50.039 – Theory and Practice of Deep learning
Homework 2

Lin Huiqing

19 February 2021

1 Hyperplanes

A linear classifier which predicts all points correctly would be a hyperplane. One possible hyperplane could be one defined by an orthogonal vector which passes through both points in D_n. This orthogonal vector, n^T would be given by:

$$\begin{aligned} n^T &= x_1 - x_2 \\ &= \begin{bmatrix} -5 \\ 1 \\ 3 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \\ -3 \end{bmatrix} \\ &= \begin{bmatrix} -7 \\ -1 \\ 6 \end{bmatrix} \end{aligned}$$

A point between x_1 and x_2, can be given by midpoint m:

$$\begin{aligned} m &= x_2 + \frac{1}{2}(x_1 - x_2) \\ &= \begin{bmatrix} 2 \\ 2 \\ -3 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} -7 \\ -1 \\ 6 \end{bmatrix} \\ &= \begin{bmatrix} -1.5 \\ 1.5 \\ 0 \end{bmatrix} \end{aligned}$$

A possible hyperplane, with orthogonal vector n^T and point m is thus given by:

$$\begin{aligned}
n \cdot (x - m) &= 0 \\
\begin{bmatrix}-7 & -1 & 6\end{bmatrix} \cdot (x - \begin{bmatrix}-1.5 \\ 1.5 \\ 0\end{bmatrix}) &= 0 \\
\begin{bmatrix}-7 & -1 & 6\end{bmatrix} x + (-10.5 + 1.5 + 0) &= 0 \\
\begin{bmatrix}-7 & -1 & 6\end{bmatrix} x - 9 &= 0
\end{aligned}$$

Hence, a linear classifier would be $f(x) = \begin{bmatrix}-7 & -1 & 6\end{bmatrix} x - 9$.

2 Basic NN

$$z = w_1 x_1 + w_2 x_2 - w_0$$

Based on the points and their outputs, we can get the following equations:

$$\begin{aligned}
w_1(0) + w_2(0) - w_0 &\geq 0 \\
-w_0 &\geq 0 \qquad (1) \\
w_1(0) + w_2(-1) - w_0 &\geq 0 \\
-w_2 - w_0 &\geq 0 \qquad (2) \\
w_1(1) + w_2(0) - w_0 &< 0 \\
w_1 - w_0 &< 0 \qquad (3) \\
w_1(1) + w_2(-1) - w_0 &< 0 \\
w_1 - w_2 - w_0 &< 0 \qquad (4)
\end{aligned}$$

By observation, consider $w_0 = w_2 = 0$, $w_1 = -1$,

$$\begin{aligned}
&(1) : -(0) = 0 \geq 0 \\
&(2) : -(0) - (0) = 0 \geq 0 \\
&(3) : (-1) - (0) = -1 < 0 \\
&(4) : (-1) - (0) - (0) = -1 < 0
\end{aligned}$$

As seen, all the conditions are fulfilled. As such, a possible set of weights for this unit is $w_0 = w_2 = 0$, $w_1 = -1$.

3 Logistic Regression

1) Given $s(a) = \frac{e^a}{1+e^a} = \frac{1}{e^{-a}+1}$,

$$\begin{aligned}
\frac{\partial log(s(a))}{\partial a} &= \frac{\partial log(\frac{1}{e^{-a}+1})}{\partial a} \\
&= -\frac{\partial log(e^{-a}+1)}{\partial a} \\
&= -\frac{1}{e^{-a}+1}\frac{\partial (e^{-a}+1)}{\partial a} \\
&= -\frac{1}{e^{-a}+1}\left(\frac{\partial e^{-a}}{\partial a} + \frac{\partial 1}{\partial a}\right) \\
&= -\frac{1}{e^{-a}+1}\left(-e^{-a}+0\right) \\
&= \frac{e^{-a}}{e^{-a}+1} \\
&= \frac{e^{-a}+1-1}{e^{-a}+1} \\
&= 1 - \frac{1}{e^{-a}+1} \\
&= 1 - s(a) \text{ (shown)}
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial log(1-s(a))}{\partial a} &= \frac{\partial log(1-\frac{1}{e^{-a}+1})}{\partial a} \\
&= \frac{\partial log(\frac{e^{-a}+1-1}{e^{-a}+1})}{\partial a} \\
&= \frac{\partial log(\frac{e^{-a}}{e^{-a}+1})}{\partial a} \\
&= \frac{\partial log(e^{-a})}{\partial a} - \frac{\partial log(e^{-a}+1)}{\partial a} \\
&= \frac{\partial -a}{\partial a} - \frac{1}{e^{-a}+1}\frac{\partial(e^{-a}+1)}{\partial a} \\
&= -1 - \frac{1}{e^{-a}+1}\left(\frac{\partial e^{-a}}{\partial a} + \frac{\partial 1}{\partial a}\right) \\
&= -1 - \frac{1}{e^{-a}+1}\left(-e^{-a}+0\right) \\
&= -1 + \frac{e^{-a}}{e^{-a}+1} \\
&= \frac{-e^{-a}-1+e^{-a}}{e^{-a}+1} \\
&= -\frac{1}{e^{-a}+1} \\
&= -s(a) \text{ (shown)}
\end{aligned}
$$

