CSCI 5521 Homework 4

John Nguyen

December 2, 2019

When I use the log function, I mean the natural log, ln

Problem 1

(a). Professor HighLowHigh is correct.

Professor HighLowHigh claims that $\mathbf{v}^t = \mathbf{x}^t$ for all $t = 1, ..., N$. In the problem, $\mathbf{x}^t$ is the input data and $\mathbf{v}^t$ is defined as $\mathbf{v}^t = W\mathbf{z}^t$, where $W \in \mathbb{R}^{D\times d}$ is the transformation matrix from the higher dimensional space $\mathbb{R}^D$ and $^t \in \mathbb{R}^D$ is the new d-dimensional features, post transformation. So W is composed of d most principle components of the covariance matrix of the original D-dimensional data, $\mathbf{x}^t$. Furthermore, $\mathbf{z}^t$ is defined as $\mathbf{z}^t = W^T\mathbf{x}^t$. Therefore, the following is equivalent to Dr. HighLowHigh's claim:

$$WW^T\mathbf{x}^t = \mathbf{x}^t$$

However, notice that $WW^T\mathbf{x}^t = \mathbf{x}^t \iff WW^T = I$, so it is sufficient to prove that $WW^T = I$. Notice that $WW^T = I$ is equivalent to starting the rows of W are orthogonal. Therefore, we want to prove that the rows of W is orthogonal. W is the d most principle eigenvectors of the covariance matrix of the original data, Σ. Since $\Sigma = \frac{1}{N}\sum_{t=1}^{N}\mathbf{x}^t(\mathbf{x})^T$ is symmetric, the eigenvectors of Σ are orthogonal. Therefore, the row/column vectors of W are orthogonal, i.e. for all $1 \leq i, j \leq d$, $i \neq d$, $< w_i, w_j >= 0$ and $< w_i, w_i >= 1$ for all i, j and where $< \cdot, \cdot >$ is the dot product of two d-dimensional vectors. Therefore, $WW^T = I$, where $I \in \mathbb{R}^{D\times D}$. Thus, $WW^T = I$. Therefore, $\mathbf{v}^t = WW^T\mathbf{x}^t = I\mathbf{x}^t = \mathbf{x}^t$, as desired.

(b). Based on the result in Problem 1(a), yes, the equality trivially holds. Since $\mathbf{v}^t = \mathbf{x}^t$:

$$\begin{aligned}
\sum_{t=1}^{N} ||\mathbf{x}^t||_2^2 - \sum_{t=1}^{N} ||\mathbf{v}^t||_2^2 &= \sum_{t=1}^{N} ||\mathbf{x}^t||_2^2 - \sum_{t=1}^{N} ||\mathbf{x}^t||_2^2 \\
&= 0 \\
&= \sum_{t=1}^{N} 0 \\
&= \sum_{t=1}^{N} (0)^2 \\
&= \sum_{t=1}^{N} \sum_{j=1}^{D} (x_j^t - x_j^t)^2 \\
&= \sum_{t=1}^{N} ||\mathbf{x}^t - \mathbf{x}^t||_2^2 \\
&= \sum_{t=1}^{N} ||\mathbf{x}^t - \mathbf{v}^t||_2^2
\end{aligned}$$

Problem 2

(a). We want to prove the update phase of gradient descent for the parameter $v_{i,h}$ updates with: $\eta \Delta_i^t z_h^t$, where $\Delta_i^t = -g'(a_i^t)\frac{\partial L(r_i^t, y_i^t)}{\partial y_i^t}$. Since we are trying to optimize the function $L(r_i^t, y_i^t)$, the gradient we want to calculate is: $\frac{\partial L(r_i^t, y_i^t)}{\partial v_{i,h}}$. However, notice that y_i^t is a function of a_i^t and a_i^t is a function of $v_{i,h}$, so we must apply the chain rule:

$$\frac{\partial L(r_i^t, y_i^t)}{\partial v_{i,h}} = \frac{\partial L(r_i^t, y_i^t)}{\partial y_i^t} \frac{\partial y_i^t}{\partial a_i^t} \frac{\partial a_i^t}{\partial v_{i,h}}$$

