Homework 2

CSCI 567: Machine Learning

Seyed Mohammad Asghari Pari

Neural networks

Solution to Question 1.1:

Calculating $\frac{\partial l}{\partial \boldsymbol{u}}$: We first calculate $\frac{\partial l}{\partial \boldsymbol{u}_m}$ as follows,

$$\frac{\partial l}{\partial \boldsymbol{u}_m} = \frac{\partial l}{\partial \boldsymbol{h}_m}\frac{\partial \boldsymbol{h}_m}{\partial \boldsymbol{u}_m} = \frac{\partial l}{\partial \boldsymbol{h}_m} H(\boldsymbol{u}_m) \tag{1}$$

where the first equality is true because $\boldsymbol{u}_m$ affects only $\boldsymbol{h}_m$ and the last equality is true by the definition of Heaviside step function. From (1), we can write

$$\frac{\partial l}{\partial \boldsymbol{u}} = \frac{\partial l}{\partial \boldsymbol{h}} \cdot * H(\boldsymbol{u}). \tag{2}$$

Now, we calculate $\frac{\partial l}{\partial \boldsymbol{h}}$ and for that we first derive $\frac{\partial l}{\partial \boldsymbol{h}_m}$ as follows,

$$\frac{\partial l}{\partial \boldsymbol{h}_m} = \sum_{k=1}^{K} \frac{\partial l}{\partial \boldsymbol{a}_k}\frac{\partial \boldsymbol{a}_k}{\partial \boldsymbol{h}_m} = \sum_{k=1}^{K} \frac{\partial l}{\partial \boldsymbol{a}_k} W^{(2)}_{km} = [W^{(2)}]^{\mathsf{T}}_{m\bullet}\frac{\partial l}{\partial \boldsymbol{a}}, \tag{3}$$

where $[W^{(2)}]_{m\bullet}$ is the m-th column of matrix $W^{(2)}$. From (3), we can write

$$\frac{\partial l}{\partial \boldsymbol{h}} = \left(W^{(2)}\right)^{\mathsf{T}}\frac{\partial l}{\partial \boldsymbol{a}}. \tag{4}$$

Finally, from (2) and (4), we get,

$$\boxed{\frac{\partial l}{\partial \boldsymbol{u}} = \left(\left(W^{(2)}\right)^{\mathsf{T}}\frac{\partial l}{\partial \boldsymbol{a}}\right) \cdot * H(\boldsymbol{u}).} \tag{5}$$

Calculating $\frac{\partial l}{\partial \boldsymbol{a}}$: We first calculate $\frac{\partial l}{\partial \boldsymbol{a}_k}$ as follows,

$$\frac{\partial l}{\partial \boldsymbol{a}_k} = \sum_{j=1}^{K} \frac{\partial l}{\partial \boldsymbol{z}_j}\frac{\partial \boldsymbol{z}_j}{\partial \boldsymbol{a}_k} = \sum_{j=1}^{K} -\frac{\boldsymbol{y}_j}{\boldsymbol{z}_j}\frac{\partial \boldsymbol{z}_j}{\partial \boldsymbol{a}_k} = \begin{bmatrix}\frac{\partial \boldsymbol{z}_1}{\partial \boldsymbol{a}_k} & \frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{a}_k} & \cdots & \frac{\partial \boldsymbol{z}_K}{\partial \boldsymbol{a}_k}\end{bmatrix}\begin{bmatrix}-\frac{\boldsymbol{y}_1}{\boldsymbol{z}_1}\\ -\frac{\boldsymbol{y}_2}{\boldsymbol{z}_2}\\ \vdots\\ -\frac{\boldsymbol{y}_K}{\boldsymbol{z}_K}\end{bmatrix} = \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{a}_k}\left(\boldsymbol{y} \cdot * \frac{\mathbf{1}}{\boldsymbol{z}}\right) \tag{6}$$

where we have used $\frac{\mathbf{1}}{\boldsymbol{z}}$ to denote $\begin{bmatrix}\frac{1}{\boldsymbol{z}_1} & \cdots & \frac{1}{\boldsymbol{z}_K}\end{bmatrix}^{\mathsf{T}}$. Note that from (6) we can write,

$$\frac{\partial l}{\partial \boldsymbol{a}} = \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{a}}\left(\boldsymbol{y} \cdot * \frac{\mathbf{1}}{\boldsymbol{z}}\right). \tag{7}$$

