MA234 · Introduction to Big Data Science

Assignment 4

刘乐奇 / 12011327

April 14, 2022

1 Exercise 1

It was done on the website.

2 Exercise 2

1. Introduce Lagrange multiplier $\alpha_i \geq 0$ and $\mu_i \geq 0$ of constraint $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$, respectively. Then we have Lagrange function

$$L(\mathbf{w}, b, \xi, \alpha, \mu) = \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^{n}\mu_i\xi_i$$

   So what we need to solve is $\min_{\mathbf{w},b,\xi}\max_{\alpha,\mu} L(\mathbf{w}, b, \alpha)$, and its dual problem is $\max_{\alpha}\min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha)$.

$$\begin{aligned}
\nabla_{\mathbf{w}} L = 0 &\implies \mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \\
\frac{\partial L}{\partial b} = 0 &\implies \sum_i \alpha_i y_i = 0 \\
\frac{\partial L}{\partial \xi_i} = 0 &\implies \alpha_i + \mu_i = C
\end{aligned}$$

   For $\alpha_i > 0$, we have $\alpha_i[y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i] = 0$. For $\mu_i > 0$, we have $\mu_i\xi_i = 0$.
   After all, the KKT condition is as below:

$$\begin{cases}
\alpha_i \geq 0 \\
y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i \geq 0 \\
\alpha_i[y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i] = 0 \\
\mu_i \geq 0 \\
\xi_i \geq 0 \\
\mu_i\xi_i = 0 \\
\sum_i \alpha_i y_i = 0 \\
\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \\
\alpha_i + \mu_i = C
\end{cases}$$

2. Back substitute the KKT condition into L, we have

$$\begin{aligned}
L &= \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^{n}\mu_i\xi_i \\
&= \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\mu_i\xi_i - \sum_{i=1}^{n}[\alpha_i y_i \mathbf{w}^T\mathbf{x}_i + \alpha_i y_i b + \alpha_i(\xi_i - 1)] \\
&= \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\mu_i\xi_i - \mathbf{w}^T\sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i - b\sum_{i=1}^{n}\alpha_i y_i - \sum_{i=1}^{n}\alpha_i\xi_i + \sum_{i=1}^{n}\alpha_i
\end{aligned}$$

   Since $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$ and $\sum_i \alpha_i y_i = 0$, we have

$$\begin{aligned}
L &= -\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\mu_i\xi_i - \sum_{i=1}^{n}\alpha_i\xi_i + \sum_{i=1}^{n}\alpha_i \\
&= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}(\alpha_i + \mu_i)\xi_i
\end{aligned}$$

Since $\alpha_i + \mu_i = C$ and $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$, we have

$$\begin{aligned} L &= \sum_{i=1}^n \alpha_i - \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ &= \sum_{i=1}^n \alpha_i - \frac{1}{2}(\sum_{i=1}^n \alpha_i y_i \mathbf{x}_i)^T(\sum_{i=1}^n \alpha_i y_i \mathbf{x}_i) \\ &= \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y_i y_j(\mathbf{x}_i^T\mathbf{x}_j) \end{aligned}$$

So the dual optimization problem is

$$\begin{aligned} \min_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y_i y_j(\mathbf{x}_i^T\mathbf{x}_j) \\ s.t. \sum_{i=1}^n \alpha_i y_i = 0,\ 0 \le \alpha_i \le C,\ i = 1, \dots, n \end{aligned}$$

3. (a) Since the inner product of vectors has commutativity, then

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j) =< \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) >=< \phi(\mathbf{x}_j), \phi(\mathbf{x}_i) >= \phi(\mathbf{x}_j)^T\phi(\mathbf{x}_i) = K(\mathbf{x}_j, \mathbf{x}_i)$$

So kernel is symmetric.

