Machine Learning (CS 6140) Homework 2

Xiao Wang

November 9, 2017

Problem 1

A. If $\boldsymbol{w}^\top \boldsymbol{x} > \mathbf{0}$, the sample is labeled as 1, otherwise -1.

Model 1

$$\begin{cases} w_1 + w_2 > 0 \\ w_1 < 0 \end{cases} \implies \begin{cases} w_1 + w_2 > 0 \\ w_1 < 0 \end{cases}$$

The constraints on $\boldsymbol{w}$ stay the same, so $\boldsymbol{w}$ doesn't change.

Model 2

$$\begin{cases} w_0 + w_1 + w_2 > 0 \\ w_0 + w_1 < 0 \\ w_0 > 0 \end{cases} \implies \begin{cases} w_0 + w_1 + w_2 > 0 \\ w_0 + w_1 < 0 \\ w_0 < 0 \end{cases}$$

From the constraints above we see that w_0 changes from positive to negative.

B. For large λ:

$$l(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N} y^i \boldsymbol{w}^\top \boldsymbol{x}^{\boldsymbol{i}} - \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$$

$$\frac{\partial l(\boldsymbol{w})}{\partial \boldsymbol{w}} = \frac{1}{2}\sum_{i=1}^{N} y^i \boldsymbol{x}^{\boldsymbol{i}} - \lambda \boldsymbol{w} = 0$$

$$\boldsymbol{w} = \frac{\sum_{i=1}^{N} y^i \boldsymbol{x}^{\boldsymbol{i}}}{2\lambda}$$

From the result above, we can see that, as λ increases, the absolute value of each term of $\boldsymbol{w}$ decreases.

Problem 2

A. **No.**
Because if $t \leq -1$, there will be mixed instances in class $+1$; if $-1 < t \leq 1$, there will be mixed instances in both classes; if $t > 1$, there will be mixed instances in class -1. Thus there is no $t \in \mathbb{R}$ that results in 0 error.

B. **Yes.**
The equation of the max-margin separator is $x_2 - 2.5 \geq 0$. The corresponding margin is 1.5.

C.
$$K(x, z) = \boldsymbol{\phi}(x) \cdot \boldsymbol{\phi}(z) = xz + x^2 z^2$$

Problem 3

A. Kernel matrix $\boldsymbol{K} = c_1 \boldsymbol{K_1} + c_2 \boldsymbol{K_2}$
$\because c_1, c_2 > 0, \boldsymbol{a}^\top \boldsymbol{K_1 a}, \boldsymbol{a}^\top \boldsymbol{K_2 a} \geq 0$
$\therefore \forall \boldsymbol{a} \in \mathbb{R}^n, \boldsymbol{a}^\top \boldsymbol{K a} = c_1 \boldsymbol{a}^\top \boldsymbol{K_1 a} + c_2 \boldsymbol{a}^\top \boldsymbol{K_2 a} \geq 0$
$\therefore$ It's still a kernel function.

B. By construction, the Kernel matrix is given by $\boldsymbol{K} = \boldsymbol{K_1} \odot \boldsymbol{K_2}$, where $\odot$ denotes the Hadamard (entrywise) product.
Given that $\boldsymbol{K_1}$ and $\boldsymbol{K_2}$ are symmetric positive semi-definite matrices, their eigendecompositions $\boldsymbol{K_1} = \sum_{i=1}^n \lambda_i \boldsymbol{u_i u_i}^\top, \boldsymbol{K_2} = \sum_{j=1}^n \mu_j \boldsymbol{v_j v_j}^\top$ have positive eigenvalues $\lambda_i \geq 0$ and $\mu_i \geq 0$.
$\boldsymbol{K} = \sum_{i=1}^n \sum_{j=1}^n \lambda_i \mu_j (\boldsymbol{u_i u_i}^\top) \odot (\boldsymbol{v_j v_j}^\top) = \sum_{i=1}^n \sum_{j=1}^n \lambda_i \mu_j (\boldsymbol{u_i} \odot \boldsymbol{v_j})(\boldsymbol{u_i} \odot \boldsymbol{v_j})^\top = \sum_{k=1}^{n^2} \gamma_k \boldsymbol{w_k w_k}^\top$ with $\gamma_k = \lambda_{\lfloor k/n \rfloor} \mu_{k \ mod \ n} \geq 0$ and $\boldsymbol{w_k} = \boldsymbol{u}_{\lfloor \boldsymbol{k/n} \rfloor} \odot \boldsymbol{v_{k \ mod \ n}}$.
$\therefore \forall \boldsymbol{a} \in \mathbb{R}^n, \quad \boldsymbol{a^T K a} = \sum_{k=1}^{n^2} \gamma_k \boldsymbol{a^T w_k w_k}^\top \boldsymbol{a} = \sum_{k=1}^{n^2} \gamma_k (\boldsymbol{w_k}^\top \boldsymbol{a})^2 \geq 0$
$\therefore$ It's still a kernel function.