Next, we calculate $\frac{\partial \boldsymbol{z}_j}{\partial \boldsymbol{a}_k}$ as follows: if $j = k$, then we have

$$\frac{\partial \boldsymbol{z}_k}{\partial \boldsymbol{a}_k} = \frac{e^{\boldsymbol{a}_k}\sum_{k'} e^{\boldsymbol{a}_{k'}} - e^{\boldsymbol{a}_k}e^{\boldsymbol{a}_k}}{\left(\sum_{k'} e^{\boldsymbol{a}_{k'}}\right)^2} = \boldsymbol{z}_k - \boldsymbol{z}_k\boldsymbol{z}_k, \tag{8}$$

and if $j \neq k$, then we have

$$\frac{\partial \boldsymbol{z}_j}{\partial \boldsymbol{a}_k} = \frac{-e^{\boldsymbol{a}_k} e^{\boldsymbol{a}_j}}{\left(\sum_{k'} e^{\boldsymbol{a}_{k'}}\right)^2} = -\boldsymbol{z}_k \boldsymbol{z}_j. \tag{9}$$

From (8) and (9), we can write,

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{a}} = diag(\boldsymbol{z}) - \boldsymbol{z}\boldsymbol{z}^{\intercal}, \tag{10}$$

where $diag(\boldsymbol{z})$ is a matrix whose diagonal is vector $\boldsymbol{z}$ and all other entries are zero. Finally, from (7) and (10), we get

$$\boxed{\frac{\partial l}{\partial \boldsymbol{a}} = \left(diag(\boldsymbol{z}) - \boldsymbol{z}\boldsymbol{z}^{\intercal}\right)\left(\boldsymbol{y} \cdot * \frac{\boldsymbol{1}}{\boldsymbol{z}}\right).} \tag{11}$$

Calculating $\frac{\partial l}{\partial W^{(1)}}$: We first calculate $\frac{\partial l}{\partial W^{(1)}_{ij}}$ as follows,

$$\frac{\partial l}{\partial W^{(1)}_{ij}} = \frac{\partial l}{\partial \boldsymbol{u}_i} \frac{\partial \boldsymbol{u}_i}{\partial W^{(1)}_{ij}} = \frac{\partial l}{\partial \boldsymbol{u}_i} \boldsymbol{x}_j, \tag{12}$$

where the first equality is true because $W^{(1)}_{ij}$ affects only $\boldsymbol{u}_i$. From (12), we can write,

$$\boxed{\frac{\partial l}{\partial W^{(1)}} = \frac{\partial l}{\partial \boldsymbol{u}} \boldsymbol{x}^{\intercal}.} \tag{13}$$

Calculating $\frac{\partial l}{\partial \boldsymbol{b}^{(1)}}$: We first calculate $\frac{\partial l}{\partial \boldsymbol{b}^{(1)}_i}$ as follows,

$$\frac{\partial l}{\partial \boldsymbol{b}^{(1)}_i} = \frac{\partial l}{\partial \boldsymbol{u}_i} \frac{\partial \boldsymbol{u}_i}{\partial \boldsymbol{b}^{(1)}_i} = \frac{\partial l}{\partial \boldsymbol{u}_i}, \tag{14}$$

where the first equality is true because $\boldsymbol{b}^{(1)}_i$ affects only $\boldsymbol{u}_i$. From (14), we can write,

$$\boxed{\frac{\partial l}{\partial \boldsymbol{b}^{(1)}} = \frac{\partial l}{\partial \boldsymbol{u}}.} \tag{15}$$

