

HW5 Handwritten Assignment

Lecturer: Pei-Yuan Wu

TAs: Chun-Lin Huang(Problem 1, 2), Yuan-Chia Chang(Problem 3, 4, 5)

December 2023, First Edition

Problem 1 (Kernel)(0.5%)

Consider the following data points:

- $c_1 = \{(3, 3), (3, -3), (-3, 3), (-3, -3)\}$
- $c_2 = \{(6, 6), (6, -6), (-6, 6), (-6, -6)\}$

The data are not linearly separable in this case. Write down a feature map and kernel function to transform the data into a new space, in which the data are linearly separable. Note that you do not just give me a feature map; please explain why.

Problem 2 (SVM with Gaussian kernel)(1.5%)

Consider the task of training a support vector machine using the Gaussian kernel $K(x, z) = \exp(-\frac{\|x-z\|^2}{\tau^2})$. We will show that as long as there are no two identical points in the training set, we can always find a value for the bandwidth parameter τ such that the SVM achieves zero training error.

Recall from class that the decision function learned by the support vector machine can be written as

$$f(x) = \sum_{i=1}^N \alpha_i y_i k(x_i, x) + b$$

Assume that the training data $\{(x_1, y_1), \dots, (x_N, y_N)\}$ consists of points which are separated by at least distance of ϵ ; that is, $\|x_j - x_i\| \geq \epsilon$, for any $i \neq j$. For simplicity, we assume $\alpha_i = 1$ for all $i = 1, \dots, m$ and $b = 0$. Find values for the Gaussian kernel width τ such that x_i is correctly classified, for all $i = 1, \dots, N$, e.g., $f(x_i)y_i > 0$ for all $i = 1, \dots, N$.

Hint: Notice that for $y \in \{-1, +1\}$ the prediction on x_i will be correct if $|f(x_i) - y_i| < 1$, so find a value of τ that satisfies this inequality for all i .

Problem 3 (Support Vector Regression)(2%)

Suppose we are given a training set $\{(x_1, y_1), \dots, (x_m, y_m)\}$, where $x_i \in \mathbb{R}^{(n+1)}$ and $y_i \in \mathbb{R}$. We would like to find a hypothesis of the form $f(x) = w^T x + b$. It is possible that no such function $f(x)$ exists to satisfy these constraints for all points. To deal with otherwise infeasible constraints, we introduce slack variables ξ_i for each point. The (convex) optimization problem is

$$\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i \tag{1}$$

$$\text{s.t. } y_i - w^T x_i - b \leq \epsilon + \xi_i \quad i = 1, \dots, m \tag{2}$$

$$w^T x_i + b - y_i \leq \epsilon + \xi_i \quad i = 1, \dots, m \tag{3}$$

$$\xi_i \geq 0 \quad i = 1, \dots, m \tag{4}$$

where $\epsilon > 0$ is a given, fixed value and $C > 0$. Denote that $\xi = (\xi_1, \dots, \xi_m)$.

- (a) Write down the Lagrangian for the optimization problem above. Consider the sets of Lagrange multiplier $\alpha_i, \alpha_i^*, \beta_i$ corresponding to the (2), (3), and (4), so that the Lagrangian would be written as $\mathcal{L}(w, b, \xi, \alpha, \alpha^*, \beta)$, where $\alpha = (\alpha_1, \dots, \alpha_m)$, $\alpha^* = (\alpha_1^*, \dots, \alpha_m^*)$, and $\beta = (\beta_1, \dots, \beta_m)$.
- (b) Derive the dual optimization problem. You will have to take derivatives of the Lagrangian with respect to w, b , and ξ
- (c) Suppose that $(\bar{w}, \bar{b}, \bar{\xi})$ and $(\bar{\alpha}, \bar{\alpha}^*, \bar{\beta})$ are the optimal solutions to a primal and dual optimization problem, respectively.

