CS680, Spring 2020, Assignment 2

Zhijie Wang, [redacted]

May 21, 2020

Exercise 1

1.

$$\begin{aligned}
&\frac{1}{2n}\|X\mathbf{w} + b\mathbf{1} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2 \\
&\frac{1}{2n}(X\mathbf{w} + b\mathbf{1} - \mathbf{y})^T(X\mathbf{w} + b\mathbf{1} - \mathbf{y}) + \lambda\mathbf{w}^T\mathbf{w} \\
&\frac{1}{2n}(\mathbf{w}^TX^T + b\mathbf{1}^T - \mathbf{y}^T)(X\mathbf{w} + b\mathbf{1} - \mathbf{y}) + \lambda\mathbf{w}^T\mathbf{w} \\
&\frac{1}{2n}(\mathbf{w}^TX^TX\mathbf{w} + b\mathbf{1}^TX\mathbf{w} - \mathbf{y}^TX\mathbf{w} + b\mathbf{w}^TX^T\mathbf{1} + b^2\mathbf{1}^T\mathbf{1} - b\mathbf{y}^T\mathbf{1} - \mathbf{w}^TX^T\mathbf{y} - b\mathbf{1}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}) + \lambda\mathbf{w}^T\mathbf{w}
\end{aligned} \tag{1}$$

Let's denote (1) as $f(\mathbf{w}, b)$. Therefore,

$$\begin{aligned}
\frac{\partial f}{\partial \mathbf{w}} \quad & \frac{1}{2n}(2X^TX\mathbf{w} + bX^T\mathbf{1} - X^T\mathbf{y} - X^T\mathbf{y}) + 2\lambda\mathbf{w} \\
& \frac{1}{n}X^T(X\mathbf{w} + b\mathbf{1} - \mathbf{y}) + 2\lambda\mathbf{w}
\end{aligned} \tag{2}$$

$$\begin{aligned}
\frac{\partial f}{\partial b} \quad & \frac{1}{2n}(\mathbf{1}^TX\mathbf{w} + \mathbf{w}^TX^T\mathbf{1} + 2b\cdot\mathbf{1}^T\mathbf{1} - \mathbf{y}^T\mathbf{1} - \mathbf{1}^T\mathbf{y}) \\
& \frac{1}{2n}(2\cdot\mathbf{1}^TX\mathbf{w} + 2b\cdot\mathbf{1}^T\mathbf{1} - 2\cdot\mathbf{1}^T\mathbf{y}) \\
& \frac{1}{n}\mathbf{1}^T(X\mathbf{w} + b\mathbf{1} - \mathbf{y})
\end{aligned} \tag{3}$$

2. Implemented in cs680-a2.ipynb.

   (a) If λ 0, training error and training loss is 17.78353985, test error is 47.4089435. The loss curve is shown as Fig. 1

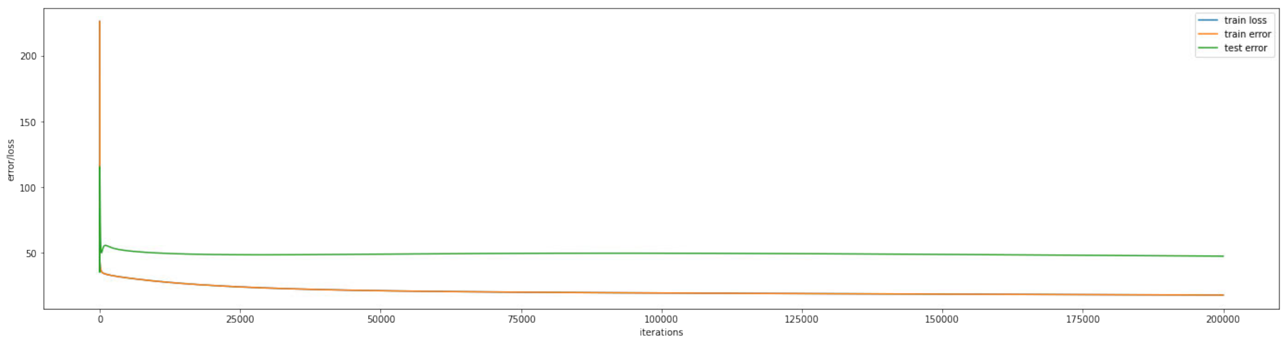

Figure 1:

