Introduction to Machine Learning
Fall 2021
University of Science and Technology of China

Lecturer: Jie Wang — Homework 6
Posted: Dec. 13, 2021 — Due: Dec. 27, 2021

Notice, to get the full credits, please show your solutions step by step.

Exercise 1: SVM for Linearly Separable Cases

Given the training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Let

$$\mathcal{D}^+ = \{(\mathbf{x}_i, y_i) \in \mathcal{D} : y_i = 1\}, \quad \mathcal{D}^- = \{(\mathbf{x}_i, y_i) \in \mathcal{D} : y_i = -1\}.$$

Assume that $\mathcal{D}^+$ and $\mathcal{D}^-$ are nonempty and the training set $\mathcal{D}$ is linearly separable. We have shown in Lecture 13 that SVM can be written as

$$\begin{aligned} \min_{\mathbf{w},b} \ & \frac{1}{2}\|\mathbf{w}\|^2, \\ \text{s.t.} \ & \min_i y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1. \end{aligned} \tag{1}$$

Moreover, we further transform Problem (1) to

$$\begin{aligned} \min_{\mathbf{w},b} \ & \frac{1}{2}\|\mathbf{w}\|^2, \\ \text{s.t.} \ & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1, i = 1, \ldots, n. \end{aligned} \tag{2}$$

We denote the feasible set of Problem (2) by

$$\mathcal{F} = \{(\mathbf{w}, b) : y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1, i = 1, \ldots, n\}.$$

1. The Euclidean distance between a linear classifier $f(\mathbf{x}; \mathbf{w}, b) = \langle \mathbf{w}, \mathbf{x} \rangle + b$ and a point $\mathbf{z}$ is

   $$d(\mathbf{z}, f) = \min_{\mathbf{x}} \{\|\mathbf{z} - \mathbf{x}\| : f(\mathbf{x}; \mathbf{w}, b) = 0\}.$$

   Please find the closed form of $d(\mathbf{z}, f)$.

2. Show that $\mathcal{F}$ is nonempty.

3. Show that Problem (2) admits an optimal solution.

4. Let $(\mathbf{w}^*, b^*)$ be the optimal solution to Problem (2). Show that $\mathbf{w}^* \neq 0$.

5. Show that Problems (1) and (2) are equivalent, that is, they share the same set of optimal solutions.

6. Let $(\mathbf{w}^*, b^*)$ be the optimal solution to Problem (2). Show that there exist at least one positive sample and one negative sample, respectively, such that the corresponding equality holds. In other words, there exist $i, j \in \{1, 2, \ldots, n\}$ such that

$$\begin{aligned} 1 = y_i =& \langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*, \\ -1 = y_j =& \langle \mathbf{w}^*, \mathbf{x}_j \rangle + b^*. \end{aligned}$$

7. Show that the optimal solution to Problem (2) is unique.

8. Can we remove the inequalities that hold strictly at the optimum to Problem (2) without affecting the solution? Please justify your claim rigorously.

9. Find the dual problem of (2) and the corresponding optimal conditions.

Solution: ■

Exercise 2: Exercises of Dual Problems

1. Consider the optimization problem

$$\begin{aligned}\min_x\ & x^2+1\\ \text{s.t.}\ & (x-2)(x-4)\le 0,\\ & x\in\mathbb{R}.\end{aligned}$$

   (a) Give the feasible set, the optimal value, and the optimal solution.

   (b) Plot the objective x^2+1 versus x. On the same plot, show the feasible set, optimal point and value, and plot the Lagrangian $L(x,\lambda)$ versus x for a few positive values of λ. Verify the lower bound property, i.e., $p^*\ge \inf_x L(x,\lambda)$ for $\lambda\ge 0$, where p^* is the optimal value. Derive and sketch the Lagrange dual function.

   (c) State the dual problem, and verify that it is a concave maximization problem. Find the dual optimal value and dual optimal solution λ^*. Does strong duality hold?

   (d) (**Sensitivity analysis**) Let $p^*(u)$ denote the optimal value of the problem

$$\begin{aligned}\min_x\ & x^2+1\\ \text{s.t.}\ & (x-2)(x-4)\le u,\\ & x\in\mathbb{R},\end{aligned}$$

   as a function of the parameter u. Please plot $p^*(u)$ and verify that $\mathrm{d}p^*(0)/\mathrm{d}u = -\lambda^*$.

