Machine Learning
Seoul National University

Homework #(**2**)
Haneul Choi

INSTRUCTIONS

- Anything that is received after the deadline will be considered to be late and we do not receive late homeworks. We do however ignore your lowest homework grade.
- Answers to every theory questions need to be submitted electronically on ETL. Only PDF generated from LaTex is accepted.
- Make sure you prepare the answers to each question separately. This helps us dispatch the problems to different graders.
- Collaboration on solving the homework is allowed. Discussions are encouraged but you should think about the problems on your own.
- If you do collaborate with someone or use a book or website, you are expected to write up your solution independently. That is, close the book and all of your notes before starting to write up your solution.

1 Q1

1.1

Since $0, 1 - y_i \mathbf{w}^\top x_i$ are both convex and differentiable by $\mathbf{w}$ with $\nabla_{\mathbf{w}}(0) = 0, \nabla_{\mathbf{w}}(1 - y_i \mathbf{w}^\top x_i) = -y_i x_i$ subgradient of $\max(0, 1 - y_i \mathbf{w}^\top x_i)$ is as following:

$$\partial(\max(0, 1 - y_i \mathbf{w}^\top x_i)) = \begin{cases} \{0\} & (y_i \mathbf{w}^\top x_i > 1) \\ \{-\theta_i y_i x_i | 0 \leq \theta \leq 1\} & (y_i \mathbf{w}^\top x_i = 1) \\ \{-y_i x_i\} & (y_i \mathbf{w}^\top x_i < 1) \end{cases} \tag{1}$$

Also, $\frac{\lambda}{2} \|\mathbf{w}\|^2$ is subdifferentiable as it's differentiable and gradient is $\lambda \mathbf{w}$. Finally, the subgradient of the loss function can be derived by summing up results above:

$$\partial_{\mathbf{w}} \text{loss}(\mathbf{w}) = \left\{ -\frac{1}{n} \sum_{i=1}^{n} \theta_i y_i x_i + \lambda \mathbf{w} \middle| \theta_i \in \Theta_i \right\} \tag{2}$$

$$\text{where } \Theta_i = \begin{cases} \{0\} & (y_i \mathbf{w}^\top x_i > 1) \\ [0, 1] & (y_i \mathbf{w}^\top x_i = 1) \\ \{1\} & (y_i \mathbf{w}^\top x_i < 1) \end{cases} \tag{3}$$

1.2

Source code is attached in q2.ipynb. Plots are attached on Figure 1.

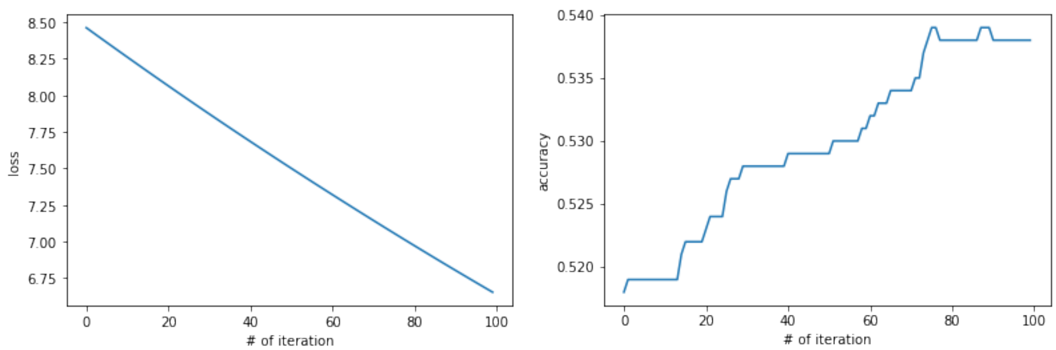

Figure 1: Result of gradient descent. b) iteration vs function value plot, c) iteration vs classification accuracy plot

2 Q2

2.1

Due to the constraint $\sum_i y_i\alpha_i = 0$, we can substitute α_1 with $-\sum_{i>1} y_i\alpha_i$. As $\alpha_3, \dots, \alpha_n$ are constants in the algorithm, reducing only α_1 from (1) by substitution results in the new objective function ℓ :

$$
\begin{aligned}
\ell(\alpha_2) &= \frac{1}{2}\alpha_2^2\left(2k(x_1,x_2) - k(x_1,x_1) - k(x_2,x_2)\right) && (4)\\
&+ \alpha_2\left(1 - y_1y_2 + y_2\sum_{i=3}^{n} y_i\alpha_i(k(x_i,x_1) - k(x_i,x_2) + k(x_1,x_2) - k(x_1,x_1))\right) && (5)
\end{aligned}
$$