Calculating $\frac{\partial l}{\partial W^{(2)}}$: We first calculate $\frac{\partial l}{\partial W^{(2)}_{ij}}$ as follows,

$$\frac{\partial l}{\partial W^{(2)}_{ij}} = \frac{\partial l}{\partial \boldsymbol{a}_i} \frac{\partial \boldsymbol{a}_i}{\partial W^{(2)}_{ij}} = \frac{\partial l}{\partial \boldsymbol{a}_i} \boldsymbol{h}_j, \tag{16}$$

where the first equality is true because $W^{(2)}_{ij}$ affects only $\boldsymbol{a}_i$. From (16), we can write,

$$\boxed{\frac{\partial l}{\partial W^{(2)}} = \frac{\partial l}{\partial \boldsymbol{a}} \boldsymbol{h}^{\intercal}.} \tag{17}$$

Solution to Question 1.2:

We know that the update equation in stochastic gradient descent for a variable X when we are going to minimize a function l is

$$X^{t+1} = X^t - \eta \frac{\partial l}{\partial X}\Big|_{X=X^t}, \tag{18}$$

where X^t is the value for variable X at the t-th iteration. Note that, (18) can be written as

$$X^{t+1} = X^0 - \sum_{s=0}^{t} \eta \frac{\partial l}{\partial X}\Big|_{X=X^s}. \tag{19}$$

Now, let X be any of variables $W^{(1)}$, $W^{(2)}$, $\boldsymbol{b}^{(1)}$. Since we have set the initial values for $W^{(1)}$, $W^{(2)}$, $\boldsymbol{b}^{(1)}$ to be zero matrices/vectors and we further know that $\frac{\partial l}{\partial W^{(1)}}$, $\frac{\partial l}{\partial W^{(1)}}$, and $\frac{\partial l}{\partial \boldsymbol{b}^{(1)}}$ are all zero matrices/vectors, we can easily conclude from (19) that $W^{(1)}$, $W^{(2)}$, $\boldsymbol{b}^{(1)}$ all stay zero matrices/vectors no matter how many iterations we perform. Therefore, no learning will happen on $W^{(1)}$, $W^{(2)}$, $\boldsymbol{b}^{(1)}$.

For the case of stochastic gradient descent with momentum, the update equation for velocity $\boldsymbol{v}$ is

$$\boldsymbol{v}^{t+1} = \alpha \boldsymbol{v}^t - \eta \frac{\partial l}{\partial X}\Big|_{X=X^t}, \tag{20}$$

$$X^{t+1} = X^t + \boldsymbol{v}^{t+1}, \tag{21}$$

which can be written as

$$\boldsymbol{v}^{t+1} = \alpha^{t+1}\boldsymbol{v}^0 - \sum_{s=0}^{t} \eta \alpha^{t-s} \frac{\partial l}{\partial X}\Big|_{X=X^s}, \tag{22}$$

$$X^{t+1} = X^0 + \sum_{s=0}^{t+1} \boldsymbol{v}^s. \tag{23}$$

Again since $\boldsymbol{v}^0$ is zero vector, if let X be any of variables $W^{(1)}$, $W^{(2)}$, $\boldsymbol{b}^{(1)}$ no learning will happen on $W^{(1)}$, $W^{(2)}$, $\boldsymbol{b}^{(1)}$. The reason again is that we have set the initial values for $W^{(1)}$, $W^{(2)}$, $\boldsymbol{b}^{(1)}$ to be zero matrices/vectors and we further know that $\frac{\partial l}{\partial W^{(1)}}$, $\frac{\partial l}{\partial W^{(1)}}$, and $\frac{\partial l}{\partial \boldsymbol{b}^{(1)}}$ are all zero matrices/vectors. Therefore, $\boldsymbol{v}$ stays zero and then from (23), X stays zero.