Denote $\bar{w} = \sum_{i=1}^m (\bar{\alpha}_i - \bar{\alpha}_i^*) x_i$

- (1) Prove that

$$\bar{b} = \arg \min_{b \in \mathbb{R}} C \sum_{i=1}^m \max(|y_i - (\bar{w}^T x_i + b)| - \epsilon, 0) \quad (5)$$

- (2) Define $e = y_i - (\bar{w}^T x_i + \bar{b})$ Prove that

$$\begin{cases} \bar{\alpha}_i = \bar{\alpha}_i^* = 0, & \bar{\xi}_i = 0, & \text{if } |e| < \epsilon \\ 0 \leq \bar{\alpha}_i \leq C, & \bar{\xi}_i = 0, & \text{if } e = \epsilon \\ 0 \leq \bar{\alpha}_i^* \leq C, & \bar{\xi}_i = 0, & \text{if } e = -\epsilon \\ \bar{\alpha}_i = C, & \bar{\xi}_i = e - \epsilon & \text{if } e > \epsilon \\ \bar{\alpha}_i^* = C, & \bar{\xi}_i = -(e + \epsilon) & \text{if } e < -\epsilon \end{cases} \quad (6)$$

- (d) Show that the algorithm can be kernelized and write down the kernel form of the decision function. For this, you have to show that

- (1) The dual optimization objective can be written in terms of inner products or training examples
- (2) At test time, given a new x the hypothesis $f(x)$ can also be computed in terms of inner products.

Problem 4 (Sparse SVM)(2%)

Given training data of N input-output pairs $\mathcal{D} = ((x_i, y_i))_{i=1}^N$, where $x_i \in \mathcal{X}$ and $y_i \in \{\pm 1\}$. One can give two types of arguments in favor of the SVM algorithm: one based on the sparsity of the support vectors, another based on the notion of margin. Suppose instead of maximizing the margin, we choose instead to maximize sparsity by minimizing the p -norm of the vector $\alpha = (\alpha_1, \dots, \alpha_N)$ that defines the weight vector \mathbf{w} , for some $p \geq 1$. In this question we consider the case $p = 2$, which leads to the following optimization problem:

$$\begin{aligned} & \text{minimize} && f(\alpha, b, \xi) = \frac{1}{2} \sum_{i=1}^N \alpha_i^2 + \sum_{i=1}^N C_i \xi_i \\ & \text{subject to} && y_i \left(\sum_{j=1}^N \alpha_j y_j \mathbf{x}_i \cdot \mathbf{x}_j + b \right) \geq 1 - \xi_i, \quad i \in 1, N \\ & \text{variables} && b \in \mathbb{R}, \alpha_i \geq 0, \xi_i \geq 0, \quad i \in 1, N \end{aligned}$$

which can be rewritten in the following primal problem:

$$\begin{aligned} & \text{minimize} && f(\alpha, b, \xi) = \frac{1}{2} \sum_{i=1}^N \alpha_i^2 + \sum_{i=1}^N C_i \xi_i \\ & \text{subject to} && \left. \begin{aligned} g_{1,i}(\alpha, b, \xi) &= 1 - \xi_i - y_i \left(\sum_{j=1}^N \alpha_j y_j \mathbf{x}_i \cdot \mathbf{x}_j + b \right) \leq 0 \\ g_{2,i}(\alpha, b, \xi) &= -\alpha_i \leq 0 \\ g_{3,i}(\alpha, b, \xi) &= -\xi_i \leq 0 \end{aligned} \right\} i \in 1, N \\ & \text{variables} && \alpha = (\alpha_1, \dots, \alpha_N) \in \mathbb{R}^N, b \in \mathbb{R}, \xi = (\xi_1, \dots, \xi_N) \in \mathbb{R}^N \end{aligned} \quad (7)$$

as well as its Lagrangian dual problem:

$$\begin{aligned} & \text{maximize} && \theta(\omega, \beta, \gamma) = \inf_{\alpha \in \mathbb{R}^N, b \in \mathbb{R}, \xi \in \mathbb{R}^N} L(\alpha, b, \xi, \omega, \beta, \gamma) \\ & \text{subject to} && \omega_i \geq 0, \beta_i \geq 0, \gamma_i \geq 0, \quad i \in 1, N \\ & \text{variables} && \omega = (\omega_1, \dots, \omega_N) \in \mathbb{R}^N, \beta = (\beta_1, \dots, \beta_N) \in \mathbb{R}^N, \gamma = (\gamma_1, \dots, \gamma_N) \in \mathbb{R}^N \end{aligned} \quad (8)$$

1. Write down the Lagrangian function $L(\alpha, b, \xi, \omega, \beta, \gamma)$ in explicit form of $\alpha, b, \xi, \omega, \beta, \gamma$.

