# Machine Learning Course 2024 Spring: Homework 2

March 31, 2024

## 1 Problem 1

**Solution:**

1. A linear SVM with $C = 0.02$.

   Solution: Correspond to Figure 1.4. The decision boundary of linear SVM is linear. In comparison with Figure 1.3 , the line does not separate two classes strictly, which corresponds to the case $C$ is small and more errors are allowed.

2. A linear SVM with $C = 20$.

   Solution: Correspond to Figure 1.3. The decision boundary of linear SVM is linear. In comparison with Figure 1.4, the line separates two classes strictly, which corresponds to the case $C$ is big.

3. A hard-margin kernel SVM with $\kappa(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} + \left(\boldsymbol{u}^\top \boldsymbol{v}\right)^2$.

   Solution: Correspond to Figure 1.5. The decision boundary of quadratic kernel is given by $f(\boldsymbol{x}) = \sum_i \alpha_i \left(\boldsymbol{x}_i^\top \boldsymbol{x} + \left(\boldsymbol{x}_i^\top \boldsymbol{x}\right)^2\right) + b$ . Hence the decision boundary is $f(\boldsymbol{x}) = 0$ . Since $f(\boldsymbol{x})$ is second order function of $\boldsymbol{x}$, the curve can be ellipse or hyperbolic curve. Figure 1.5 is hyperbolic curve.

4. A hard-margin kernel SVM with $\kappa(\boldsymbol{u}, \boldsymbol{v}) = \exp\left(-5\|\boldsymbol{u} - \boldsymbol{v}\|^2\right)$ .

   Solution: Correspond to Figure 1.6. We can write out the decision function as $f(\boldsymbol{x}) = \sum_i \alpha_i \exp\left(-\gamma \left\|\boldsymbol{x}_i - \boldsymbol{x}\right\|^2\right) + b$ . If $\gamma$ is large, the kernel value is quite small even if the distance between the $\boldsymbol{x}$ and $\boldsymbol{x}_i$ is small. This makes the classification hard with few supporting vectors, Hence Figure 1.6 corresponds to $\gamma = 5$ .

5. A hard-margin kernel SVM with $\kappa(\boldsymbol{u}, \boldsymbol{v}) = \exp\left(-\frac{1}{5}\|\boldsymbol{u} - \boldsymbol{v}\|^2\right)$ .

   Solution: Correspond to Figure 1.1.

□

# 2 Problem 2

**Solution:**

1. Construct the Lagrangian:

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \mu) = -\sum_{i=1}^{n} \log(\alpha_i + x_i) + \sum_{i=1}^{n} \lambda_i(-x_i) + \mu(\sum_{i=1}^{n} x_i - 1), \tag{1}$$

   where $\boldsymbol{x} = [x_1, \ldots, x_n]^\top, \boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_n]^\top \in \mathbb{R}^n$, and $\mu \in \mathbb{R}$.

   The KKT conditions are:

   (1) $-x_i^* \leq 0, i = 1, \ldots, n$

   (2) $\sum_{i=1}^{n} x_i^* - 1 = 0$

   (3) $\lambda_i^* \geq 0, i = 1, \ldots, n$

   (4) $\lambda_i^* x_i^* = 0, i = 1, \ldots, n$

   (5) $\nabla_{x_i^*} L(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \mu^*) = -\frac{1}{\alpha_i + x_i^*} - \lambda_i^* + \mu^* = 0, i = 1, \ldots, n$

2. According to the KKT conditions (3)-(5), we have

$$x_i^*(\mu^* - \frac{1}{\alpha_i + x_i^*}) = 0, i = 1, \ldots, n \tag{2}$$

$$\mu^* \geq \frac{1}{\alpha_i + x_i^*}, i = 1, \ldots, n \tag{3}$$

   (1) If $\mu^* < \frac{1}{\alpha_i}$, Eq.(3) is satisfied when $x_i^* > 0$. Then, according to Eq.(2), we have $x_i^* = \frac{1}{\mu^*} - \alpha_i$;

   (2) If $\mu^* \geq \frac{1}{\alpha_i}$, according to Eq.(2), we have $x_i^* = 0$.

