_m}\frac{\partial x_m}{\partial w_{ji}} \\ &= \sum_m y_j(-\frac{t_m}{x_m})x_m(\delta(i-m) - x_i) \\ &= y_j(\sum_m t_m \cdot x_i - t_i)\end{aligned}$$

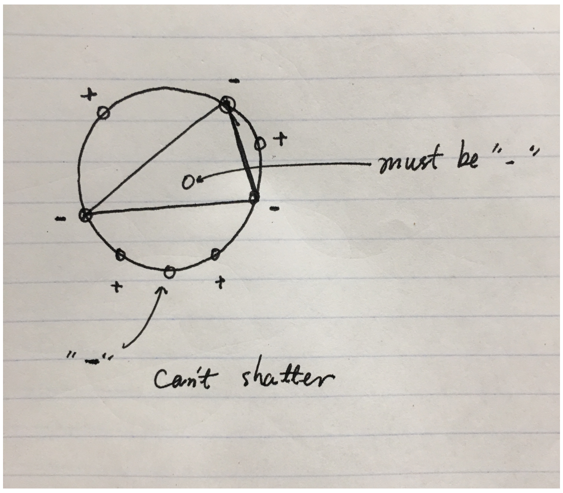

Figure 4.1: illustration

$$\begin{aligned}\frac{\partial E}{\partial w_{kj}} &= \sum_i(\sum_m(\frac{\partial E}{\partial x_m}\frac{\partial x_m}{\partial y_j})\frac{\partial y_j}{\partial w_{kj}}) \\ &= \sum_i(w_{ji}(\sum_m t_m x_i - t_i))y_j(1-y_j)z_k \\ &= \sum_i(\sum_m t_m x_i - t_i)(w_{ji}y_j(1-y_j))z_k\end{aligned}$$

Here $\delta_j^i = \sum_m t_m \cdot x_i - t_i$, $\delta_k^j = \sum_i(\sum_m t_m x_i - t_i)(w_{ji}y_j(1-y_j))$. For both cases, $w^{t+1} = w^t - \eta\frac{\partial E}{\partial w^t}$.

4 Problem 4: VC dimension

Sol:

The VC dimension is 7. we can draw a circumcircle in the 2D plane and arrange the points on the circle, for a set of 7 points labeled by $\{+1, -1\}$, the max cardinality of convex hull is $\left\lceil \frac{7}{2} \right\rceil = 3$, let the 3 points be the vertexes, which indicates that we can always find a triangle that contains the 3 points and has no intersection with other points; For a set of 8 points, the max convex hull becomes 4 points, take the labeling {+1, -1, +1, -1, +1, -1, +1, -1} which is the most complex condition for example, there is no triangle that can shatter them, because there will be one point either outside or inside the convex hull that can't be determined, as illustrated in 4.1.

Therefore, $VC(H) = 7$.