Quiz 2: Introduction to Machine Learning

University of Science and Technology of China

Dec. 2, 2019

Name: ID:

START HERE: Instructions

- This exam has **14** pages (including this one) and **4** Problems. Please make sure that none of the pages is missing. Fill in your name and ID above.
- Note that the problems vary in difficulty. Make sure to look over the entire exam before you start. It might be a good idea to answer the easier ones first.
- **Notice,** to get the full credits, please show your solutions step by step.

Question	**Point**	**Score**
1	10	
2	30	
3	30	
4	30	
Total	100	
Bonus	20	

Problem 1: Naive Bayes Classifier 10pts

Train a naive bayes classifier on the dataset in Table 1. The dataset in Table 1 consists of 15 samples. For each sample, we can observe the features $X^{(1)} \in \{1, 2, 3\}, X^{(2)} \in \{S, M, L\}$ and the corresponding response $Y \in \{-1, 1\}$. Note that $X^{(1)}, X^{(2)}$ are independient conditioned on the Y.

1. Compute the number of probabilities you need to estimate when training the classifier.
2. Predict the class of $x = (2, S)^\top$.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$X^{(1)}$	1	1	1	2	2	3	1	1	2	2	2	3	3	3	3
$X^{(2)}$	S	M	S	S	M	L	M	S	M	L	L	L	M	M	L
Y	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1	1

Table 1: The training data

Solution:

Since $X^{(1)}, X^{(2)}$ are independient conditioned on the Y,

$$\mathbb{P}(X^{(1)}, X^{(2)}|Y) = \mathbb{P}(X^{(1)}|Y)\mathbb{P}(X^{(2)}|Y)$$

1. The number of probabilities need to estimate is

$$\begin{aligned} N &= 2 + 2 \times (3 + 3) \\ &= 14 \end{aligned}$$

2. We have

$$\begin{aligned} \mathbb{P}(y = -1|x = (2, S)^T) &\propto \mathbb{P}(y = -1)\mathbb{P}(x^{(1)} = 2|y = -1)\mathbb{P}(x^{(2)} = S|y = -1) \\ &= \frac{1}{15}, \\ \mathbb{P}(y = 1|x = (2, S)^T) &\propto \mathbb{P}(y = 1)\mathbb{P}(x^{(1)} = 2|y = 1)\mathbb{P}(x^{(2)} = S|y = 1) \\ &= \frac{1}{45}. \end{aligned}$$

Since $\frac{\mathbb{P}(y=-1|x=(2,S)^T)}{\mathbb{P}(y=1|x=(2,S)^T)} = \frac{45}{15} > 1$, the class of x is -1. ■

Problem 2: Logistic Regression 30pts

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

$$\begin{aligned}\mathcal{I}^+ =&\{i : i \in [n], y_i = 1\},\\ \mathcal{I}^- =&\{i : i \in [n], y_i = -1\},\end{aligned}$$

where $[n] = \{1, 2, \ldots, n\}$. We assume that $\mathcal{I}^+$ and $\mathcal{I}^-$ are nonempty.

Then, we can formulate the logistic regression as:

$$\min_{\mathbf{w}} \; L(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \log(1 + \exp(-\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle)), \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^{d+1}$ is the model parameter to be estimated and $\overline{\mathbf{x}}_i^\top = y_i \cdot (1, \mathbf{x}_i^\top)$.

1. (15pts) Please find the dual problem of (1). (*Hint: add constraints by letting* $q_i = \frac{1}{n}(-\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle)$, $i = 1, \ldots, n$.)

2. (15pts) Suppose that the training data is linearly separable, that is, there exists $\hat{\mathbf{w}} \in \mathbb{R}^{d+1}$ such that

   $$\langle \hat{\mathbf{w}}, \overline{\mathbf{x}}_i\rangle > 0, \, \forall \, i = 1, \ldots, n.$$

   Show that problem (1) has no solution.

3. (**Bonus** 10pts) Suppose that the training data is NOT linearly separable, that is,

   $$\forall \, \mathbf{w} \in \mathbb{R}^{d+1}, \exists \, i \in \{1, \ldots, n\} \text{ s.t. } \langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle < 0.$$

   Show that problem (1) always admits a solution.