2. (**Optional**) Please find the dual of the following problems respectively.

   (a) The entropy maximization problem

$$\begin{aligned}\min_{\mathbf{p}}\ & \sum_{i=1}^n p_i\log p_i\\ \text{s.t.}\ & \mathbf{A}\mathbf{p}\preceq\mathbf{b},\\ & \mathbf{1}^\top\mathbf{p}=1,\\ & \mathbf{p}\in\mathbb{R}^n_+,\end{aligned}$$

   where $\mathbf{A}\in\mathbb{R}^{m\times n}$ and $\mathbf{b}\in\mathbb{R}^m$.

   (b) The projection problem

$$\begin{aligned}\min_{\mathbf{x}}\ & \|\mathbf{z}-\mathbf{x}\|^2\\ \text{s.t.}\ & \mathbf{A}\mathbf{x}=\mathbf{0},\\ & \mathbf{x}\in\mathbb{R}^n,\end{aligned}$$

   where $\mathbf{A}\in\mathbb{R}^{m\times n}$ and $\mathbf{z}\in\mathbb{R}^n$.

(c) The problem of minimizing a nonconvex quadratic function over the unit ball

$$\begin{aligned}\min_{\mathbf{x}} \ & \mathbf{x}^\top \mathbf{A}\mathbf{x} + 2\mathbf{b}^\top \mathbf{x} \\ \text{s.t.} \ & \mathbf{x}^\top \mathbf{x} \le 1, \\ & \mathbf{x} \in \mathbb{R}^n,\end{aligned}$$

where $\mathbf{A} \in \mathbf{S}^n, \mathbf{A} \not\succeq 0$ and $\mathbf{b} \in \mathbb{R}^n$

3. Please use the duality to show that in three-dimensional space, the (minimum) distance from the origin to a line is equal to the maximum over all (minimum) distances of the origin from planes that contain the line.

4. (**Optional**) A Boolen linear program is an optimization problem of the form

$$\begin{aligned}\min_{\mathbf{x}\in\mathbb{R}^n} \ & \mathbf{c}^\top \mathbf{x} \\ \text{s.t.} \ & \mathbf{A}\mathbf{x} \preceq \mathbf{b} \\ & x_i \in \{0, 1\}, \ i = 1, \cdots, n,\end{aligned}$$

and is, in general, difficult to solve. We can study the LP relaxation of this problem,

$$\begin{aligned}\min_{\mathbf{x}\in\mathbb{R}^n} \ & \mathbf{c}^\top \mathbf{x} \\ \text{s.t.} \ & \mathbf{A}\mathbf{x} \preceq \mathbf{b} \\ & 0 \le x_i \le 1, \ i = 1, \cdots, n,\end{aligned} \tag{3}$$

which is far easier to solve, and gives a lower bound on the optimal value of the Boolean LP.

(a) The Boolen LP can also be reformulated as the problem

$$\begin{aligned}\min_{\mathbf{x}\in\mathbb{R}^n} \ & \mathbf{c}^\top \mathbf{x} \\ \text{s.t.} \ & \mathbf{A}\mathbf{x} \preceq \mathbf{b} \\ & x_i(1 - x_i) = 0, \ i = 1, \cdots, n,\end{aligned}$$

which has quadratic equality constraints. Find the Lagrange dual of this problem. The optimal value of the dual problem (which is convex) gives a lower bound on the optimal value of the Boolean LP. This method of finding a lower bound on the optimal value is called Lagrangian relaxation.

(b) Show that the lower bound obtained via Lagrangian relaxation, and via the LP relaxation (3), are the same.(**Hint:** Derive the dual of the LP relaxation (3).)

5. We can transform an unconstrained optimization problem to a constrained optimization problem via some intermediate variables.