(b)

$$\begin{aligned} \mathbf{z}^T\mathbf{A}\mathbf{z} &= \sum_{i=1}^n\sum_{j=1}^n z_i \mathbf{A}(i,j) z_j \\ &= \sum_{i=1}^n\sum_{j=1}^n z_i \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j) z_j \\ &= \sum_{i=1}^n\sum_{j=1}^n z_i \sum_{k=1}^n \phi_k(\mathbf{x}_i)\phi_k(\mathbf{x}_j) z_j \\ &= \sum_{k=1}^n\sum_{i=1}^n\sum_{j=1}^n z_i \phi_k(\mathbf{x}_i)\phi_k(\mathbf{x}_j) z_j \\ &= \sum_{k=1}^n(\sum_{i=1}^n z_i\phi_k(\mathbf{x}_i))^2 \ge 0 \end{aligned}$$

So the kernel matrix is semi-positive definite.

3 Exercise 3

1. Since $y_i sign(f_{\mathbf{w}}(\mathbf{x}_i)) > 0$ for $i = 1, \dots, n$, i.e., $y_i sign(f_{\mathbf{w}}(\mathbf{x}_i)) = 1$, then the 0/1-loss minimization is to be $\min_{\mathbf{w}} 0$.
Since $y_i sign(f_{\mathbf{w}}(\mathbf{x}_i)) > 0$ for $i = 1, \dots, n$, without loss of generality, we have $y_i f_{\mathbf{w}}(\mathbf{x}_i) = y_i \mathbf{w}^T\mathbf{x}_i \ge 1$ (if not, just take $\mathbf{w} = \lambda\mathbf{w}$). Since $y_i\mathbf{w}^T\mathbf{x}_i = y_i\sum_{j=1}^n w_j x_{ij} = \sum_{j=1}^n y_i x_{ij} w_j =\ge 1$, then the constraint can be $\sum_{i=1}^n\sum_{j=1}^n y_i x_{ij} w_j =\ge \mathbf{1} \implies \mathbf{A}\mathbf{w} \ge \mathbf{1}$, where $A_{ij} = y_i x_{ij}, \mathbf{1} = (1, \dots, 1)^T \in \mathbb{R}$.
Therefore, the minimization can be formulated as

$$\min_{\mathbf{w}} 0\ s.t.\ \mathbf{A}\mathbf{w} \ge \mathbf{1}$$

2. Let

$$L(\mathbf{w}) = \sum_{i=1}^n (y_i - \mathbf{w}^T\mathbf{x}_i)^2$$

then

$$\begin{aligned}\nabla_{\mathbf{w}} L &= 2\sum_{i=1}^{n} \mathbf{x}_i^T (y_i - \mathbf{w}^T \mathbf{x}_i) \\ &= 2(\sum_{i=1}^{n} \mathbf{x}_i^T y_i - \mathbf{w}^T \sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i)\end{aligned}$$

Let it equal to 0, then we have

$$\mathbf{w} = \frac{\sum_{i=1}^{n} y_i \mathbf{x}_i}{\sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i}$$

3. For SVM, the primal optimization problem is

$$\begin{aligned}&\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{n} \xi_i \\ s.t.\ & y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, i = 1, \ldots, n \\ & \xi_i \geq 0, i = 1, \ldots, n\end{aligned}$$

We can rewrite the constraint as

$$\begin{aligned}&\xi_i \geq 1 - y_i(\mathbf{w}^T \mathbf{x}_i + b), i = 1, \ldots, n \\ &\xi_i \geq 0, i = 1, \ldots, n\end{aligned}$$

For hinge loss function as $L(y, f) = [1 - yf]_+ = \max\{1 - yf, 0\}$, we have

$$\xi_i = [1 - y_i(\mathbf{w}^T \mathbf{x}_i + b)]_+$$

Therefore, the optimization problem can be

$$\min_{\mathbf{w},b} \sum_{i=1}^{n} [1 - y_i(\mathbf{w}^T \mathbf{x}_i + b)]_+ + \lambda||\mathbf{w}||_2^2$$