Solution: 1. Let $q_i = \frac{1}{n}\left(-\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle\right)$ and $\mathbf{q} = (q_1, \ldots, q_n)$. The primal problem can be formulated as

$$\begin{aligned}&\min_{\mathbf{w},\mathbf{q}} \frac{1}{n}\sum_{i=1}^n \log(1 + \exp(nq_i)),\\ &\text{s.t. } \frac{1}{n}\left(-\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle\right) - q_i = 0, \, i = 1, \ldots, n.\end{aligned}$$

We first construct the Lagrangian:

$$\mathcal{L}(\mathbf{w}, \mathbf{q}, \alpha) = \frac{1}{n}\sum_{i=1}^n \log(1 + \exp(nq_i)) + \sum_{i=1}^n \alpha_i \left(\frac{1}{n}\left(-\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle\right) - q_i\right).$$

We next find the dual function:

$$\begin{aligned} h(\alpha) &= \inf_{\mathbf{w},\mathbf{q}} \mathcal{L}(\mathbf{w}, \mathbf{q}, \alpha) \\ &= \inf_{\mathbf{w}}(-\frac{1}{n}\sum_{i=1}^{n} \alpha_i(\langle \mathbf{w}, \overline{\mathbf{x}}_i \rangle)) + \sum_{i=1}^{n} \inf_{q_i}(\frac{1}{n}\log(1+\exp(nq_i)) - \alpha_i q_i) \end{aligned}$$

Let

$$l(\mathbf{w}) = -\frac{1}{n}\sum_{i=1}^{n} \alpha_i(\langle \mathbf{w}, \overline{\mathbf{x}}_i \rangle))$$

and

$$h_{\alpha_i}(q_i) = \frac{1}{n}\log(1+\exp(nq_i)) - \alpha_i q_i.$$

Then,

$$h(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \inf_{q_i} h_{\alpha_i}(q_i) + \inf_{\mathbf{w}} l(\mathbf{w}).$$

For $\mathbf{w}$,

$$\nabla_{\mathbf{w}} l(\mathbf{w}) = \sum_{i=1}^{n} \alpha_i \overline{\mathbf{x}}_i$$

If $\sum_{i=1}^{n} \alpha_i \overline{\mathbf{x}}_i = 0$, then

$$\inf_{\mathbf{w}} l(\mathbf{w}) = 0.$$

If $\sum_{i=1}^{n} \alpha_i \overline{\mathbf{x}}_i \neq 0$, then

$$\inf_{\mathbf{w}} l(\mathbf{w}) = -\infty.$$

For q_i,

$$\frac{\partial h_{\alpha_i}(q_i)}{\partial q_i} = \frac{\exp(nq_i)}{1+\exp(nq_i)} - \alpha_i.$$

Thus, if $\alpha_i \in (0,1)$, then

$$\begin{aligned} &\frac{\exp(nq_i)}{1+\exp(nq_i)} - \alpha_i = 0 \\ \Rightarrow &q_i = \frac{1}{n}\log(\frac{1}{1-\alpha_i} - 1). \end{aligned}$$

Thus,

$$\begin{aligned}\inf_{q_i} h_{\alpha_i}(q_i) &= h_{\alpha_i}(q_i)|_{q_i=\frac{1}{n}\log(\frac{1}{1-\alpha_i}-1)} \\ &= -\frac{1}{n}(\alpha_i \log \alpha_i + (1-\alpha_i)\log(1-\alpha_i));\end{aligned}$$

if $\alpha_i = 0$, then

$$\begin{aligned}\inf_{q_i} h_{\alpha_i}(q_i) &\lim_{q_i \to -\infty}(\frac{1}{n}\log(1+\exp(nq_i))) \\ &= 0.\end{aligned}$$

If $\alpha_i = 1$, then we show that $h_{\alpha_i}(q_i)$ is decreasing. Note that

$$\begin{aligned}h'_{\alpha_i}(q_i) &= \frac{\exp(nq_i)}{1+\exp(nq_i)} - 1 \\ &< 0.\end{aligned}$$

Thus,

$$\begin{aligned}\inf_{q_i} h_{\alpha_i}(q_i) &= \inf_{q_i}(\frac{1}{n}\log(1+\exp(nq_i)) - q_i) \\ &= \lim_{q_i \to +\infty}(\frac{1}{n}\log(1+\exp(nq_i)) - q_i) \\ &= 0.\end{aligned}$$