(a) (**Optional**) Consider the unconstrained geometric program problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \log\left(\sum_{i=1}^{m} \exp\left(\mathbf{a}_i^\top \mathbf{x} + b_i\right)\right),$$

where the $\mathbf{a}_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$ for $i = 1, 2, \cdots, m$. It is equivalent to the problem

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n,\mathbf{y}\in\mathbb{R}^m} \ & \sum_{i=1}^{m} y_i \\ \text{s.t. } & \mathbf{a}_i^\top \mathbf{x} = \log y_i - b_i,\ i = 1, \cdots, m. \end{aligned} \tag{4}$$

Please find the dual problem of (4) and show that the dual problem is an entropy maximization problem.

(b) (**Optional**) Consider the convex piecewise-linear minimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \max_{1\le i\le m} \left(\mathbf{a}_i^\top \mathbf{x} + b_i\right),$$

where $\mathbf{a}_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$. Please find its dual problem by letting $y_i = \mathbf{a}_i^\top \mathbf{x} + b_i$.

(c) Recall that Lasso takes the form of

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1,$$

where $\mathbf{y} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n\times d}$ and $\lambda > 0$. Please find its dual problem by letting $\mathbf{z} = \mathbf{y} - \mathbf{X}\mathbf{w}$.

Solution: ■

Exercise 3: Discussions on Geometric Multiplier and Duality Gap

Consider the primal problem

$$\begin{aligned} \min_{\mathbf{x}} \ & f(\mathbf{x}) \\ \text{s.t. } & g_i(\mathbf{x}) \leq 0, \ i = 1, \cdots, m, \\ & h_i(\mathbf{x}) = 0, i = 1, \cdots, p, \\ & \mathbf{x} \in X. \end{aligned} \tag{5}$$

Let

$$S = \{(\mathbf{g}(\mathbf{x}), \mathbf{h}(\mathbf{x}), f(\mathbf{x})) : \mathbf{x} \in X\} \subset \mathbb{R}^{m+p+1}. \tag{6}$$

Are the following claims on the geometric multiplier and the duality gap for the primal problem correct? Justify the claims rigorously if they are correct. Otherwise please give a counterexample for each.

1. The geometric multiplier for the primal problem (5) always exists.
2. If the geometric multiplier exists, then it is unique.
3. If the geometric multiplier exists, then the duality gap is zero.
4. If the duality gap is zero, there exists at least one geometric multiplier.
5. Let (λ^*, μ^*) be a geometric multiplier. Then, the problem $\mathbf{argmin}_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \mu^*)$ always admits at least one solution, where $L(\mathbf{x}, \lambda, \mu)$ is the Lagrangian for (5).
6. If (λ^*, μ^*) is a geometric multiplier and the problem $\mathbf{argmin}_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \mu^*)$ admits a solution $\mathbf{x}^*$, then $\mathbf{x}^*$ is feasible.
7. Let (λ^*, μ^*) be a geometric multiplier. Then, $\mathbf{x}^*$ is a global minimum of the primal problem if and only if $\mathbf{x}^*$ is feasible and $\mathbf{x}^* \in \mathbf{argmin}_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \mu^*)$.
8. (λ^*, μ^*) is a geometric multiplier if and only if $\lambda^* \geq 0$ and among all hyperplanes with normal $(\lambda^*, \mu^*, 1)$ that contain the set S in their positive halfspace, the highest attained level of interception of the vertical axis is f^*, where

$$f^* = \inf\{f(\mathbf{x}) : \mathbf{g}(\mathbf{x}) \leq 0, \mathbf{h}(\mathbf{x}) = 0, \mathbf{x} \in X)\}.$$

Exercise 4: Inequality Form Linear Programming and its Dual (Optional)

Let $\mathbf{c}, \mathbf{x} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$, and $A \in \mathbb{R}^{m \times n}$. Consider the inequality form LP

$$\begin{aligned} \min_{\mathbf{x}} \; & \mathbf{c}^\top \mathbf{x} \\ \text{s.t. } & A\mathbf{x} \preceq \mathbf{b}. \end{aligned} \tag{7}$$

1. Find the dual problem of Problem (7).

2. (**A Geometric Interpretation of the Dual LP**) Let $\mathbf{w} \in \mathbb{R}^m_+$. If $\mathbf{x}$ is feasible for the LP, i.e., satisfies $A\mathbf{x} \preceq \mathbf{b}$, then it also satisfies the inequality

$$\mathbf{w}^\top A\mathbf{x} \leq \mathbf{w}^\top \mathbf{b}.$$

   Geometrically, for any $\mathbf{w} \succeq \mathbf{0}$, the halfspace $H_{\mathbf{w}} = \{\mathbf{x} : \mathbf{w}^\top A\mathbf{x} \leq \mathbf{w}^\top \mathbf{b}\}$ contains the feasible set for the LP. Therefore, if we minimize the objective $\mathbf{c}^\top \mathbf{x}$ over the halfspace $H_{\mathbf{w}}$, we get a lower bound on f^*.