The figure is as below

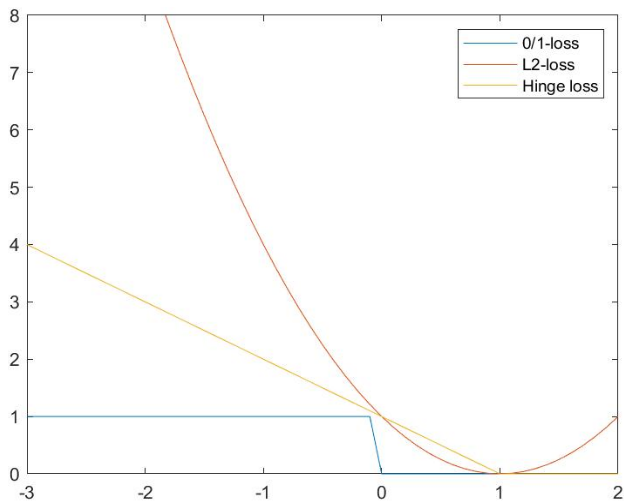

Figure 1: 0/1-loss, L2-loss, Hinge loss

4 Exercise 4

1. For Model 1, input the third example, we have

$$
\begin{aligned}
P(Y=1|\mathbf{X},w_1,w_2) &= g(w_1X1+w_2X2) = g(0) = \frac{1}{2} \\
P(Y=0|\mathbf{X},w_1,w_2) &= 1-P(Y=1|\mathbf{X},w_1,w_2) = \frac{1}{2}
\end{aligned}
$$

with the probabilities of positive and negative on the decision boundary is the same as 0.5, it will not be different if change the label to -1.
For Model 2, input the third example, we have

$$
\begin{aligned}
P(Y=1|\mathbf{X},w_1,w_2) &= g(w_0+w_1X1+w_2X2) = g(w_0) = \frac{1}{1+e^{-w_0}} \\
P(Y=0|\mathbf{X},w_1,w_2) &= 1-P(Y=1|\mathbf{X},w_1,w_2) = \frac{e^{-w_0}}{1+e^{-w_0}}
\end{aligned}
$$

Unless $w_0 = 0$, or the value of $\mathbf{w} = (w_1, w_2)$ will be different if change the label to -1.

2. Let

$$
\begin{aligned}
L(\mathbf{w}) &= \sum_i \log g(y^{(i)}\mathbf{w}^T\mathbf{x}^{(i)}) - \frac{\lambda}{2}||\mathbf{w}||^2 \\
&= \sum_i \frac{1}{2}y^{(i)}\mathbf{w}^T\mathbf{x}^{(i)} - \frac{\lambda}{2}||\mathbf{w}||^2
\end{aligned}
$$

then

$$
\begin{aligned}
\nabla_{\mathbf{w}}L &= \sum_i \frac{1}{2}y^{(i)}\mathbf{x}^{(i)} - \lambda\mathbf{w} = 0 \\
\implies \mathbf{w} &= \frac{1}{2\lambda}\sum_i y^{(i)}\mathbf{x}^{(i)}
\end{aligned}
$$

As λ increases, $\mathbf{w}$ will get closer to $\mathbf{0}$.

5 Exercise 5

1.

$$
\left.\begin{aligned}
\frac{\partial}{\partial w_{ij}^{(l)}}J(W,b;\mathbf{x},y) &= \frac{\partial J(W,b;\mathbf{x},y)}{\partial z_i^{(l+1)}}\frac{\partial z_i^{(l+1)}}{\partial w_{ij}^{(l)}} \\
\delta_i^{(l+1)} &= \frac{\partial}{\partial z_i^{(l+1)}}J(W,b;\mathbf{x},y) \\
\frac{\partial z_i^{(l+1)}}{\partial w_{ij}^{(l)}} &= \frac{\partial}{\partial w_{ij}^{(l)}}\sum_{j=1}^n w_{ij}^{(l)}a_j^{(l)} + b_i^{(l)} = a_j^{(l)}
\end{aligned}\right\} \implies \frac{\partial}{\partial w_{ij}^{(l)}}J(W,b;\mathbf{x},y) = \delta_i^{(l+1)}a_j^{(l)}
$$