If $\alpha_i > 1$, then we first show that

$$\log(1+\exp(nx)) < nx + 1$$

for all $x > 0$. Suppose that $g(x) = \log(1+\exp(nx)) - nx - 1$, then

$$\begin{aligned}g'(x) &= \frac{n\exp(nx)}{1+\exp(nx)} - n \\ &< 0.\end{aligned}$$

Thus, $g(x)$ is strictly decreasing, which implies that

$$\begin{aligned}g(x) &< g(0) \\ &= 0\end{aligned}$$

for all $x > 0$. Thus,

$$h_{\alpha_i}(q_i) < q_i + \frac{1}{n} - \alpha_i q_i$$

for all $q_i > 0$. That is,

$$\begin{aligned}\lim_{q_i \to +\infty} h_{\alpha_i}(q_i) &\leq \lim_{q_i \to +\infty} (q_i + \frac{1}{n} - \alpha_i q_i) \\ &= -\infty,\end{aligned}$$

Thus,

$$\begin{aligned}\inf_{q_i} h_{\alpha_i}(q_i) &= \lim_{q_i \to +\infty} h_{\alpha_i}(q_i) \\ &= -\infty.\end{aligned}$$

If $\alpha_i < 0$, then

$$\begin{aligned}\inf_{q_i} h_{\alpha_i}(q_i) &= \lim_{q_i \to -\infty} (\frac{1}{n}\log(1 + \exp(nq_i)) - \alpha_i q_i) \\ &= -\infty.\end{aligned}$$

Thus

$$\textbf{dom } h(\alpha) = \{\alpha : h(\alpha) > -\infty\} = \{\alpha : \alpha_i \in [0,1], \sum_{i=1}^{n} \alpha_i \overline{\mathbf{x}}_i = 0\}.$$

Therefore, the dual problem is

$$\begin{aligned}\max_{\alpha} \ & -\frac{1}{n}\sum_{i=1}^{n}(\alpha_i \log \alpha_i + (1-\alpha_i)\log(1-\alpha_i)), \\ \text{s.t. } & \alpha_i \in [0,1], \\ & \sum_{i=1}^{n} \alpha_i \overline{\mathbf{x}}_i = 0,\end{aligned}$$

where $0\log(0) = 0$.

2. Since the training data is linearly separable, there exists $\hat{\mathbf{w}} \in \mathbb{R}^{d+1}$ such that

$$\langle \hat{\mathbf{w}}, \bar{\mathbf{x}}_i \rangle > 0,\ i = 1, \ldots, n.$$

Consider $\lambda\hat{\mathbf{w}}$ with $\lambda > 0$. Then,

$$\lim_{\lambda \to +\infty} \langle \lambda\hat{\mathbf{w}}, \bar{\mathbf{x}}_i \rangle = +\infty,\ i = 1, \ldots, n.$$

$$\begin{aligned}&\lim_{\lambda \to +\infty} L(\lambda\hat{\mathbf{w}}) = \lim_{\lambda \to +\infty} \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-\langle \lambda\hat{\mathbf{w}}, \overline{\mathbf{x}}_i \rangle)) = 0 \\ &\Rightarrow \inf_{\mathbf{w}} L(\mathbf{w}) \leq 0.\end{aligned}$$

Let $\mathbf{w}^*$ be a solution to (1). Thus

$$\begin{aligned} & L(\mathbf{w}^*) \leq 0 \\ \Rightarrow & \log(1+\exp(-\langle \mathbf{w}^*, \overline{\mathbf{x}}_i \rangle)) = 0, \, \forall i = 1, \ldots, n. \end{aligned} \tag{2}$$

Therefore, the equation (2) has no solution and hence the problem (1) has no solution.

3. The training data is NOT linearly separable, that is,

$$\forall \, \mathbf{w} \in \mathbb{R}^{d+1}, \exists \, i \in [n] \;\; \text{s.t.} \;\; \langle \mathbf{w}, \bar{\mathbf{x}}_i \rangle < 0. \tag{3}$$

Note that the inverse proposition of (3) does not imply the training data is linearly separable.

