

Homework 5

Student Name: Andrew Choi

UID: 205348339

- Feel free to talk to other students in the class when doing the homework. You should, however, write down your solution yourself. You also must indicate on each homework with whom you collaborated and cite any other sources you use including Internet sites.
- You will write your solution in LaTeX and submit the **pdf file through Gradescope**. You also need to submit the **zipped LaTeX files to CCLE**. We will grade your homework based on the final version of the pdf file submitted to Gradescope. We will not grade the zipped Latex files on CCLE. However, failure to submitting your LaTeX files to CCLE will incur 2 points penalty out of 100 points.
- The homework (both pdf and zipped Latex source files) is due at **1:59 PM before the class**.

1. Exercise 15.1

First, as the hint indicates, we denote $\mathcal{G} = \{(w, b) : \forall i, y_i((w \bullet x_i) + b) > 0\}$. Hard-SVM implies that this set is non-empty. Note that for every $(w, b) \notin \mathcal{G}$, $y_i((w \bullet x_i) + b) \leq 0$ which implies the following.

$$\operatorname{argmax}_{(w,b): \|w\|=1} \min_{i \in [m]} y_i((w \bullet x_i) + b) \subseteq \mathcal{G}$$

This then implies that $\forall (w, b) \in \mathcal{G}$,

$$\min_{i \in [m]} y_i((w \bullet x_i) + b) = \min_{i \in [m]} |(w \bullet x_i) + b|$$

Therefore, both optimization problems are equivalent.

2. Exercise 15.2

For a training set S which is linearly separable with margin γ and is contained within a ball of radius ρ , the margin assumption implies that there exists $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that $y_i(w \bullet x_i) + b \geq \gamma \forall i \in [m]$. Assume that $\|w\| = 1$. This can then be normalized to the following.

$$y_i((w/\gamma) \bullet x_i) + b/\gamma \geq 1 \forall i \in [m]$$

From here we can simply use Theorem 9.1 to obtain the maximal number of updates. Let $w^* = w/\gamma$ which results in $\|w^*\| = 1/\gamma$. Then since $\rho = \max_i \|x_i\|$, the Perceptron algorithm stops after at most $(\rho/\gamma)^2$ iterations.

3. Exercise 16.2

For Kernelized Perceptron, denote a vector $\alpha^{(t)} \in \mathbb{R}^m$. This vector will be updated at each time step t , such that

$$w^{(t)} = \sum_{i=1}^m \alpha_i^{(t)} \psi(x_i)$$

Now, note that the update index selection rule for Kernalized Perceptron is i s.t. $y_i(w \bullet \psi(x_i)) \leq 0$. Using the Representer Theorem (16.1), this is equivalent to finding i s.t. $y_i \sum_{j=1}^m \alpha_j^{(t)} K(x_i, x_j) \leq 0$ which satisfies the limitation of only accessing instances via the kernel function. The update rule $w^{(t+1)} = w^{(t)} + y_i x_i$ can be represented as $\alpha^{(t+1)} = \alpha^{(t)} + y_i e_i$. Finally, denote $\alpha^{(T)}$ as the final output of the algorithm. Then, given a new instance x , we can represent the prediction rule $\hat{y} = \text{sign}(w^{(T)} \bullet \psi(x))$ with the following equation below using Theorem 16.1 once more.

$$\hat{y} = \text{sign} \left(\sum_{i=1}^m \alpha_i^{(T)} K(x_i, x) \right)$$

4. Exercise 16.3

- a. As the hint indicates, the Representer Theorem tells us that exists a vector $\alpha^* \in \mathbb{R}^m$ such that $w^* = \sum_{i=1}^m \alpha_i \psi(x_i)$ is a minimizer of $f(w)$. We can then use this definition and Equation 16.8 to arrive at the following ERM objective.

$$\min_{\alpha \in \mathbb{R}^m} \lambda \left\| \sum_{i=1}^m \alpha_i \psi(x_i) \right\|^2 + \frac{1}{2m} \sum_{i=1}^m \left(\left(\sum_{i=1}^m \alpha_i \psi(x_i) \right) \bullet \psi(x_i) - y_i \right)^2$$

It can be seen that the above equation is equivalent to the ERM objective of $g(\alpha)$ shown below.

$$\min_{\alpha \in \mathbb{R}^m} \lambda \alpha^T G \alpha + \frac{1}{2m} \sum_{i=1}^m ((\alpha \bullet G_{:,i}) - y_i)^2$$

This proves that if α^* minimizes Equation 16.9, it also minimizes Equation 16.8.