   (a) Derive an expression for the minimum value of $\mathbf{c}^\top \mathbf{x}$ over the halfspace $H_{\mathbf{w}}$ (which will depend on the choice of $\mathbf{w}$).

   (b) Formulate the problem of finding the best such bound, by maximizing the lower bound over $\mathbf{w} \succeq \mathbf{0}$.

   (c) Relate the results of (a) and (b) to the Lagrange dual of the LP.

3. (**Strong Duality in LP**) We want to show that the **strong duality** holds for Problem (7) and its dual provided **at least one of the two problems above is feasible**. In other words, the only possible exception to the strong duality occurs when $f^* = \infty$ and $q^* = -\infty$, where f^* and q^* denotes the primal and the dual optimal values, respectively. Let b_i be the ith component of $\mathbf{b}$ and $\mathbf{a}_i^\top$ be the ith row of A.

   (a) Suppose that f^* is finite and $\mathbf{x}^*$ is an optimal solution. Let $I \subset \{1, 2, \ldots, m\}$ be the set of active constraints at $\mathbf{x}^*$:

$$\begin{aligned} &\mathbf{a}_i^\top \mathbf{x}^* = b_i, \; i \in I, \\ &\mathbf{a}_i^\top \mathbf{x}^* < b_i, \; i \notin I. \end{aligned}$$

   Show that there exists a point $\mathbf{z} \in \mathbb{R}^m$ such that

$$\begin{aligned} &z_i \geq 0, \; i \in I, \\ &z_i = 0, \; i \notin I, \\ &\sum_{i \in I} z_i \mathbf{a}_i + \mathbf{c} = 0, \end{aligned}$$

   where z_i is the i^{th} component of $\mathbf{z}$. Show that $\mathbf{z}$ is a dual optimal solution with the objective value $\mathbf{c}^\top \mathbf{x}^*$.

   (b) Suppose that $f^* = \infty$ and the dual problem is feasible. Show that $q^* = \infty$.

(c) Consider the example

$$\begin{aligned} &\min_{x \in \mathbb{R}} \; x \\ &\text{s.t.} \; \begin{pmatrix} 0 \\ 1 \end{pmatrix} x \preceq \begin{pmatrix} -1 \\ 1 \end{pmatrix}. \end{aligned}$$

Formulate the dual problem, and solve the primal and the dual problems. Show that $f^* = \infty$ and $q^* = -\infty$.

Solution: ■

Exercise 5: Piecewise-linear Minimization (Optional)

We consider the convex piecewise-linear minimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \max_{1\le i\le m} (\mathbf{a}_i^\top \mathbf{x} + b_i), \tag{8}$$

where $\mathbf{a}_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$ for $i = 1, \ldots, n$.

1. Derive a dual problem, based on the Lagrange dual of the equivalent problem

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}\in\mathbb{R}^n} \ & \max_{1\le i\le m} y_i \\ \text{s.t. } & \mathbf{a}_i^\top \mathbf{x} + b_i = y_i,\ i = 1, \ldots, m. \end{aligned}$$

2. Formulate the piecewise-linear minimization problem (8) as an LP, and form the dual of the LP. Relate the LP dual to the dual obtained in part 1.

3. Suppose that we approximate the objective function in (8) by the smooth function

$$f_0(\mathbf{x}) = \log\left(\sum_{i=1}^m \exp(\mathbf{a}_i^\top \mathbf{x} + b_i)\right),$$

   and solve the unconstrained geometric program

$$\min_{\mathbf{x}\in\mathbb{R}^n} \ f_0(\mathbf{x}). \tag{9}$$

   (a) Derive a dual of this problem by letting $y_i = \exp(\mathbf{a}_i^\top \mathbf{x} + b_i)$, $i = 1, \ldots, m$.