Let $\overline{\mathbf{X}} = (\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, \ldots, \overline{\mathbf{x}}_n)$. Suppose $\mathbf{rank}(\overline{\mathbf{X}}) = r \leq d+1$. Let $[\,\cdot\,]_{1:r}$ denote the first r entries of a vector.

WLOG, we assume that $\overline{\mathbf{X}}_r = ([\overline{\mathbf{x}}_1]_{1:r} \quad \ldots \quad [\overline{\mathbf{x}}_r]_{1:r})^\top$ and $\mathbf{rank}(\overline{\mathbf{X}}_r) = r$. Thus $\overline{\mathbf{X}} = \overline{\mathbf{X}}_r(\mathbf{I}_r \quad \mathbf{M})$, where $\mathbf{M} \in \mathbf{R}^{r \times (d+1-r)}$. Define

$$L^*(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} \log(1+\exp(-\langle \mathbf{u}, [\overline{\mathbf{x}}_i]_{1:r} \rangle))$$

Note that $L^*((\mathbf{I}_r \quad \mathbf{M})\mathbf{w}) = L(\mathbf{w})$ and $\{([\mathbf{x}_i]_{1:r}, y_i)\}_{i=1}^n$ is not linearly separable.

Now we show that $L^*(\mathbf{u})$ always admits a solution by following steps.

(a) Let $U = \{\mathbf{u} : \|\mathbf{u}\| = 1\}$. Let $g : U \to R$ defined by

$$g(\mathbf{u}) = \max_i(-\langle \mathbf{u}, [\mathbf{x}_i]_{1:r} \rangle), \; \mathbf{u} \in U.$$

Note that $g(\mathbf{u}) > 0$. As $g(\mathbf{u})$ is continuous and U is a compact set, g attains its minimum in U, i.e.

$$\exists \, \mathbf{u}_0 \in U, \, g(\mathbf{u_0}) = \min_{\mathbf{u}} g(\mathbf{u}).$$

(b) Since $\lim_{\alpha \to +\infty} \frac{1}{n} \log(1+\exp(\alpha g(\mathbf{u}_0))) = +\infty$, $\exists \, \alpha_0 > 0$ such that

$$\frac{1}{n} \log(1+\exp(\alpha_0 g(\mathbf{u}_0))) \geq L^*(0).$$

For all $\alpha > \alpha_0, \mathbf{u} \in U$, we have

$$\begin{aligned} L^*(\alpha\mathbf{u}) &\geq \frac{1}{n}\log(1+\exp(g(\alpha\mathbf{u}))) \\ &= \frac{1}{n}\log(1+\exp(\alpha g(\mathbf{u}))) \\ &\geq \frac{1}{n}\log(1+\exp(\alpha g(\mathbf{u}_0))) \\ &\geq \frac{1}{n}\log(1+\exp(\alpha_0 g(\mathbf{u}_0))) \\ &\geq L^*(0) \end{aligned}$$

Define $B_{\alpha_0}(0) = \{\mathbf{u} : \|\mathbf{u}\| \leq \alpha_0\}$. Thus, for all $u \notin B_{\alpha_0}(0)$,

$$L^*(\mathbf{u}) \geq L^*(0).$$

That is, $L^*(\mathbf{u})$ attains its minimum in $B_{\alpha_0}(0)$.

Let $(\mathbf{w}^*) = \begin{pmatrix} \mathbf{u}^* \\ \mathbf{0} \end{pmatrix}$. Next, we show that $L(\mathbf{w}^*) = \min_{\mathbf{w}} L(\mathbf{w})$. We have

$$\begin{aligned} L(\mathbf{w}) &= L^*((\mathbf{I}_r \quad \mathbf{M})\mathbf{w}) \\ &\geq L^*(\mathbf{u}^*) \\ &= L(\mathbf{w}^*). \end{aligned}$$

Therefore, $L(\mathbf{w})$ attains its minimum at $\mathbf{w}^*$.

Indeed, the minimum of $L(\mathbf{w})$ is not unique. If $\mathbf{M} = \mathbf{0}$, then $(\mathbf{w}^*) = \begin{pmatrix} \mathbf{u}^* \\ \mathbf{1} \end{pmatrix}$ is also an optimal solution.

It follows from the exercise 1.2 that the solution is unique if $\mathbf{rank}(\overline{\mathbf{X}}) = d + 1$.