- b. To find a closed form expression for α^* , first, we rearrange the equation from part a. to the following form.

$$\min_{\alpha \in \mathbb{R}^m} \lambda \alpha^T G \alpha + \frac{1}{2m} (G \alpha - y)^T (G \alpha - y)$$

As this equation is convex, we can then solve for α using the gradient of the rearranged equation with respect to α set to 0. Denote the equation above as $f(\alpha)$ and $2m\lambda$ as λ' .

$$\begin{aligned} \nabla_{\alpha} f(\alpha) &= \lambda' G \alpha + G G^T \alpha - G y = 0 \\ (\lambda' G + G G^T) \alpha^* - G y &= 0 \end{aligned}$$

This can be rewritten as the following as G is symmetric.

$$\begin{aligned} G(\lambda' I + G) \alpha &= G y \\ (\lambda' I + G) \alpha &= y \end{aligned}$$

Finally, since $m\lambda > 0$ and G is positive semi-definite, we know that $m\lambda I + G$ is positive definitive and therefore invertible. Therefore, we can arrive at the closed form expression below.

$$\alpha^* = (\lambda' I + G)^{-1} y$$

5. Exercise **16.4**

Consider the mapping where $\psi(i)$ produces a vector $(1^i; 0^{N-i}) \in \mathbb{R}^N$ where the first i elements are 1 and the remaining $N - 1$ elements are 0. Then, by computing the inner product $(\psi(i) \bullet \psi(j))$, we see that the output will be the smaller value between i and j . In other words,

$$(\psi(i) \bullet \psi(j)) = \min\{i, j\} = K(i, j)$$

Therefore, K is a valid kernel.

6. Exercise **16.6**

- a. First we translate the piecewise function $h(x)$ into the analytical format shown below.

$$h(x) = \text{sign}(\|\psi(x) - c_-\|^2 - \|\psi(x) - c_+\|^2)$$

Using this definition, we then derive the proof as follows.

$$\begin{aligned} h(x) &= \text{sign}(\|\psi(x) - c_-\|^2 - \|\psi(x) - c_+\|^2) \\ &= \text{sign}(2(\psi(x) \bullet c_+) - 2(\psi(x) \bullet c_-) + \|c_-\|^2 - \|c_+\|^2) \\ &= \text{sign}(2(\psi(x) \bullet w) + 2b) \\ &= \text{sign}((\psi(x) \bullet w) + b) \end{aligned}$$

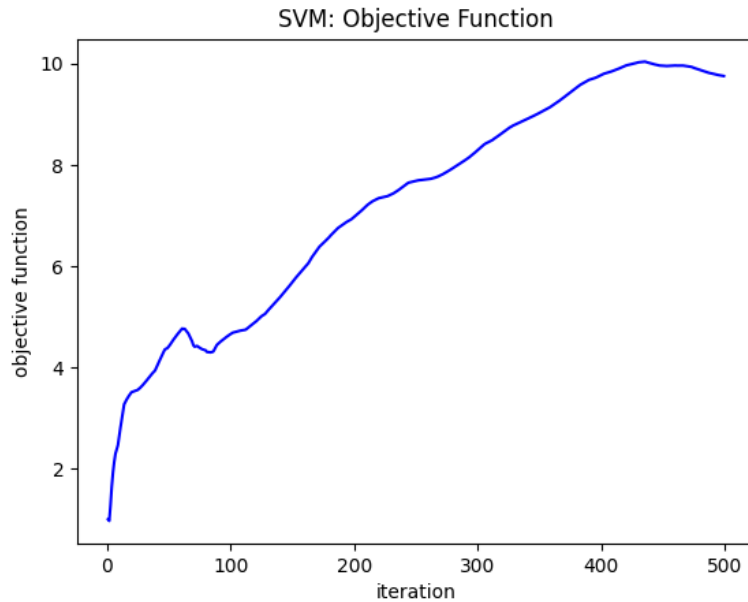
- b. $h(x)$ can be expressed using the kernel function and without accessing individual entries of $\psi(x)$ or w by simply replacing the dot product $(\psi(x) \bullet w)$ shown in the definition above with the equality shown below where we use the definition $w = c_+ - c_-$.