   (b) Let p^*_{pwl} and p^*_{gp} be the optimal values of (8) and (9), respectively. Show that

$$0 \le p^*_{\text{gp}} - p^*_{\text{pwl}} \le \log m.$$

   (c) Derive similar bounds for the difference between p^*_{pwl} and the optimal value of

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{\gamma} \log\left(\sum_{i=1}^m \exp\gamma(\mathbf{a}_i^\top \mathbf{x} + b_i)\right),$$

   where $\gamma > 0$ is a parameter. What happens as we increase γ?

Exercise 6: Duality Gap of the Knapsack Problem (Optional)

Given objects $i = 1, 2, \ldots, n$ with positive weights w_i and values v_i, we want to assemble a subset of the objects so that the sum of the weights of the subset does not exceed a given $A > 0$, and the sum of the values of the subset is maximized.

1. Show that this problem can be written as

$$\begin{aligned} \max_{x_1,\ldots,x_n} \ & \sum_{i=1}^n v_i x_i \\ \text{s.t.} \ & \sum_{i=1}^n w_i x_i \le A, \\ & x_i \in \{0,1\},\ i = 1, 2, \ldots, n. \end{aligned} \tag{10}$$

2. Find the dual problem of (10) where a Lagrange multiplier is assigned to the constraint $\sum_{i=1}^n w_i x_i \le A$.

3. Suppose that the objects are sorted in the order $\frac{v_1}{w_1} \le \frac{v_2}{w_2} \le \cdots \le \frac{v_n}{w_n}$. Please solve the dual problem above. (Hint: the objective function $q(\lambda)$ is piecewise linear.)

4. Consider a slack version of the primal problem:

$$\begin{aligned} \max_{x_1,\ldots,x_n} \ & \sum_{i=1}^n v_i x_i \\ \text{s.t.} \ & \sum_{i=1}^n w_i x_i \le A, \\ & x_i \in [0,1],\ i = 1, 2, \ldots, n. \end{aligned} \tag{11}$$

   Let f_R^* be the optimal value of problem (11) and and q_R^* be the optimal value of the dual problem, respectively. Show that

$$f_R^* = q_R^*.$$

5. Let f^* be the optimal value of problem (10) and and q^* be the optimal value of the dual problem, respectively. Show that

$$0 \le q^* - f^* \le \max_{i=1,\ldots,n} v_i.$$

   (Hint: find the relationship between (f^*, q^*) and (f_R^*, q_R^*).)

6. Consider the problem where A is multiplied by a positive integer k and each object is replaced by k replicas of itself, while the object weights and values stay the same. Let $f^*(k)$ and $q^*(k)$ be the corresponding optimal primal and dual values. Show that

$$\frac{q^*(k) - f^*(k)}{f^*(k)} \leq \frac{1}{k} \frac{\max_{i=1,\ldots,n} v_i}{f^*}.$$

Thus, the relative value of the duality gap tends to 0 as $k \to \infty$.

Solution: ■

Exercise 7: The Dual Problem of SVM

Suppose that the training set is $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Let

$$\mathcal{D}^+ = \{(\mathbf{x}_i, y_i) \in \mathcal{D} : y_i = 1\}, \quad \mathcal{D}^- = \{(\mathbf{x}_i, y_i) \in \mathcal{D} : y_i = -1\}.$$

Assume that $\mathcal{D}^+$ and $\mathcal{D}^-$ are nonempty. The soft margin SVM takes the form of

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \ & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^n \xi_i, \\ \text{s.t. } & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i, i = 1, \ldots, n, \\ & \xi_i \geq 0, \ i = 1, \ldots, n, \end{aligned} \tag{12}$$

The corresponding dual problem is

$$\begin{aligned} \min_{\alpha} \ & \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \sum_{i=1}^n \alpha_i \\ \text{s.t. } & \sum_{i=1}^n \alpha_i y_i = 0, \\ & \alpha_i \in [0, C], i = 1, \ldots, n. \end{aligned} \tag{13}$$

1. Show that the problems (12) and (13) always admit optimal solutions.
2. Let $(\mathbf{w}^*, b^*)$ be the solution to (12) and α^* be the corresponding solution to (13).
   (a) When does α_i^* equals to C, $i = 1, \ldots, n$? Please give an example and find the corresponding solutions.
   (b) When dose $\mathbf{w}^*$ equal to 0? Please give an example and find the corresponding solutions.