   (b) If λ 10, training error is 22.83620562, training loss is 26.25654515, test error is 48.18984303. The loss curve is shown as Fig. 2

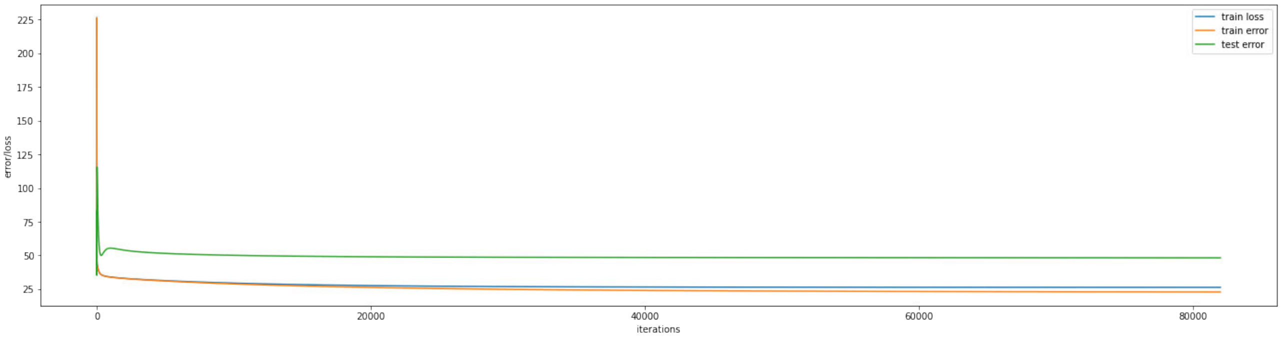

Figure 2:

3. No, they don't converge to the same solution. Actually, with this re-implementation, when λ 0, the training loss decrease to 13.40029951, the test error is 24.8248982, respectively. When λ 10, the training loss decrease to 22.42940329, the test error is 26.29971308.

   (a) If λ 0, the loss curve is shown as Fig. 3

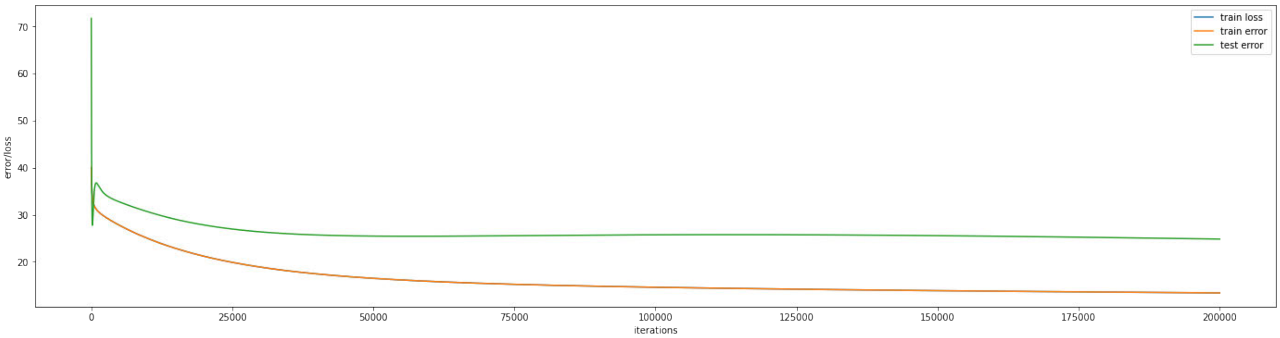

Figure 3:

   (b) If λ 10, the loss curve is shown as Fig. 4

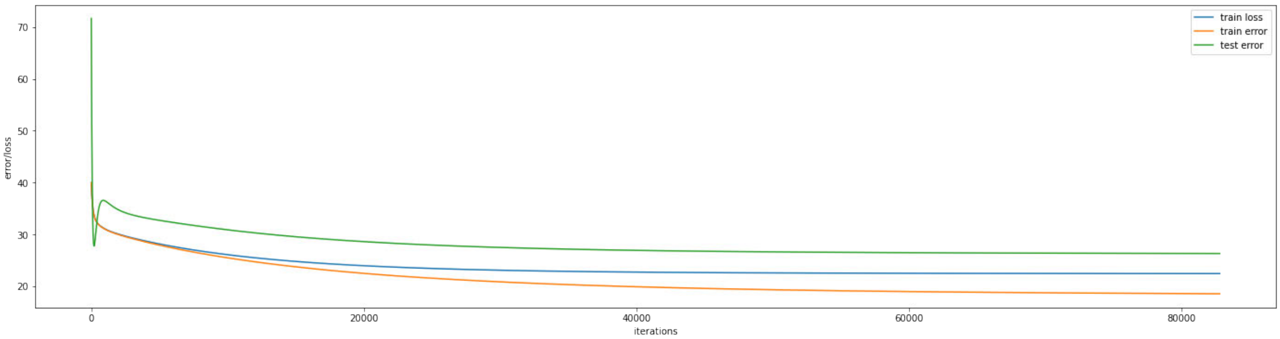

Figure 4:

4. The optimal b should be 0. We can verify this result by implement it.

   (a) If λ 0, the loss curve is shown as Fig. 5, b is finally 0.04580368.

   (b) If λ 10, the loss curve is shown as Fig. 6, b is finally 0.01814166.

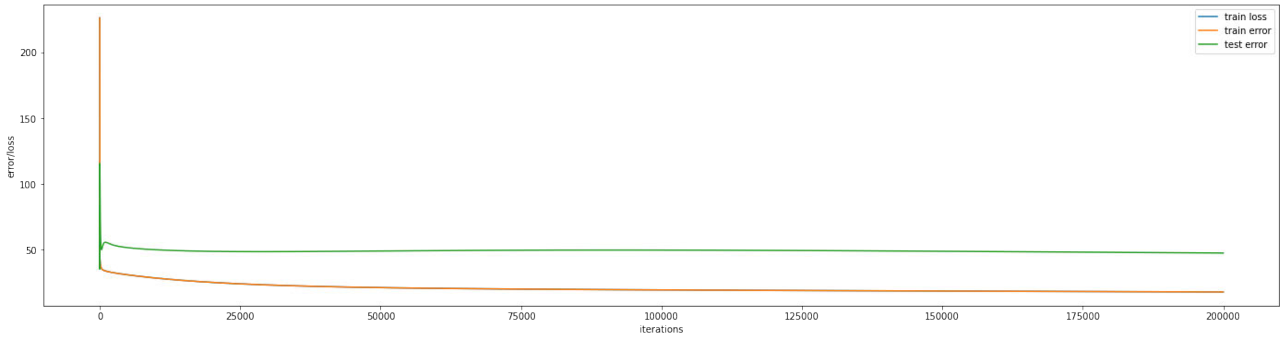

Figure 5:

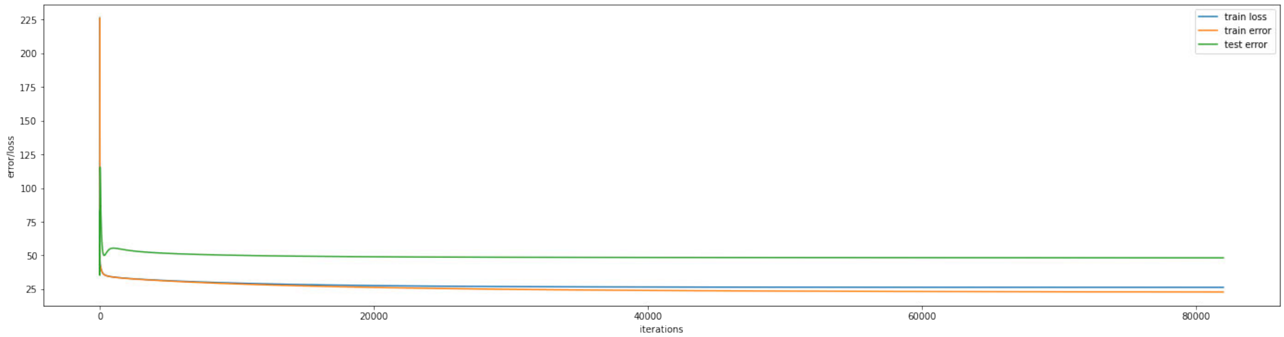

Figure 6:

Exercise 2

1. Take the Lagrangian.

$$
\begin{aligned}
\mathcal{L} \quad \frac{1}{2}\|\mathbf{x}-\mathbf{z}\|_2^2 + \lambda(\mathbf{w}^T\mathbf{x}+b) \quad & \frac{1}{2}(\mathbf{x}-\mathbf{z})^T(\mathbf{x}-\mathbf{z}) + \lambda(\mathbf{w}^T\mathbf{x}+b) \\
& \frac{1}{2}(\mathbf{x}^T-\mathbf{z}^T)(\mathbf{x}-\mathbf{z}) + \lambda(\mathbf{w}^T\mathbf{x}+b) \\
& \frac{1}{2}(\mathbf{x}^T\mathbf{x}-\mathbf{z}^T\mathbf{x}-\mathbf{x}^T\mathbf{z}+\mathbf{z}^T\mathbf{z}) + \lambda(\mathbf{w}^T\mathbf{x}+b)
\end{aligned} \tag{4}
$$

Let the derivative of the Lagrangian equal to 0, then

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{x}} \quad & \frac{1}{2}(2\mathbf{x}-\mathbf{z}-\mathbf{z}) + \lambda\mathbf{w} \quad 0 \\
\mathbf{x} \quad & \mathbf{z}-\lambda\mathbf{w}
\end{aligned} \tag{5}
$$

Put this expression of $\mathbf{x}$ into the equation of hyper-plane, then

$$
\begin{aligned}
\mathbf{w}^T(\mathbf{z}-\lambda\mathbf{w})+b \quad & 0 \\
\mathbf{w}^T\mathbf{z}-\lambda\mathbf{w}^T\mathbf{w}+b \quad & 0 \\
\lambda \quad & \frac{\mathbf{w}^T\mathbf{z}+b}{\mathbf{w}^T\mathbf{w}}
\end{aligned} \tag{6}
$$

So the distance is

$$
\|\mathbf{x}-\mathbf{z}\|_2 \quad \|\mathbf{z}-\lambda\mathbf{w}-\mathbf{z}\|_2 \quad |\lambda|\|\mathbf{w}\|_2 \quad \frac{|\mathbf{w}^T\mathbf{z}+b|}{\|\mathbf{w}\|_2} \tag{7}
$$

2. If $\mathbf{w}^T\mathbf{z}+b \leqslant 0$, which means that the point is in the halfspace, then the distance should be 0. If $\mathbf{w}^T\mathbf{z}+b \quad 0$, the distance to the halfpace is equal to the distance to hyperplane $\{\mathbf{x}\in\mathbb{R}^d : \mathbf{w}^T\mathbf{x}+b \quad 0\}$, therefore, the distance is $\frac{|\mathbf{w}^T\mathbf{z}+b|}{\|\mathbf{w}\|_2}$ as in (2).

3. To separate these two data points, we have two hyperplanes $\{\mathbf{x}\in\mathbb{R}^d : \mathbf{w}^T\mathbf{x}+b \quad 1\}$, where all points above this hyperplane have +1 label, and $\{\mathbf{x}\in\mathbb{R}^d : \mathbf{w}^T\mathbf{x}+b \quad -1\}$, where all points below have -1 label, which have a distance of $\frac{2}{\|\mathbf{w}\|_2}$. Therefore, to maximize the distance we need to minimize $\|\mathbf{w}\|_2$. Thus, the question become:

$$
\begin{aligned}
\min_{\mathbf{w}} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\
\text{subject to} \quad & \mathbf{w}^T\mathbf{x}_1+b \geqslant 1 \\
& \mathbf{w}^T\mathbf{x}_2+b \leqslant -1
\end{aligned} \tag{8}
$$

Now take the Lagrangian.

$$
\mathcal{L} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + \alpha[(\mathbf{w}^T\mathbf{x}_1+b)-1] + \beta[-(\mathbf{w}^T\mathbf{x}_2+b)-1] \tag{9}
$$

Let the derivative of the Lagrangian equal to 0, then

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{w}} \quad & \mathbf{w}+\alpha\mathbf{x}_1-\beta\mathbf{x}_2 \quad 0 \\
\frac{\partial \mathcal{L}}{\partial \mathbf{b}} \quad & \alpha-\beta \quad 0
\end{aligned} \tag{10}
$$

Thus, let $\alpha \quad \beta \quad c$, where c is a constant. Hence, $\mathbf{w} \quad c(\mathbf{x}_2-\mathbf{x}_1)$, and b does not affect the distance of margin.