■

Problem 3: Duality Gap 30pts

Consider the primal problem as follows.

$$\begin{aligned} \min_{\mathbf{x}} \; & f(\mathbf{x}) \\ \text{s.t. } & \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \\ & \mathbf{h}(\mathbf{x}) = \mathbf{0}, \end{aligned} \tag{4}$$

where $\mathbf{x} \in \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$, and $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^p$. The functions f, $\mathbf{g}$, and $\mathbf{h}$ are continuously differentiable.

The Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ associated with the problem in (4) takes the form of

$$L(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \mu_i h_i(\mathbf{x}). \tag{5}$$

The dual function $q : \mathbb{R}^m \times \mathbb{R}^p$ is defined by

$$q(\lambda, \mu) = \inf_{\mathbf{x}} L(\mathbf{x}, \lambda, \mu).$$

The dual problem is

$$\begin{aligned} \sup_{\lambda,\, \mu} \; & q(\lambda, \mu), \\ \text{s.t. } & \lambda \geq 0. \end{aligned} \tag{6}$$

Duality gap is defined by

$$f^* - q^*.$$

1. (a) (10pts) Show that there is no duality gap if the set of geometric multipliers is nonempty.

   (b) (10pts) Is the set of optimal dual solutions nonempty if there is no duality gap? If so, please show it is true. Otherwise, please give a counterexample.

2. (10pts) Show that the set of geometric multipliers is equal to the set of dual optimal solutions if there is no duality gap.

Solution:

1. (a) Suppose (λ^*, μ^*) is a geometric multiplier, then

$$\inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \mu^*) = f^*.$$

As the weak duality $q^* \leq f^*$ holds, we have

$$q^* = q(\lambda^*, \mu^*) = f^*$$

Thus,

$$f^* - q^* = 0$$

That means there is no duality gap.

(b) The set of optimal dual solutions may be empty even if the duality gap is 0, and we can support this claim by a counter example.

The primal problem is:

$$\begin{aligned} &\inf_x -x \\ &\text{s.t. } x^2 \leq 0, \end{aligned}$$

and its dual problem is

$$\sup_{\lambda>0} -\frac{1}{4\lambda}.$$

Clearly, $f^* = q^* = 0$ and the duality gap is 0, and the optimal solution to the primal problem exists. However, the dual optimal solution doesn't exist. Thus, the set of optimal dual solutions may be empty even if the duality gap is 0.

2. Duality gap is zero implies

$$f^* = q^* \tag{7}$$

We first show that the set of geometric multipliers belongs to the set of dual optimal solutions. Let (λ^*, μ^*) be a geometric multiplier. Thus

$$q(\lambda^*, \mu^*) = \inf_{x \in X} L(\mathbf{x}, \lambda^*, \mu^*) = f^* = q^*.$$

Therefore, (λ^*, μ^*) is a dual optimal solution.

We next show that the set of dual optimal solutions belongs to the set of geometric multipliers. Let (λ^*, μ^*) be a dual optimal solution.

Since

$$\inf_{x \in X} L(\mathbf{x}, \lambda^*, \mu^*) = q(\lambda^*, \mu^*) = q^* = f^*,$$

(λ^*, μ^*) is a geometric multiplier for the primal problem.

Therefore, the set of geometric multipliers is equal to the set of dual optimal solutions.

■

Problem 4: Support Vector Machine (SVM) 30pts

Given the training sample $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. The soft margin SVM takes the form of

$$\begin{aligned} &\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n \xi_i, \\ &\text{s.t. } y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i, i = 1, \ldots, n, \\ &\quad \xi_i \geq 0,\ i = 1, \ldots, n, \end{aligned} \tag{8}$$

where $C > 0$. The corresponding dual problem is

$$\begin{aligned} &\min_{\alpha} \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j\rangle - \sum_{i=1}^n \alpha_i \\ &\text{s.t. } \sum_{i=1}^n \alpha_i y_i = 0, \\ &\quad \alpha_i \in [0, C], i = 1, \ldots, n. \end{aligned} \tag{9}$$

