HW4 Handwritten Assignment

Lecturer: Pei-Yuan Wu
TAs: Po-Yang Hsieh, Le-Rong Hsu

November 2024

Problem 1 (True or false)(0.6%)

You don't need to give any explanation for this problem. Just determine the correctness of the following statement and answer "T" for true and "F" for false. The notation here please refer to the course video.

1. (0.2 %) The strong duality holds only when the primal problem is convex and satisfies the Slater's conditions.
2. (0.2%) The complementary slackness condition, $u_i g_i(x) = 0, \forall i = 1, \dots m$, in a minimization problem, implies that "whenever $g(\bar{x}) = 0$, then $u_i > 0$".
3. (0.2%) The dual function $\theta(\mathbf{u}, \mathbf{v})$ gives a lower bound of the optimal value of the primal problem (as a convex minimization problem in standard form) when $\theta(\mathbf{u}, \mathbf{v}) > -\infty$ and $\mathbf{u} \geq 0$.

Problem 2 (SVM with Gaussian kernel)(0.9%)

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 $\{(x_1, y_1), \cdots, (x_N, y_N)\}$ 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 $i = 1, \cdots, m$ and $b = 0$. Find values for the Gaussian kernel width τ such that x_i is correctly classified, for all $i = 1, \cdots, N$, e.g., $f(x_i)y_i > 0$ for all $i = 1, \cdots, N$.

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)(1.5%)

Suppose we are given a training set $\{(x_1, y_1), \cdots, (x_m, y_m)\}$, 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

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

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

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

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

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

(a) Write down the Lagrangian for the optimization problem above. Consider the sets of Lagrange multiplier α_i, α_i^*, β_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 $\alpha = (\alpha_1, \cdots, \alpha_m)$, $\alpha^* = (\alpha_1^*, \cdots, \alpha_m^*)$, and $\beta = (\beta_1, \cdots, \beta_m)$.

(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

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

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

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

(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 produce.

Problem 4 (Hinge loss with L^1 regularization)(1.5%)

Given data points $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^m$ as well as their labels $y_1, \ldots, y_N \in \{\pm 1\}$ and penalty coefficients $C_1, \ldots, C_N > 0$, where each $\mathbf{x}_i = [x_{i,1}, \ldots, x_{i,m}]^T$ is a column vector, consider the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \|\mathbf{w}\|_1 + \sum_{i=1}^N C_i \xi_i \\ \text{subject to} & \left.\begin{array}{l} y_i(\mathbf{w}^T \mathbf{x}_i + b) \ge 1 - \xi_i \\ \xi_i \ge 0 \end{array}\right\} \; i = 1, \ldots, N \\ \text{variables} & \mathbf{w} \in \mathbb{R}^m, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N \end{array} \tag{7}$$

Note that in this formulation, we replace the L^2-regularization term $\frac{1}{2}\|\mathbf{w}\|^2$ by the L^1-regularization term $\|\mathbf{w}\|_1 = \sum_{j=1}^m |w_j|$.

(a) (0.5%) Show that $(\bar{\mathbf{w}}, \bar{b}, \bar{\boldsymbol{\xi}})$ is an optimal solution of (7) if and only if $\bar{\mathbf{w}} = \bar{\mathbf{u}} - \bar{\mathbf{v}}$, where $(\bar{\mathbf{u}}, \bar{\mathbf{v}}, \bar{b}, \bar{\boldsymbol{\xi}})$ is an optimal solution of the following problem:

$$\begin{array}{ll} \text{minimize} & f(\mathbf{u}, \mathbf{v}, b, \boldsymbol{\xi}) = \sum_{j=1}^m (u_j + v_j) + \sum_{i=1}^N C_i \xi_i \\ \text{subject to} & y_i((\mathbf{u} - \mathbf{v})^T \mathbf{x}_i + b) \ge 1 - \xi_i, \quad i = 1, \ldots, N \\ & \xi_i \ge 0, \quad i = 1, \ldots, N \\ & u_j \ge 0, \quad v_j \ge 0, \quad j = 1, \ldots, m \\ \text{variables} & \mathbf{u} \in \mathbb{R}^m, \mathbf{v} \in \mathbb{R}^m, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N \end{array} \tag{8}$$