Thus we have,

$$x_i^* = \max\{0, \frac{1}{\mu^*} - \alpha_i\}. \tag{4}$$

Substituting Eq.(4) into the condition (2), we can obtain

$$\sum_{i=1}^{n} \max\{0, \frac{1}{\mu^*} - \alpha_i\} = 1. \tag{5}$$

Therefore, the optimal solution of the problem is $x_i^* = \max\{0, \frac{1}{\mu^*} - \alpha_i\}, i = 1, \ldots, n$, where $\mu^*$ satisfies Eq.(5)

□

# 3 Problem 3

1. **Solution:**

$$\begin{aligned} & ||\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)||^2 \\ = & \phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_i) + \phi(\boldsymbol{x}_j)^\top \phi(\boldsymbol{x}_j) - 2\phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_j) \\ = & \kappa(\boldsymbol{x}_i, \boldsymbol{x}_i) + \kappa(\boldsymbol{x}_j, \boldsymbol{x}_j) - 2\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ = & 1 + 1 - 2\exp(-\frac{1}{2}||\boldsymbol{x}_i - \boldsymbol{x}_j||^2) \\ < & 2 \end{aligned}$$

□

2. **Proof:**

Consider a simpler kernel function $\kappa'(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^\top \boldsymbol{x}_j + c$ first. For any dataset $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m\}$, the kernel matrix $\mathbf{K}'$ is:

$$\mathbf{K}' = \begin{pmatrix} \kappa'(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & \kappa'(\boldsymbol{x}_1, \boldsymbol{x}_m) \\ \vdots & \ddots & \vdots \\ \kappa'(\boldsymbol{x}_m, \boldsymbol{x}_1) & \cdots & \kappa'(\boldsymbol{x}_m, \boldsymbol{x}_m) \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^\top \boldsymbol{x}_1 + c & \cdots & \boldsymbol{x}_1^\top \boldsymbol{x}_m + c \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}_m^\top \boldsymbol{x}_1 + c & \cdots & \boldsymbol{x}_m^\top \boldsymbol{x}_m + c \end{pmatrix}. \tag{6}$$

The kernel matrix can be further rewritten as

$$\begin{aligned}\mathbf{K}' &= \begin{pmatrix} \boldsymbol{x}_1^\top \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_1^\top \boldsymbol{x}_m \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}_m^\top \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_m^\top \boldsymbol{x}_m \end{pmatrix} + \begin{pmatrix} c & \cdots & c \\ \vdots & \ddots & \vdots \\ c & \cdots & c \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_m \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_m \end{pmatrix} + c \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \\ &= \mathbf{X}\mathbf{X}^\top + c\mathbf{1}\mathbf{1}^\top, \end{aligned} \tag{7}$$

where $\mathbf{X} \in \mathbb{R}^{m \times d}$ is the data matrix. Considering the positive definiteness of matrix $\mathbf{K}'$, for any non-zero vector $\boldsymbol{w} \in \mathbb{R}^m$, we have

$$\begin{aligned}\boldsymbol{w}^\top \mathbf{K}' \boldsymbol{w} &= \boldsymbol{w}^\top (\mathbf{X}\mathbf{X}^\top + c\mathbf{1}\mathbf{1}^\top)\boldsymbol{w} \\ &= \boldsymbol{w}^\top \mathbf{X}\mathbf{X}^\top \boldsymbol{w} + c\boldsymbol{w}^\top \mathbf{1}\mathbf{1}^\top \boldsymbol{w} \\ &= (\mathbf{X}^\top \boldsymbol{w})^\top (\mathbf{X}^\top \boldsymbol{w}) + c(\mathbf{1}^\top \boldsymbol{w})^\top (\mathbf{1}^\top \boldsymbol{w}) \\ &= ||\mathbf{X}^\top \boldsymbol{w}||^2 + c(\mathbf{1}^\top \boldsymbol{w})^2. \end{aligned} \tag{8}$$

(a) When $c \geq 0$, we have $\boldsymbol{w}^\top \mathbf{K}' \boldsymbol{w} \geq 0$, thus, $\mathbf{K}'$ is semi-positive definite. According to Mercer's Theorem (refer to page 55 of Lecture 6 slides), $\kappa'(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is a kernel function. According to the kernels' composition rule (refer to page 57 of Lecture 6 slides), $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\kappa'(\boldsymbol{x}_i, \boldsymbol{x}_j))^N$ is also a kernel function as a multiplication of $N$ kernel functions.

(b) When $c < 0$, it cannot be guaranteed that $\boldsymbol{w}^\top \mathbf{K}' \boldsymbol{w} \geq 0$, thus, $\mathbf{K}'$ is not semi-positive definite. Consequently, $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is not a kernel function.

□