$$\begin{aligned} (\psi(x) \bullet w) &= (\psi(x) \bullet (c_+ - c_-)) \\ &= \frac{1}{m_+} \sum_{i:y_i=y} (\psi(x) \bullet \psi(x_i)) - \frac{1}{m_-} \sum_{i:y_i \neq y} (\psi(x) \bullet \psi(x_i)) \\ &= \frac{1}{m_+} \sum_{i:y_i=y} K(x, x_i) - \frac{1}{m_-} \sum_{i:y_i \neq y} K(x, x_i) \end{aligned}$$

7. The training and testing errors obtained by **scikit-learn's Linear SVM model** are listed below.

- Training error: 0.183
- Testing error: 0.125

Below is a plot of the objective value $J(\bar{w}^{(t)})$ with respect to the number of iterations t during training.



The training and testing errors obtained by **my own Linear SVM model** are listed below.

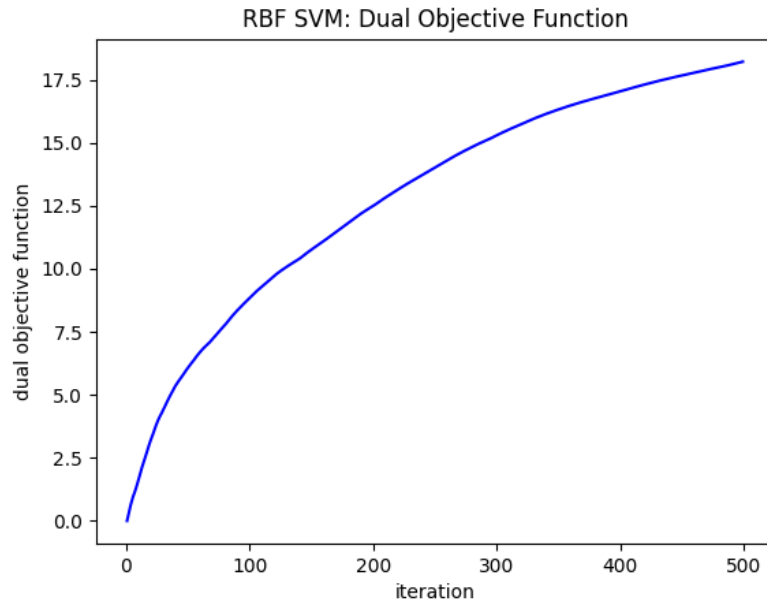
- Training error: 0.167
- Testing error: 0.100

The hyperparameters $\lambda = 2.0$ and $T = 500$ were used as they resulted in improved performance from scikit-learn's implementation.

8. The training and testing error obtained by **scikit-learn's Soft-SVM model with RBF kernel** are listed below.

- Training error: 0.067
- Testing error: 0.025

Below is a plot of the objective function of dual SVM $\Theta(\bar{\alpha}^{(t)})$ with respect to the number of iterations t during training.



The training and testing errors obtained by **my own RBF SVM model** are listed below.

- Training error: 0.017
- Testing error: 0.025

The hyperparameters were set to $\gamma = 2$, $\lambda = 2.0$, and $T = 500$ as they resulted in an identical testing error with scikit-learn's implementation.