   Notice that, you need to find all the primal and dual optimal solutions if they are not unique.

Solution: ■

Exercise 8: An Example of the Soft Margin SVM

Recall that the soft margin SVM takes the form of

$$\begin{aligned}
&\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i, \\
&\text{s.t. } y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i, i = 1, \ldots, n, \\
&\quad\quad \xi_i \geq 0,\, i = 1, \ldots, n,
\end{aligned} \tag{14}$$

where $C > 0$.

1. The function of the slack variables used in the optimization problem for soft margin hyperplanes takes the form $\sum_{i=1}^{n}\xi_i$. We could also use $\sum_{i=1}^{n}\xi_i^p$, where $p > 1$. The soft margin SVM becomes

$$\begin{aligned}
&\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i^p, \\
&\text{s.t. } y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i, i = 1, \ldots, n, \\
&\quad\quad \xi_i \geq 0,\, i = 1, \ldots, n.
\end{aligned} \tag{15}$$

   Please find the dual problem of (15) and the corresponding optimal conditions.

As shown in Figure 1, the training set is $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{11}$, where $\mathbf{x}_i \in \mathbb{R}^2$ and $y_i \in \{+1, -1\}$. Suppose that we use the soft margin SVM to classify the data points and get the optimal parameters $\mathbf{w}^*$, b^*, and $\boldsymbol{\xi}^*$ by solving the problem (15).

2. Please write down the equations of the separating hyperplane (H_0) and the marginal hyperplanes (H_1 and H_2) in terms of $\mathbf{w}^*$ and b^*.

3. Please find the support vectors and the non-support vectors.

4. Please find the values (or ranges) of the optimal slack variables ξ_i^* for $i = 1, 2, \ldots, 11$. (*Hint: The possible answers are* $\xi_i^* = 0$, $0 < \xi_i^* < 1$, $\xi_i^* = 1$, *and* $\xi_i^* > 1$). How do the slack variables change when the parameter C increases and decreases?

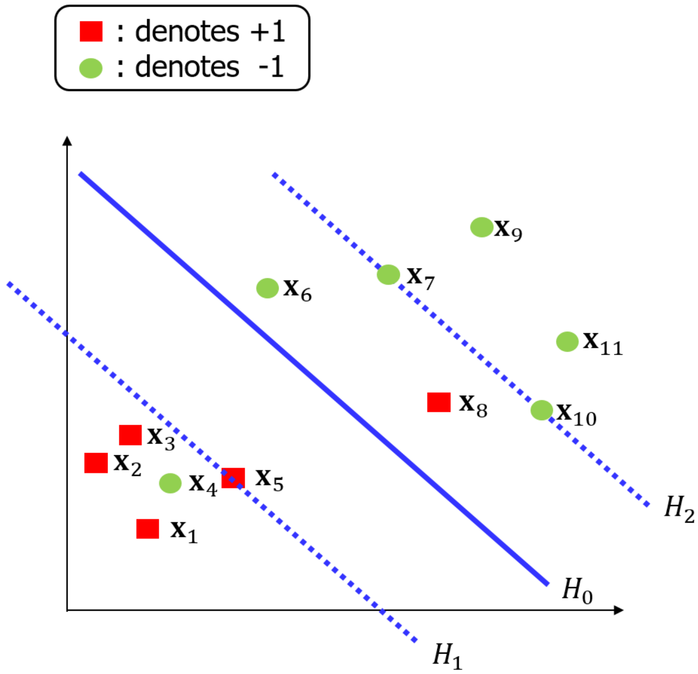

Figure 1: Classifying the data points using the soft margin SVM. H_0 is the separating hyperplane. H_1 and H_2 are the marginal hyperplanes.