Since $\sum_{i\geq 3} y_i\alpha_i = -y_1\alpha_1^{(t)} - y_2\alpha_2^{(t)}$,

$$
\begin{aligned}
\ell(\alpha_2) &= \frac{1}{2}\alpha_2^2\left(2k(x_1,x_2) - k(x_1,x_1) - k(x_2,x_2)\right) && (6)\\
&+ \alpha_2\left(1 - y_1y_2 + (\alpha_2^{(t)} + y_1y_2\alpha_1^{(t)})(k(x_1,x_1) - k(x_1,x_2)) + y_2\sum_{i=3}^{n} y_i\alpha_i(k(x_i,x_1) - k(x_i,x_2))\right) && (7)
\end{aligned}
$$

2.2

α_2 should satisfy $0 \leq \alpha_2 \leq C$, and we should consider the constraint of reduced variable α_1, too:

$$
\begin{aligned}
0 \leq \alpha_1 \leq C &\iff 0 \leq -\sum_{i=2}^{n} y_1y_i\alpha_i \leq C && (8)\\
&\iff -\sum_{i=3}^{n} y_1y_i\alpha_i - C \leq y_1y_2\alpha_2 \leq -\sum_{i=3}^{n} y_1y_i\alpha_i && (9)\\
&\iff \begin{cases} -\sum_{i=3}^{n} y_2y_i\alpha_i - C \leq \alpha_2 \leq -\sum_{i=3}^{n} y_2y_i\alpha_i & (y_1y_2 = 1)\\ -\sum_{i=3}^{n} y_2y_i\alpha_i \leq \alpha_2 \leq -\sum_{i=3}^{n} y_2y_i\alpha_i + C & (y_1y_2 = -1)\end{cases} && (10)
\end{aligned}
$$

Therefore, we can derive the value of L, U as following:

$$L = \begin{cases} \max(0, -\sum_{i=3}^{n} y_2 y_i \alpha_i - C) & (y_1 y_2 = 1) \\ \max(0, -\sum_{i=3}^{n} y_2 y_i \alpha_i) & (y_1 y_2 = -1) \end{cases} \tag{11}$$

$$U = \begin{cases} \min(C, -\sum_{i=3}^{n} y_2 y_i \alpha_i) & (y_1 y_2 = 1) \\ \min(C, -\sum_{i=3}^{n} y_2 y_i \alpha_i + C) & (y_1 y_2 = -1) \end{cases} \tag{12}$$

Since $\sum_{i \geq 3} y_i \alpha_i = -y_1 \alpha_1^{(t)} - y_2 \alpha_2^{(t)}$, we can simplify L, U as following:

$$L = \begin{cases} \max(0, \alpha_1^{(t)} + \alpha_2^{(t)} - C) & (y_1 y_2 = 1) \\ \max(0, -\alpha_1^{(t)} + \alpha_2^{(t)} & (y_1 y_2 = -1) \end{cases} \tag{13}$$

$$U = \begin{cases} \min(C, \alpha_1^{(t)} + \alpha_2^{(t)}) & (y_1 y_2 = 1) \\ \min(C, -\alpha_1^{(t)} + \alpha_2^{(t)} + C) & (y_1 y_2 = -1) \end{cases} \tag{14}$$

2.3

Twice differentiating (6), we can derive η:

$$\eta = 2k(x_1, x_2) - k(x_1, x_1) - k(x_2, x_2) \tag{15}$$

When $\eta < 0$, ℓ is a concave function and thus is maximized by α_2 satisfies $\frac{\partial \ell}{\partial \alpha_2} = 0$. Before differentiating ℓ, replacing summation in ℓ with E_1, E_2, we can simplify ℓ as following:

$$\ell(\alpha_2) = \frac{1}{2}\eta\alpha_2^2 + \alpha_2(y_2(E_1 - E_2) - \alpha_2^{(t)}\eta) \tag{16}$$

Now we can find α_2^* without constraint:

$$\frac{\partial \ell}{\partial \alpha_2} = \eta\alpha_2 + y_2(E_1 - E_2) - \alpha_2^{(t)}\eta = 0 \tag{17}$$

$$\alpha_2^* = \alpha_2^{(t)} - \frac{y_2(E_1 - E_2)}{\eta} \tag{18}$$

Since α_2 should satisfy constraints, the value should be clipped to be between L and U:

$$\therefore \alpha_2^* = \max\left(L, \min\left(U, \alpha_2^{(t)} - \frac{y_2(E_1 - E_2)}{\eta}\right)\right) \tag{19}$$

Corresponding α_1^* can be obtained by that $y_1\alpha_1 + y_2\alpha_2$ should be constant since $\sum_{i=1}^{n} y_i \alpha_i = 0$.

$$y_1\alpha_1^* + y_2\alpha_2^* = y_1\alpha_1^{(t)} + y_2\alpha_2^{(t)} \tag{20}$$

$$\therefore \alpha_1^* = \alpha_1^{(t)} + y_1 y_2(\alpha_2^{(t)} - \alpha_2^*) \tag{21}$$