Following (a), we can now rewrite (8) as the following primal problem:

$$\begin{array}{ll} \text{minimize} & f(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi}) = \sum_{j=1}^{m}(u_j+v_j) + \sum_{i=1}^{N} C_i\xi_i \\ \text{subject to} & g_i^1(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi}) = 1-\xi_i - y_i((\mathbf{u}-\mathbf{v})^T\mathbf{x}_i + b) \le 0, \quad i=1,\ldots,N \\ & g_i^2(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi}) = -\xi_i \le 0, \quad i=1,\ldots,N \\ & g_i^3(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi}) = -u_j \le 0, \quad j=1,\ldots,m \\ & g_i^4(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi}) = -v_j \le 0, \quad j=1,\ldots,m \\ \text{variables} & \mathbf{u}\in\mathbb{R}^m, \mathbf{v}\in\mathbb{R}^m, b\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^N \end{array} \tag{9}$$

as well as its Lagrangian dual problem:

$$\begin{array}{ll} \text{maximize} & \theta(\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu}) = \inf\left\{L(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi},\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu}) : \mathbf{u}\in\mathbb{R}^m, \mathbf{v}\in\mathbb{R}^m, b\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^N\right\} \\ \text{subject to} & \alpha_i \ge 0, \beta_i \ge 0 \quad , i=1,\ldots,N \\ & \mu_j \ge 0, \nu_j \ge 0 \quad , j=1,\ldots,m \\ \text{variables} & \alpha\in\mathbb{R}^N, \beta\in\mathbb{R}^N, \boldsymbol{\mu}\in\mathbb{R}^m, \boldsymbol{\nu}\in\mathbb{R}^m \end{array} \tag{10}$$

where L denotes the Lagranian function.

(b) (0.2%) Associate dual variables $\alpha_i, \beta_i, \mu_j, \nu_j$ to constraints $g_i^1, g_i^2, g_j^3, g_j^4$, respectively. Show that L can be written in the following explicit form:

$$\begin{aligned} &L(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi},\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu}) \\ &= \mathbf{1}^T(\mathbf{u}+\mathbf{v}) + \sum_{i=1}^{N} C_i\xi_i + \sum_{i=1}^{N}\alpha_i\left(1-\xi_i - y_i((\mathbf{u}-\mathbf{v})^T\mathbf{x}_i+b)\right) + \sum_{i=1}^{N}\beta_i(-\xi_i) - \boldsymbol{\mu}^T\mathbf{u} - \boldsymbol{\nu}^T\mathbf{v}, \end{aligned} \tag{11}$$

where $\mathbf{1}$ denotes the all-one vector.

(c) (0.1%) Show that (9) satisfies Slater's condition.

(d) (0.2%) Show that:

(i) (0.1%) $\theta(\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu}) = -\infty$ unless the following conditions hold:

$$\sum_{i=1}^{N}\alpha_i y_i = 0, \quad \boldsymbol{\mu} = \mathbf{1} - \sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i, \quad \boldsymbol{\nu} = \mathbf{1} + \sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i, \tag{12}$$

and

$$\alpha_i + \beta_i = C_i, \quad \forall i = 1,\ldots,N, \tag{13}$$

at which case $\theta(\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu}) = \sum_{i=1}^{N}\alpha_i$.

(ii) (0.1%) The stationary condition holds if and only if (12) and (13) are satisfied.

(e) (0.2%) Show that the dual problem (10) can be simplified as:

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{N}\alpha_i \\ \text{subject to} & \sum_{i=1}^{N}\alpha_i y_i = 0 \\ & -\mathbf{1} \le \sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i \le \mathbf{1} \\ \text{variables} & 0 \le \alpha_i \le C_i, \quad i=1,\ldots,N \end{array} \tag{14}$$

(f) (0.2%) Write down the KKT conditions for the primal and dual problems (9)(10).

(g) (0.1%) Is it true that $\bar{\mathbf{w}}$ must be a linear combination of $\mathbf{x}_1,\ldots,\mathbf{x}_N$? Justify your answer.