Solution: ■

Exercise 9: Neural Networks

1. The softmax function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ is defined by:

$$f_i(\mathbf{x}) = \frac{\exp(x_i)}{\sum_{k=1}^{n} \exp(x_k)}, i = 1, \ldots, n,$$

   where x_i is the i^{th} component of $\mathbf{x} \in \mathbb{R}^n$. The function $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_n(\mathbf{x}))^\top$ converts each input $\mathbf{x}$ into a probability (stochastic) vector in which all entries are nonnegative and add up to one.

   (a) Please find the gradient and Jacobian matrix of $\mathbf{f}(\mathbf{x})$, i.e., $\nabla \mathbf{f}(\mathbf{x})$ and $J\mathbf{f}(\mathbf{x})$.

   (b) Show that $\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{x} - c\mathbf{1})$, where $c = \max\{x_1, x_2, ..., x_n\}$ and $\mathbf{1}$ is a vector all of whose components are one. When do we need this transformation?

2. Consider the neural network with a single hidden layer in Figure 2. Let $\mathbf{x} \in \mathbb{R}^3$ be an input vector, and $\mathbf{y}$ be its corresponding output of the network. f implies that there exist four units in the hidden layer, each of which is followed by a sigmoid activation function σ, converting its input $\mathbf{z}$ to output $\mathbf{a}$. Suppose that the ground truth label vector of $\mathbf{x}$ is $[0, 0, 1]^\top$ and we use the cross entropy introduced in Lecture 15 as the loss function.

   (a) Please find the update formula for the j^{th} weight of the i^{th} hidden unit , i.e., w_{ij}^1 where $i \in \{1, 2, 3\}$ and $j \in \{1, 2, 3, 4\}$.

   (b) Can we initialize all the parameters, i.e., weights and bias, of the neural network to zero? Please state you conclusion.

3. Consider a convolutional neural network as shown in Table 1.

   (a) The convolutional layer parameters are denoted as "conv⟨filter size⟩-⟨number of filters⟩".

   (b) The fully connected layer parameters are denoted as "FC⟨number of neurons⟩".

   (c) The window size of pooling layers is 2.

   (d) The stride of convolutinal layers is 1.

   (e) The stride of pooling layers is 2.

   (f) You may want to use padding in both convolutional and pooling layers if necessary.

   (g) For convenience, we assume that there is no activation function and bias.

   Suppose that the input is a $\mathbf{210} \times \mathbf{160}$ **RGB** image. Please derive the size of all feature maps and the number of parameters.

conv3-32	conv5-32	max pool	conv3-64	conv5-64	max pool	FC-128	FC-10

Table 1: The architecture of convolutional neural network

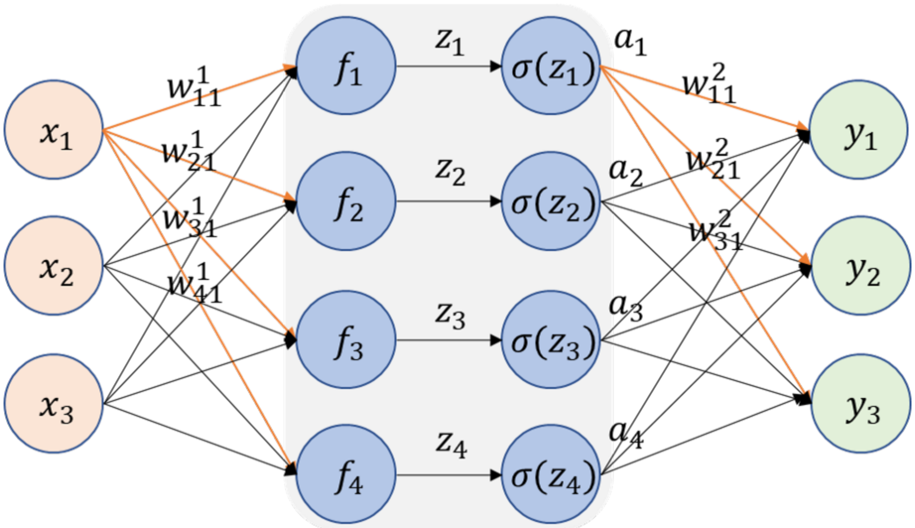

Figure 2: A neural network with a single hidden layer.

Solution:

■