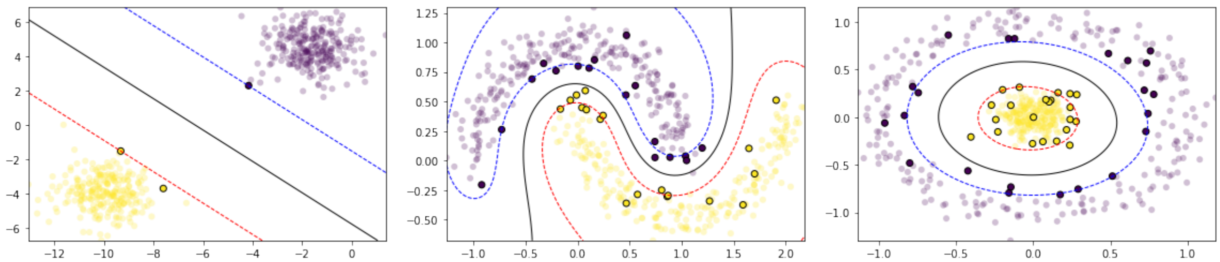

Figure 2: Result of running exp1, exp2, and exp3 in exp.py

2.4

Source code is attached in q2.py

2.5

Plots are attached on Figure 2. Circled points denotes support vectors, and each red/blue dotted lines denote the hyperplane $f(\mathbf{x}; \alpha) = 1$, $f(\mathbf{x}; \alpha) = -1$.

3 Q3

3.1

$$\begin{aligned} J_\lambda(\beta) &= \frac{1}{2}\|y - X\beta\|^2 + \lambda \|\beta\|_1 && (22) \\ &= \frac{1}{2}(y - X\beta)^\top (y - X\beta) + \lambda \|\beta\|_1 && (23) \\ &= \frac{1}{2}(y^\top y + \beta^\top \beta - 2y^\top X\beta) + \lambda \|\beta\|_1 && (24) \\ &= \frac{1}{2}\|y\|^2 + \sum_{i=1}^{d}\left(\frac{1}{2}\beta_i^2 - y^\top X_{.i}\beta_i + \lambda|\beta_i|\right) && (25) \end{aligned}$$

3.2

Assuming $\beta_j^* > 0$:

$$\begin{aligned} \beta_j^* &= \arg\min_{\beta_j} J_\lambda(\beta) && (26) \\ &= \arg\min_{\beta_j} f(X_{.j}, y, \beta_i, \lambda) && (27) \\ &= \arg\min_{\beta_j} \left(\frac{1}{2}\beta_j^2 - y^\top X_{.i}\beta_j + \lambda\beta_j\right) && (28) \\ &= y^\top X_{.j} - \lambda && (29) \end{aligned}$$

By assumption, $y^\top X_{.j} - \lambda > 0$ holds.

3.3

Assuming $\beta_j^* < 0$:

$$\beta_j^* = \arg\min_{\beta_j} J_\lambda(\beta) \tag{30}$$
$$= \arg\min_{\beta_j} f(X_{.j}, y, \beta_i, \lambda) \tag{31}$$
$$= \arg\min_{\beta_j} \left(\frac{1}{2}\beta_j^2 - y^\top X_{.i}\beta_j - \lambda\beta_j\right) \tag{32}$$
$$= y^\top X_{.j} + \lambda \tag{33}$$

By assumption, $y^\top X_{.j} + \lambda < 0$ holds.

3.4

Since $f(X_{.j}, y, 0, \lambda) = 0$ holds, to $\beta_j^* = 0$ to hold, $f \geq 0$ should hold for every β_j. As $f(X_{.j}, y, \beta_i, \lambda) = \frac{1}{2}\beta_j^2 - y^\top X_{.i}\beta_j + \lambda|\beta_j|$, following inequality should hold to satisfy $f \geq 0$ for all positive or negative β_j:

$$-\lambda \leq y^\top X_{.j} \leq \lambda \tag{34}$$

This condition implies that $y^\top X_{.j}$ is near 0 or $y, X_{.j}$ is almost perpendicular, which means that j-th feature of train data has little contribution to the output. In other words, β_j will be set to 0 if it has little effect to the output.

3.5

If regularization term is changed to L2-norm, following holds:

$$\beta_j^* = \arg\min_{\beta_j} J_\lambda(\beta) \tag{35}$$
$$= \arg\min_{\beta_j} f(X_{.j}, y, \beta_j, \lambda) \tag{36}$$
$$= \arg\min_{\beta_j} \left(\frac{1}{2}\beta_j^2 - y^\top X_{.j}\beta_j + \frac{1}{2}\lambda\beta_j^2\right) \tag{37}$$
$$= \frac{y^\top X_{.j}}{\lambda + 1} \tag{38}$$

Therefore, $y^\top X_{.j} = 0$ is the condition for $\beta_j^* = 0$ to be satisfied. This is much strict condition compared to the condition from Q3.4, as $y^\top X_{.j}$ should exactly be 0 in the ridge case, while it is okay to just be near 0 in the lasso case.