C. Since the powers of $\boldsymbol{K_1}$ are products of $\boldsymbol{K_1}$ by itself and thus valid kernels, their linear combination is also a valid kernel.

D. We can expand the exponential function with Taylor series which is essentially a polynomial function in Question C. Thus it's also a valid kernel.

E. Because the kernel matrix is also semidefinite, $K(x, z) = x^\top A z$ is also a valid kernel.

F.
$$\begin{aligned} k(\boldsymbol{x}, \boldsymbol{z}) &= \exp\left(-\|\boldsymbol{x} - \boldsymbol{z}\|_2^2\right) = \exp\left(-\|\boldsymbol{x}\|_2^2 - \|\boldsymbol{z}\|_2^2 + 2\boldsymbol{x}^\top \boldsymbol{z}\right) \\ &= \left[\exp\left(-\|\boldsymbol{x}\|_2^2\right) \exp\left(-\|\boldsymbol{z}\|_2^2\right)\right] \exp\left(2\boldsymbol{x}^\top \boldsymbol{z}\right) \end{aligned}$$

Because $g(\boldsymbol{x})g(\boldsymbol{z})$ is a kernel; $\exp(k_1(\boldsymbol{x}, \boldsymbol{z}))$ is a kernel according to Question D; the production of two kernels is kernel according to Question B. Thus this is also a valid kernel.

Problem 4

A. number of misclassified instances $\leq \sum_{i=1}^{n} \lfloor \xi_i \rfloor$

B. C is a regularization parameter:
- $C \to 0$ allows constraints to be easily ignored $\to$ large margin
- large C makes constraints hard to ignore $\to$ narrow margin
- $C \to \infty$ enforces all constraints $\to$ hard margin

C. During training,

$$\begin{aligned}\text{minimize}_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_j) \\ = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)\end{aligned}$$

Find some i s.t. $0 < \alpha_i < C$, so that $\phi(\boldsymbol{x}_i)$ lies on the boundary of the margin in the transformed space, and then solve

$$\begin{aligned}\boldsymbol{b} = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x_i}) - \boldsymbol{y_i} &= \left[\sum_{k=1}^{n} \alpha_k y_k \boldsymbol{\phi}(\boldsymbol{x_k})^\top \boldsymbol{\phi}(\boldsymbol{x_i}) \right] - \boldsymbol{y_i} \\ &= \left[\sum_{k=1}^{n} \alpha_k y_k K(\boldsymbol{x_k}, \boldsymbol{x_i}) \right] - \boldsymbol{y_i}\end{aligned}$$

During prediction,

$$\boldsymbol{x} \mapsto \operatorname{sgn} \left[\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{b} \right] = \operatorname{sgn} \left\{ \left[\sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x_i}, \boldsymbol{x}) \right] - \boldsymbol{b} \right\}$$

Problem 5

Define an infinite step function:

$$I(u) = \begin{cases} 0 & \text{, if } u \le 0 \\ \infty & \text{, otherwise} \end{cases}$$