2. Show that the duality gap between (7) and (8) is zero.
3. Derive $\theta(\boldsymbol{\omega}, \beta, \gamma)$ in explicit form of dual variables $\boldsymbol{\omega}, \beta, \gamma$.
4. Show that the dual problem can be simplified as

$$\begin{aligned} & \text{maximize} && \sum_{i=1}^N \omega_i - \frac{1}{2} \sum_{i=1}^N \left(\sum_{j=1}^N \omega_j y_j y_i \mathbf{x}_j \cdot \mathbf{x}_i + \beta_i \right)^2 \\ & \text{subject to} && \sum_{i=1}^N \omega_i y_i = 0 \\ & \text{variables} && 0 \leq \omega_i \leq C_i, \quad i = 1, \dots, N \end{aligned} \quad (9)$$

5. Suppose $(\bar{\alpha}, \bar{b}, \bar{\xi})$ and $(\bar{\omega}, \bar{\beta}, \bar{\gamma})$ are the optimal solutions to problems (7) and (8) respectively. Denote $\bar{\mathbf{w}} = \sum_{j=1}^N \bar{\alpha}_j y_j \mathbf{x}_j$.

(a) Prove that

$$\bar{\alpha}_i = \max \left(\sum_{j=1}^N \bar{\omega}_j y_j y_i \mathbf{x}_j \cdot \mathbf{x}_i, 0 \right) \quad \forall i = 1, \dots, N \quad (10)$$

(b) Prove that

$$\bar{b} = \min_{b \in \mathbb{R}} \sum_{i=1}^N C_i \max(1 - y_i (\bar{\mathbf{w}} \cdot \mathbf{x}_i + b), 0), \quad (11)$$

(c) Prove that $\bar{\xi}_i = \max(1 - y_i (\bar{\mathbf{w}} \cdot \mathbf{x}_i + \bar{b}), 0)$ for all $i = 1, N$.

(d) Prove that

$$\left. \begin{aligned} \bar{\omega}_i &= C_i, && \text{if } y_i (\bar{\mathbf{w}} \cdot \mathbf{x}_i + \bar{b}) < 1 \\ \bar{\omega}_i &= 0, && \text{if } y_i (\bar{\mathbf{w}} \cdot \mathbf{x}_i + \bar{b}) > 1 \\ 0 \leq \bar{\omega}_i &\leq C_i, && \text{if } y_i (\bar{\mathbf{w}} \cdot \mathbf{x}_i + \bar{b}) = 1 \end{aligned} \right\} \quad \forall i = 1, \dots, N$$

Problem 5 (Spherical one class SVM)(2%)(Bonus)

Suppose we aim to fit a hypersphere which encompasses a majority of data points $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^M$ by considering the following optimization problem: (here $\boldsymbol{\mu}$ and each \mathbf{x}_i are considered as column vectors)

$$\begin{aligned} & \text{minimize} && R^2 + \frac{1}{\nu} \sum_{i=1}^N C_i \xi_i \\ & \text{subject to} && \left. \begin{aligned} \|\mathbf{x}_i - \boldsymbol{\mu}\|^2 &\leq R^2 + \xi_i \\ \xi_i &\geq 0 \end{aligned} \right\} \quad \forall i \in 1, N \\ & \text{variables} && R \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\xi} = (\xi_1, \dots, \xi_N) \in \mathbb{R}^N \end{aligned} \quad (12)$$

where $C_i > 0$ for each $i \in 1, N$, and $0 < \nu < \sum_{i=1}^N C_i$. Let $\rho = R^2$ and rewrite (12) in the form of primal problem:

$$\begin{aligned} & \text{minimize} && f(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = \rho + \frac{1}{\nu} \sum_{i=1}^N C_i \xi_i \\ & \text{subject to} && \left. \begin{aligned} g_{1,i}(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) &= \|\mathbf{x}_i - \boldsymbol{\mu}\|^2 - \rho - \xi_i \leq 0 \\ g_{2,i}(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) &= -\xi_i \leq 0 \end{aligned} \right\} \quad \forall i \in 1, N \\ & \text{variables} && \rho \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\xi} \in \mathbb{R}^N \end{aligned} \quad (13)$$

as well its Lagrangian dual problem:

$$\begin{aligned} & \text{maximize} && \theta(\alpha, \beta, \gamma) = \inf_{\rho \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\xi} \in \mathbb{R}^N} L(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}, \alpha, \beta, \gamma) \\ & \text{subject to} && \left. \begin{aligned} \alpha_i &\geq 0, \beta_i \geq 0 \quad \forall i \in 1, N \\ \gamma &\geq 0 \end{aligned} \right\} \\ & \text{variables} && \alpha = (\alpha_1, \dots, \alpha_N) \in \mathbb{R}^N, \beta = (\beta_1, \dots, \beta_N) \in \mathbb{R}^N, \gamma \in \mathbb{R} \end{aligned} \quad (14)$$

1. Write down the Lagrangian function $L(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}, \alpha, \beta, \gamma)$ in explicit form of $\rho, \boldsymbol{\mu}, \boldsymbol{\xi}, \alpha, \beta, \gamma$.
2. Show that the duality gap between (13) and (14) is zero.

3. Derive $\theta(\alpha, \beta, \gamma)$ in explicit form of dual variables α, β, γ .
4. Show that the dual problem can be simplified as

$$\begin{aligned}
& \text{maximize} && \|\alpha\|_1 \left(\sum_{i=1}^N \hat{\alpha}_i \|\mathbf{x}_i\|^2 - \sum_{1 \leq i, j \leq N} \hat{\alpha}_i \hat{\alpha}_j \mathbf{x}_i^T \mathbf{x}_j \right) \\
& \text{subject to} && \sum_{i=1}^N \alpha_i \leq 1 \\
& \text{variables} && 0 \leq \alpha_i \leq \frac{C_i}{\nu}, i \in 1, N
\end{aligned} \tag{15}$$

where $\|\alpha\|_1 = \sum_{i=1}^N \alpha_i$ and $\alpha_i = \|\alpha\|_1 \hat{\alpha}_i$.

5. Suppose $(\bar{\rho}, \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\xi}})$ and $(\bar{\alpha}, \bar{\beta}, \bar{\gamma})$ are optimal solutions to problems (13) and (14), respectively.

- (a) Show that $\|\bar{\alpha}\|_1 \bar{\boldsymbol{\mu}} = \sum_{i=1}^N \bar{\alpha}_i \mathbf{x}_i$.
- (b) Show that

$$\bar{\rho} \in_{\rho \geq 0} \left(\rho + \frac{1}{\nu} \sum_{i=1}^N C_i \max(\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 - \rho, 0) \right).$$

- (c) Show that

$$\min \left\{ \rho \geq 0 : \sum_{i: \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 > \rho} C_i \leq \nu \right\} \leq \bar{\rho} \leq \min \left\{ \rho \geq 0 : \sum_{i: \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 > \rho} C_i < \nu \right\}. \tag{16}$$

- (d) Prove that $\bar{\xi}_i = \max(\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 - \bar{\rho}, 0)$ for each $i \in 1, N$.
- (e) Prove that

$$\begin{cases} \bar{\alpha}_i = C_i/\nu & , \text{ if } \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 > \bar{\rho} \\ \bar{\alpha}_i = 0 & , \text{ if } \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 < \bar{\rho} \\ 0 \leq \bar{\alpha}_i \leq C_i/\nu & , \text{ if } \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 = \bar{\rho} \end{cases}.$$

6. Suppose $C_i = 1/n$ for each $i \in 1, n$. What is the physical meaning of ν ?

Version Description

1. First Edition: Finish Problem 1 to 5