Problem 5 (Spherical one class SVM)(1.5%)

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

$$\begin{array}{ll} \text{minimize} & R^2 + \frac{1}{\nu}\sum_{i=1}^{N} C_i\xi_i \\ \text{subject to} & \left.\begin{array}{l} \|\mathbf{x}_i - \boldsymbol{\mu}\|^2 \le R^2 + \xi_i \\ \xi_i \ge 0 \end{array}\right\} \ \forall i \in [1, N] \\ & R \ge 0 \\ \text{variables} & R \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\xi} = (\xi_1, ..., \xi_N) \in \mathbb{R}^N \end{array} \tag{15}$$

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 (15) in the form of primal problem:

$$\begin{array}{ll} \text{minimize} & f(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = \rho + \frac{1}{\nu}\sum_{i=1}^{N} C_i\xi_i \\ \text{subject to} & \left.\begin{array}{l} g_{1,i}(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = \|\mathbf{x}_i - \boldsymbol{\mu}\|^2 - \rho - \xi_i \le 0 \\ g_{2,i}(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = -\xi_i \le 0 \end{array}\right\} \ \forall i \in [1, N] \\ & g_3(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = -\rho \le 0 \\ \text{variables} & \rho \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\xi} \in \mathbb{R}^N \end{array} \tag{16}$$

as well its Lagrangian dual problem:

$$\begin{array}{ll} \text{maximize} & \theta(\alpha, \beta, \gamma) = \inf_{\rho \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\xi} \in \mathbb{R}^N} L(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}, \alpha, \beta, \gamma) \\ \text{subject to} & \alpha_i \ge 0, \beta_i \ge 0 \ \forall i \in [1, N] \\ & \gamma \ge 0 \\ \text{variables} & \alpha = (\alpha_1, ..., \alpha_N) \in \mathbb{R}^N, \beta = (\beta_1, ..., \beta_N) \in \mathbb{R}^N, \gamma \in \mathbb{R} \end{array} \tag{17}$$

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 (16) and (17) 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

$$\begin{array}{ll} \text{maximize} & \|\alpha\|_1 \left(\sum_{i=1}^{N} \hat{\alpha}_i \|\mathbf{x}_i\|^2 - \sum_{1 \le i,j \le N} \hat{\alpha}_i \hat{\alpha}_j \mathbf{x}_i^T \mathbf{x}_j \right) \\ \text{subject to} & \sum_{i=1}^{N} \alpha_i \le 1 \\ \text{variables} & 0 \le \alpha_i \le \frac{C_i}{\nu}, i \in [1, N] \end{array} \tag{18}$$

   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 (16) and (17), respectively.
   (a) Show that $\|\bar{\alpha}\|_1 \bar{\boldsymbol{\mu}} = \sum_{i=1}^{N} \bar{\alpha}_i \mathbf{x}_i$.
   (b) Show that
   $$\bar{\rho} \in \arg min_{\rho \ge 0} \left(\rho + \frac{1}{\nu} \sum_{i=1}^{N} C_i \max(\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 - \rho, 0) \right).$$
   (c) Show that
   $$\min \left\{ \rho \ge 0 : \sum_{i:\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 > \rho} C_i \le \nu \right\} \le \bar{\rho} \le \min \left\{ \rho \ge 0 : \sum_{i:\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 > \rho} C_i < \nu \right\}. \tag{19}$$
   (d) Prove that $\bar{\xi}_i = \max\left(\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 - \bar{\rho}, 0\right)$ for each $i \in [1, N]$.
   (e) Prove that
   $$\begin{cases} \bar{\alpha}_i = C_i/\nu & , \text{if } \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 > \bar{\rho} \\ \bar{\alpha}_i = 0 & , \text{if } \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 < \bar{\rho} \\ 0 \le \bar{\alpha}_i \le C_i/\nu & , \text{if } \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 = \bar{\rho} \end{cases}.$$
6. Suppose $C_i = 1/n$ for each $i \in [1, n]$. What is the physical meaning of ν?