Define objective function as following:

$$\begin{aligned} J(\boldsymbol{w}) &= \begin{cases} f(\boldsymbol{w}) & \text{, if } g_j(\boldsymbol{w}) \le 0 \\ \infty & \text{, otherwise} \end{cases} \\ &= f(\boldsymbol{w}) + \sum_j I\left(g_j(\boldsymbol{w})\right) \end{aligned}$$

$$\begin{gathered} \because \alpha_j g_j(\boldsymbol{w}) \text{ is a lower bound of } I\left(g_j(\boldsymbol{w})\right) \\ \therefore L(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \text{ is a lower bound of } J(\boldsymbol{w}) \\ \therefore L(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \le J(\boldsymbol{w}), \forall \boldsymbol{a} > \boldsymbol{0} \\ \therefore \min_w L(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \le \min_w J(\boldsymbol{w}) \\ \therefore \max_{\alpha \ge 0, \beta} \min_w L(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \le \min_w \max_{\alpha \ge 0, \beta} L(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \end{gathered}$$

Problem 6

A.

$$L(\boldsymbol{x}, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{P_0 x} + \boldsymbol{q_0}^\top \boldsymbol{x} + r_0 + \sum_{i=1}^{m} \alpha_i \left(\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{P_i x} + \boldsymbol{q_i}^\top \boldsymbol{x} + r_i \right)$$

B.

$$\begin{gathered} g(\boldsymbol{\alpha}) = \min_x L(\boldsymbol{x}, \boldsymbol{\alpha}) \\ \frac{\partial L(\boldsymbol{x}, \boldsymbol{\alpha})}{\partial x} = \boldsymbol{P_0 x} + \boldsymbol{q_0} + \sum_{i=1}^{m} \alpha_i \left(\boldsymbol{P_i x} + \boldsymbol{q_i} \right) = \boldsymbol{0} \\ \boldsymbol{x = -\left(P_0 + \sum_{i=1}^{m} \alpha_i P_i \right)^{-1} \left(q_0 + \sum_{i=1}^{m} \alpha_i q_i \right)} \end{gathered}$$

C.

Problem 7

A. Please see code attached in email.

B. **Data1**

```
Optimization terminated successfully.
         Current function value: 0.066343
         Iterations: 15
         Function evaluations: 16
         Gradient evaluations: 16
Weight Vec:  [ 2.79426699 -1.09593905]
Training Err:  0.0
Test Err:  0.0
```

C. **Data2**

```
Optimization terminated successfully.
         Current function value: 0.177556
         Iterations: 13
         Function evaluations: 14
         Gradient evaluations: 14
Weight Vec:  [ 3.50145858 -0.08341908]
Training Err:  0.047619047619
Test Err:  0.0
```

Dataset2 has bigger training error than dataset1 because, as we can see from the graphs, the samples in training set 2 are not linearly separable.

Problem 8

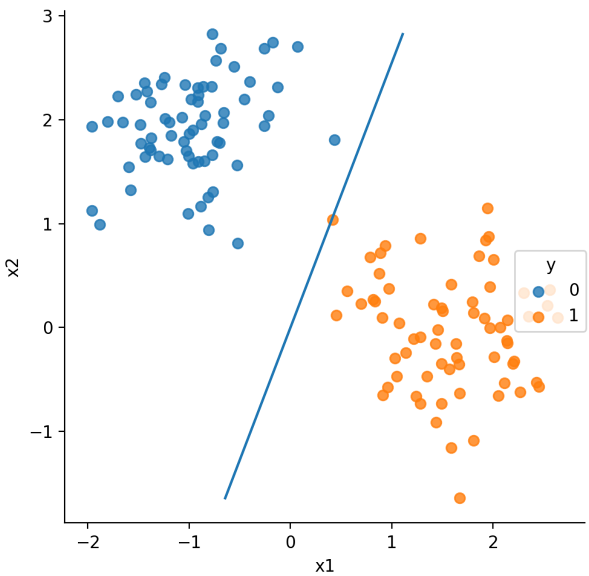
A. **Data1**

```
Training Err:  0.007352941176
Test Err:  0.0
```