2)

$$
\begin{aligned}
\nabla_w L = \nabla_w &\left((-1)\cdot\sum_{i=1}^{n} y_i log(h(x_i)) + (1-y_i)log(1-h(x_i))\right) \\
&= -\sum_{i=1}^{n} \nabla_w(y_i log(s(w_i\cdot x_i)) + (1-y_i)log(1-s(w_i\cdot x_i))) \\
&= -\sum_{i=1}^{n}(\nabla_w(y_i log(s(w_i\cdot x_i))) + \nabla_w((1-y_i)log(1-s(w_i\cdot x_i)))) \\
&= -\sum_{i=1}^{n}(y_i\nabla_w(log(s(w_i\cdot x_i))) + (1-y_i)\nabla_w(log(1-s(w_i\cdot x_i))))
\end{aligned}
$$

Substituting equations shown in part (1),

$$\begin{aligned}
\nabla_w L &= -\sum_{i=1}^{n}(y_i \nabla_w(log(s(w_i \cdot x_i))) + (1-y_i)\nabla_w(log(1-s(w_i \cdot x_i)))) \\
&= -\sum_{i=1}^{n}(((1-s(w_i \cdot x_i))x_i y_i + (1-y_i)(-s(w_i \cdot x_i)x_i))) \\
&= -x_i \sum_{i=1}^{n}(((1-s(w_i \cdot x_i))y_i + (1-y_i)(-s(w_i \cdot x_i)))) \\
&= -x_i \sum_{i=1}^{n}((y_i - s(w_i \cdot x_i)y_i - s(w_i \cdot x_i) + s(w_i \cdot x_i)y_i)) \\
&= \sum_{i=1}^{n}(-x_i(y_i - s(w_i \cdot x_i))) \\
&= \sum_{i=1}^{n}(x_i(s(w_i \cdot x_i) - y_i))
\end{aligned}$$

Since $h(x) = s(w \cdot x)$,

$$\begin{aligned}
\nabla_w L &= \sum_{i=1}^{n}(x_i(s(w_i \cdot x_i) - y_i)) \\
&= \sum_{i=1}^{n}(x_i(h(x_i) - y_i)) \text{ (shown)}
\end{aligned}$$

4 Back-Propagation

As $\frac{\partial L}{\partial n_2(x)}$ and $\frac{\partial L}{\partial n_3(x)}$ can be computed directly from L, we can consider them as given values. As such, we compute the other gradients as follows:

$$\begin{aligned}
\frac{\partial L}{\partial n_4(x)} &= \frac{\partial L}{\partial n_2(x)}\frac{\partial n_2(x)}{\partial n_4(x)} + \frac{\partial L}{\partial n_3(x)}\frac{\partial n_3(x)}{\partial n_4(x)} \\
&= \frac{\partial L}{\partial n_2(x)} g'(w_{2,4}n_4(x))w_{2,4} + \frac{\partial L}{\partial n_3(x)} g'(w_{3,4}n_4(x) + w_{3,6}n_6(x))w_{3,4} \\
\frac{\partial L}{\partial n_5(x)} &= \frac{\partial L}{\partial n_4(x)}\frac{\partial n_4(x)}{\partial n_5(x)} \\
&= \frac{\partial L}{\partial n_4(x)} g'(w_{4,5}n_5(x) + w_{4,6}n_6(x))w_{4,5} \\
\frac{\partial L}{\partial n_6(x)} &= \frac{\partial L}{\partial n_4(x)}\frac{\partial n_4(x)}{\partial n_6(x)} + \frac{\partial L}{\partial n_3(x)}\frac{\partial n_3(x)}{\partial n_6(x)} \\
&= \frac{\partial L}{\partial n_4(x)} g'(w_{4,5}n_5(x) + w_{4,6}n_6(x))w_{4,6} + \frac{\partial L}{\partial n_3(x)} g'(w_{3,4}n_4(x) + w_{3,6}n_6(x))w_{3,6} \\
\frac{\partial L}{\partial n_7(x)} &= \frac{\partial L}{\partial n_5(x)}\frac{\partial n_5(x)}{\partial n_7(x)} + \frac{\partial L}{\partial n_6(x)}\frac{\partial n_6(x)}{\partial n_7(x)} \\
&= \frac{\partial L}{\partial n_5(x)} g'(w_{5,7}n_7(x))w_{5,7} + \frac{\partial L}{\partial n_6(x)} g'(w_{6,7}n_7(x))w_{6,7} \\
\frac{\partial L}{\partial x} &= \frac{\partial L}{\partial n_7(x)}\frac{\partial n_7(x)}{\partial x} \\
&= \frac{\partial L}{\partial n_7(x)} g'(w_{7,x}x)w_{7,x}
\end{aligned}$$

5 Some Einsum

For $C_{jk} = \sum_i A_{ijk}b_i$, einsum notation in code: `"ijk, i -> jk", [A, b]`.

For $A_{ik} = \sum_{j,l} A_{ijkl}$, einsum notation in code: `"ijkl -> ik", [A]`.

For $A_{ki} = \sum_{j,l} A_{ijkl}$, einsum notation in code: `"ijkl -> ki", [A]`.

For $C_i = \sum_{j,k} A_{ijk}A_{ijk}$, einsum notation in code: `"ijk, ijk -> i", [A, A]`.

For $C = AG^TB, A \in \mathbb{R}^{d\times e}, 2-tensor, G \in \mathbb{R}^{f\times e}, 2-tensor, B \in \mathbb{R}^{f\times l}$, einsum notation in code: `"de, fe, fl -> dl", [A, G, B]`.