First, notice that $y = g(a_i^t)$, so $\frac{\partial y_i^t}{\partial a_i^t} = g'(a_i^t)$. Furthermore, notice that $a_i^t = \sum_{h=1}^{H} v_{i,h} z_h^t + v_{i,0}$, so a_i^t is a linear function of $v_{i,h}$. Therefore, $\frac{\partial a_i^t}{\partial v_{i,h}} = z_h^t$. Lastly, recall that η is a parameter which scales the gradient. When using gradient descent, we subtract the gradient, which is where the -1 term comes in from $-g'(a_i^t)$. Thus in conclusion, the gradient updating process is:

$$\begin{aligned}
v_{i,h}^{new} &= v_{i,h}^{old} + \Delta v_{i,h} \\
&= v_{i,h}^{old} - \eta \frac{\partial L(r_i^t, y_i^t)}{\partial y_i^t} \frac{\partial y_i^t}{\partial a_i^t} \frac{\partial a_i^t}{\partial v_{i,h}} \\
&= v_{i,h}^{old} - \eta \left(\frac{\partial L(r_i^t, y_i^t)}{\partial y_i^t} \right) g'(a_i^t) z_h^t \\
&= v_{i,h}^{old} + \eta \Delta_i^t z_h^t
\end{aligned}$$

as desired.

(b). We will use a similar process as Problem 2(a). In this problem we want to optimize the objective function $L(r_i^t, y_i^t)$ with respect to w_h, j. Notice that $L(r_i^t, y_i^t)$ is a function of y_i^t, for all $i = 1, ..., k$, which is a function of

a_i^t, for each respective i, which is a function of z_h^t, which is a function of $w_{h,j}$. In Problem 2(a), we calculated the gradient for a specific value of i. In this case, we must consider the gradient for every possible value of i. Therefore in order to calculate the gradient, we apply the chain rule to find that we want to calculate:

$$\frac{\partial L(r_i^t, y_i^t)}{\partial w_{h,j}} = \sum_{i=1}^{k} \frac{\partial L(r_i^t, y_i^t)}{\partial y_i^t} \frac{\partial y_i^t}{\partial a_i^t} \frac{\partial a_i^t}{\partial z_h^t} \frac{\partial z_h^t}{\partial a_h^t} \frac{\partial a_h^t}{\partial w_{h,j}}$$

In Problem 2(a), we have calculated that $\frac{\partial y_i^t}{\partial a_i^t} = g'(a_i^t)$.

$\frac{\partial a_i^t}{\partial z_h^t}$ is the derivative of a_i^t with respect to z_h^t. $a_i^t = \sum_{h=1}^{H} v_{i,h} z_h^t + v_{i,0}$, so by taking the derivative with respect to z_h^t, we find: $\frac{\partial a_i^t}{\partial z_h^t} = v_{i,h}$.

Next, we want to calculate $\frac{\partial z_h^t}{\partial a_h^t}$. Notice that $z_h^t = g(a_h^t)$, so by chain rule, $\frac{\partial z_h^t}{\partial a_h^t} = g'(a_h^t)$.

We want to calculate $\frac{\partial a_h^t}{\partial w_{h,j}}$, where $a_h^t = \sum_{j=1}^{d} w_{h,j} x_j^t + w_0$, which is a linear function with respect to $w_{h,j}$. Therefore, $\frac{\partial a_h^t}{\partial w_{h,j}} = x_j^t$.

So, in conclusion,

$$\begin{aligned}
w_{h,j}^{new} &= w_{h,j}^{old} - \eta \frac{\partial L(r_i^t, y_i^t)}{\partial w_{h,j}} \\
&= w_{h,j}^{old} - \eta \sum_{i=1}^{k} \frac{\partial L(r_i^t, y_i^t)}{\partial y_i^t} \frac{\partial y_i^t}{\partial a_i^t} \frac{\partial a_i^t}{\partial z_h^t} \frac{\partial z_h^t}{\partial a_h^t} \frac{\partial a_h^t}{\partial w_{h,j}} \\
&= w_{h,j}^{old} - \eta \sum_{i=1}^{k} \left(\frac{\partial L(r_i^t, y_i^t)}{\partial y_i^t} \right) \left(g'(a_i^t) \right) \left(v_{i,h} \right) \left(g'(a_h^t) \right) \left(x_j^t \right) \\
&= w_{h,j}^{old} + \eta \left(g'(a_h^t) \right) \sum_{i=1}^{k} \left(\frac{\partial L(r_i^t, y_i^t)}{\partial y_i^t} \right) \left(-g'(a_i^t) \right) \left(v_{i,h} \right) \left(x_j^t \right) \\
&= w_{h,j}^{old} + \eta \left(g'(a_h^t) \right) \sum_{i=1}^{k} \Delta_i^t \left(v_{i,h} \right) \left(x_j^t \right) \\
&= w_{h,j}^{old} + \eta \Delta_h^t x_j^t
\end{aligned}$$