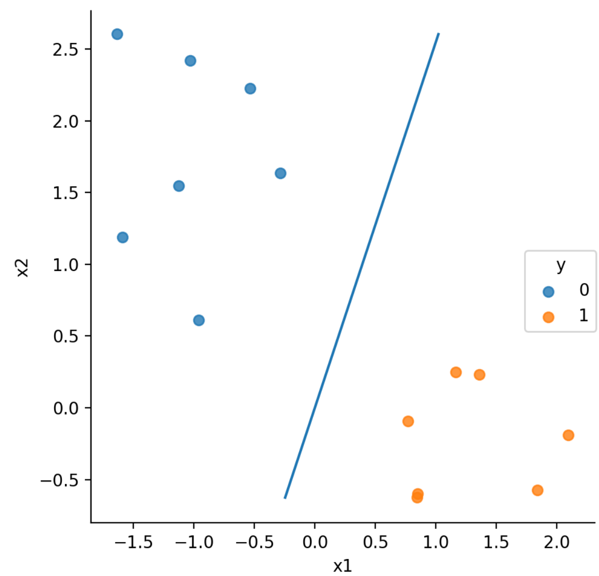
B. **Data2**

```
Training Err:  0.023809523810
Test Err:  0.0
```

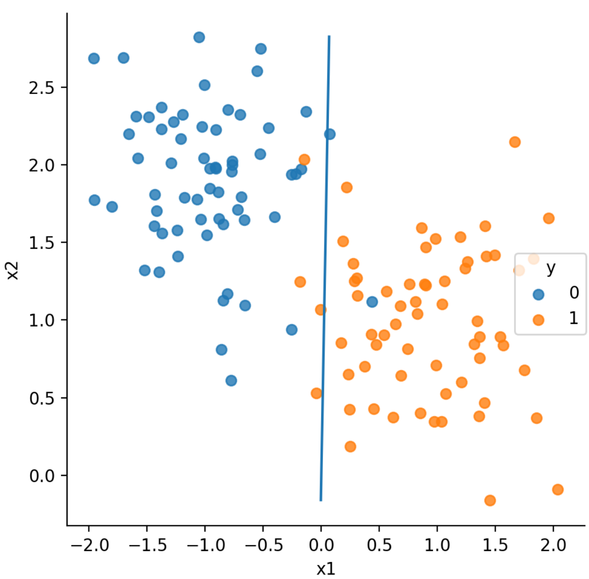
Dataset2 has bigger training error than dataset1 because, as we can see from the graphs, the samples in training set 2 are not linearly separable.



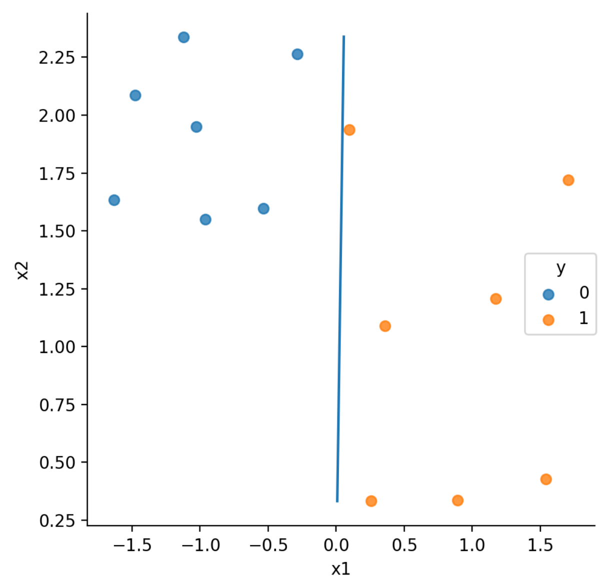
(a) Training set 1



(b) Test set 1



(c) Training set 2



(d) Test set 2