$$
\left.\begin{aligned}
\frac{\partial}{\partial b_i^{(l)}}J(W,b;\mathbf{x},y) &= \frac{\partial J(W,b;\mathbf{x},y)}{z_i^{(l+1)}}\frac{z_i^{(l+1)}}{\partial b_i^{(l)}} \\
\delta_i^{(l+1)} &= \frac{\partial}{\partial z_i^{(l+1)}}J(W,b;\mathbf{x},y) \\
\frac{\partial z_i^{(l+1)}}{\partial b_i^{(l)}} &= \frac{\partial}{\partial b_i^{(l)}}\sum_{j=1}^n w_{ij}^{(l)}a_j^{(l)} + b_i^{(l)} = 1
\end{aligned}\right\} \implies \frac{\partial}{\partial b_i^{(l)}}J(W,b;\mathbf{x},y) = \delta_i^{(l+1)}
$$

2.

$$
\delta_i^{(L)} = \frac{\partial}{\partial z_i^{(L)}}J(W,b;\mathbf{x},y) = \frac{\partial J(W,b;\mathbf{x},y)}{\partial a_i^{(L)}}\frac{\partial a_i^{(L)}}{\partial z_i^{(L)}} = \frac{\partial J(W,b;\mathbf{x},y)}{\partial a_i^{(L)}}\frac{\partial f(z_i^{(L)})}{\partial z_i^{(L)}} = \frac{\partial J(W,b;\mathbf{x},y)}{\partial a_i^{(L)}}f'(z_i^{(L)})
$$

Since $J(W,b;\mathbf{x},y) = \frac{1}{2}||h_{W,b}(\mathbf{x}) - y||^2$ and $h_{W,b}(\mathbf{x}) = a^{(L)}$, then we have

$$\begin{aligned}
\delta_i^{(L)} &= \frac{\partial J(W,b;\mathbf{x},y)}{\partial a_i^{(L)}} f'(z_i^{(L)}) \\
&= \frac{\partial}{\partial a_i^{(L)}} (\frac{1}{2}||h_{W,b}(\mathbf{x}) - y||^2) f'(z_i^{(L)}) \\
&= \frac{\partial}{\partial a_i^{(L)}} (\frac{1}{2}||a^{(L)} - y||^2) f'(z_i^{(L)}) \\
&= \frac{\partial}{\partial a_i^{(L)}} (\frac{1}{2}\sum_{i=1}^{n} ||a_i^{(L)} - y||^2) f'(z_i^{(L)}) \\
&= (a_i^{(L)} - y) f'(z_i^{(L)})
\end{aligned}$$

Therefore, $\delta_i^{(L)} = (a_i^{(L)} - y) f'(z_i^{(L)})$.

$$\delta_i^{(l)} = \frac{\partial}{\partial z_i^{(l)}} J(W,b;\mathbf{x},y) = \sum_{j=1}^{S_{l+1}} \frac{\partial J(W,b;\mathbf{x},y)}{\partial z_j^{(l+1)}} \frac{\partial z_j^{(l+1)}}{\partial z_i^{(l)}} = \sum_{j=1}^{S_{l+1}} \delta_j^{(l+1)} \frac{\partial z_j^{(l+1)}}{\partial z_i^{(l)}}$$

Since

$$\begin{aligned}
z_j^{(l+1)} &= (\sum_{i=1}^{n} w_{ji}^{(l)} a_i^{(l)}) + b_j^{(l)} = (\sum_{i=1}^{n} w_{ji}^{(l)} f(z_i^{(l)})) + b_j^{(l)} \\
\implies \frac{\partial z_j^{(l+1)}}{\partial z_i^{(l)}} &= w_{ji}^{(l)} f'(z_i^{(l)})
\end{aligned}$$

we have

$$\delta_i^{(l)} = \sum_{j=1}^{S_{l+1}} \delta_j^{(l+1)} \frac{\partial z_j^{(l+1)}}{\partial z_i^{(l)}} = \sum_{j=1}^{S_{l+1}} \delta_j^{(l+1)} w_{ji}^{(l)} f'(z_i^{(l)})$$

Therefore, $\delta_i^{(l)} = \sum_{j=1}^{S_{l+1}} \delta_j^{(l+1)} w_{ji}^{(l)} f'(z_i^{(l)})$.