Solution to Question 1.3:

If $\boldsymbol{a} = W^{(2)}\boldsymbol{u} + \boldsymbol{b}^{(2)}$, then since $\boldsymbol{u} = W^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$ we have

$$\boldsymbol{a} = W^{(2)}\big(W^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\big) + \boldsymbol{b}^{(2)} = W^{(2)}W^{(1)}\boldsymbol{x} + W^{(2)}\boldsymbol{b}^{(1)} + \boldsymbol{b}^{(2)}. \tag{24}$$

Hence, we have

$$\boxed{\boldsymbol{U} = W^{(2)}W^{(1)}, \quad \boldsymbol{v} = W^{(2)}\boldsymbol{b}^{(1)} + \boldsymbol{b}^{(2)}.} \tag{25}$$

Kernel methods

Solution to Question 2.1:

In this question, we have

$$J(\boldsymbol{w}) = \sum_{n=1}^{N} l(\boldsymbol{w}^\intercal \phi(\boldsymbol{x}_n), y_n) + \frac{\lambda}{2} \parallel \boldsymbol{w} \parallel_2^2 . \tag{26}$$

If we take the derivative of $J(\boldsymbol{w})$ with respect to $\boldsymbol{w}$ and set it to zero, we get

$$\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}} = 0, \quad \implies \quad \sum_{n=1}^{N} \frac{\partial l(s_n, y_n)}{\partial s_n} \frac{\partial s_n}{\partial \boldsymbol{w}}\Big|_{s_n = \boldsymbol{w}^\intercal \phi(\boldsymbol{x}_n)} + \lambda \boldsymbol{w} = 0 \tag{27}$$

$$\implies \quad \sum_{n=1}^{N} \frac{\partial l(s_n, y_n)}{\partial s_n}\Big|_{s_n = \boldsymbol{w}^\intercal \phi(\boldsymbol{x}_n)} \phi(\boldsymbol{x}_n) + \lambda \boldsymbol{w} = 0 \tag{28}$$

$$\implies \quad \boldsymbol{w} = \sum_{n=1}^{N} \frac{-1}{\lambda} \frac{\partial l(s_n, y_n)}{\partial s_n}\Big|_{s_n = \boldsymbol{w}^\intercal \phi(\boldsymbol{x}_n)} \phi(\boldsymbol{x}_n) \tag{29}$$

$$\implies \quad \boxed{\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n \phi(\boldsymbol{x}_n),} \tag{30}$$

where we have denoted $\frac{-1}{\lambda}\frac{\partial l(s_n, y_n)}{\partial s_n}\Big|_{s_n=\boldsymbol{w}^\intercal \phi(\boldsymbol{x}_n)}$ by α_n.

Solution to Question 2.2:

Note that from (30), we can write

$$\boldsymbol{w} = \Phi^\intercal \boldsymbol{\alpha}, \tag{31}$$

where $\Phi = [\phi(\boldsymbol{x}_1)^\intercal \quad \dots \quad \phi(\boldsymbol{x}_N)^\intercal]^\intercal$. Now, by substituting (31) in (26), we get

$$J(\boldsymbol{\alpha}) = \sum_{n=1}^{N} l(\boldsymbol{\alpha}^\intercal \Phi \phi(\boldsymbol{x}_n), y_n) + \frac{\lambda}{2} \parallel \Phi^\intercal \boldsymbol{\alpha} \parallel_2^2 = \sum_{n=1}^{N} l(\boldsymbol{\alpha}^\intercal \Phi \phi(\boldsymbol{x}_n), y_n) + \frac{\lambda}{2} \boldsymbol{\alpha}^\intercal \Phi \Phi^\intercal \boldsymbol{\alpha} \tag{32}$$

which can be written as,

$$\boxed{J(\boldsymbol{\alpha}) = \sum_{n=1}^{N} l(\boldsymbol{\alpha}^\intercal [K]_{n\bullet}, y_n) + \frac{\lambda}{2} \boldsymbol{\alpha}^\intercal K \boldsymbol{\alpha},} \tag{33}$$

where $K = \Phi\Phi^\intercal$ and $[K]_{n\bullet}$ is the n-th column of matrix K, that is,

$$[K]_{n\bullet} = \begin{bmatrix} \phi(x_1)^\intercal \phi(x_n) \\ \vdots \\ \phi(x_N)^\intercal \phi(x_n) \end{bmatrix}. \tag{34}$$