as desired.

Problem 3

Error Rates for MySVM2 on Boston50						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.42	0.44	0.44	0.40	0.17	0.37	0.10

Error Rates for MySVM2 on Boston75						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.24	0.32	0.55	0.16	0.04	0.26	0.17

Error Rates for Logistic Regression on Boston50						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.13	0.12	0.09	0.28	0.22	0.17	0.07

Error Rates for Logistic Regression on Boston75						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.10	0.08	0.13	0.11	0.05	0.09	0.03

Problem 4

(a). Notice that $g(u) = \max(0, u)$ can be rewritten as a piecewise function:

$$g(u) = \begin{cases} 0 & u \leq 0 \\ u & u > 0 \end{cases}$$

Therefore if we differentiate $g(u)$ with respect to a, then the derivative is the derivative of each piecewise component:

$$g'(u) = \begin{cases} 0 & u \leq 0 \\ 1 & u > 0 \end{cases}$$

So in the context of Formula 1 (from the homework), the gradient update is:

$$\begin{aligned} v_{i,h}^{new} &= v_{i,h}^{old} + \eta \Delta_i^t z_h^t \\ &= v_{i,h}^{old} - \eta \left(\frac{\partial L(r_i^t, y_i^t)}{\partial y_i^t} \right) g'(a_i^t) z_h^t \\ &= v_{i,h}^{old} - \eta \left(2y_i^t - 2r_i^t \right) g'(a_i^t) z_h^t \end{aligned}$$

We must break this formula into 2 cases: $a_i^t \leq u$ or $a_i^t > u$. First, assume $a_i^t \leq u$, so $g'(a_i^t) = 0$. Then the update is:

$$v_{i,h}^{old} - \eta \left(2y_i^t - 2r_i^t \right) (0) z_h^t = v_{i,h}^{old}$$

Now assume $a_i^t > u$, so $g'(a_i^t) = 1$. Then the update is:

$$v_{i,h}^{old} - \eta \left(2y_i^t - 2r_i^t \right) (1) z_h^t = v_{i,h}^{old} - \eta \left(2y_i^t - 2r_i^t \right) z_h^t$$

(b). In order to calculate $g'(a)$, we must decompose $g(a)$ into a piecewise function. From the definition of max and min functions,

$$g(a) = \begin{cases} a & a > 0 \\ 0 & a = 0 \\ \alpha a & a < 0 \end{cases}$$

Notice that when $a = 0$, $a = 0 = \alpha a$, so we can merge the $a = 0$ into either of the other two cases. Thus $g(a)$ as a piecewise function can be written as:

$$g(a) = \begin{cases} a & a \geq 0 \\ \alpha a & a < 0 \end{cases}$$

Reacll that the derivative of a piecewise function is the derivative of each piecewise component. Now, we differentiate $g(a)$ with respect to a to find:

$$g'(a) = \begin{cases} 1 & a \geq 0 \\ \alpha & a < 0 \end{cases}$$

(c). If we take $\alpha = 1$, the Formula 5 (in the homework) degenerates to the linear function $g(a) = a$. We prove this below.

First, take $\alpha = 1$, so $g(a) = \max(0, a) + (1)\min(0, a) = \max(0, a) + \min(0, a)$. We rewrite $g(a)$ as a piecewise function:

$$g(a) = \begin{cases} a & a \geq 0 \\ a & g < 0 \end{cases}$$

By writing $g(a)$ this way, we see that $g(a) = a$ for all $a \in \mathbb{R}$, which is a linear function of a.