1. (10pts) Show that the problems (8) and (9) always admit optimal solutions.
2. (10pts) Let $(\mathbf{w}^*, b^*, \xi^*)$ be an optimal solution to the problem (8).
   (a) Show that $\mathbf{w}^*$ is unique.
   (b) Is b^* unique? If so, please show that this claim is true. Otherwise, please give a counterexample.
3. (10pts) Let α^* be one of the optimal solutions to the problem (9). Suppose that α_k^* is one of the entries of α^* and $\alpha_k^* \in (0, C)$. Please find a primal optimal solution of (8).
4. (**Bonus** 10pts) Suppose that the training sample $\mathcal{D}$ is linearly separable and there exist $i, j \in \{1, \ldots, n\}$ such that $y_i = 1$ and $y_j = -1$. Let $(\mathbf{w}_1^*, b_1^*)$ be an optimal solution to the hard margin SVM and $(\mathbf{w}_2^*, b_2^*, \xi^*)$ be an optimal solution to the soft margin SVM. Is $(\mathbf{w}_1^*, b_1^*) = (\mathbf{w}_2^*, b_2^*)$ true? If so, please show it is true. Otherwise, please give a counterexample.

Solution:

1. First, we show that the feasible set is nonempty. For any $\overline{\mathbf{w}} \in \mathbb{R}^n, \overline{b} \in \mathbb{R}$, let $\overline{\xi}_i = \max\{0, 1 - y_i(\langle \overline{\mathbf{w}}, \mathbf{x}_i\rangle + \overline{b})\}$. Clearly, $(\overline{\mathbf{w}}, \overline{b}, \overline{\xi})$ is a feasible solution.

   Indeed, $f(\mathbf{w}, b, \xi) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n \xi_i$ is a convex quadratic function, the optimal value f^* is finite since

$$\begin{cases} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n \xi_i \geq 0 & \Rightarrow f^* > -\infty, \\ \text{the feasible set is nonempty} & \Rightarrow f^* < +\infty. \end{cases}$$

and $\mathbf{dom}(f)$ is a polyhedral.

Thus, the Linear and Quadratic Programming Duality Theorem holds for the soft margin SVM, which means problems (8) and (9) always admit optimal solutions.

2. (a) Suppose $(\mathbf{w}_1^*, b_1^*, \xi_1^*)$ and $(\mathbf{w}_2^*, b_2^*, \xi_2^*)$ are two different optimal solutions, i.e.,

$$f(\mathbf{w}_1^*, b_1^*, \xi_1^*) = f(\mathbf{w}_2^*, b_2^*, \xi_2^*) = \min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i = f^*.$$

Let

$$(\mathbf{w}^*, b^*, \xi^*) = \frac{1}{2}(\mathbf{w}_1^*, b_1^*, \xi_1^*) + \frac{1}{2}(\mathbf{w}_2^*, b_2^*, \xi_2^*).$$

Clearly, $(\mathbf{w}^*, b^*, \xi^*)$ is also a feasible solution to the primal problem.
Then we have

$$\begin{aligned}
0 &\le f(\mathbf{w}^*, b^*, \xi^*) - f^* \\
&= f(\mathbf{w}^*, b^*, \xi^*) - \frac{1}{2}f(\mathbf{w}_1^*, b_1^*, \xi_1^*) - \frac{1}{2}f(\mathbf{w}_2^*, b_2^*, \xi_2^*) \\
&= \frac{1}{2}\left(\|\frac{1}{2}\mathbf{w}_1^* + \frac{1}{2}\mathbf{w}_2^*\|^2 - \frac{1}{2}\|\mathbf{w}_1^*\|^2 - \frac{1}{2}\|\mathbf{w}_2^*\|^2\right) \\
&= -\frac{1}{8}\|\mathbf{w}_1^* - \mathbf{w}_2^*\|^2 \\
&\le 0.
\end{aligned}$$

That is

$$\mathbf{w}_1^* = \mathbf{w}_2^*$$

Thus, $\mathbf{w}^*$ is unique.

(b) Counterexample:
Consider the training data consists of four data instances: $\mathbf{x}_1 = (2, 3)$, $\mathbf{x}_2 = (1, 2)$, $\mathbf{x}_3 = (1, 3)$, and $\mathbf{x}_4 = (2, 2)$, and the corresponding labels are $y_1 = y_2 = 1$ and $y_3 = y_4 = -1$. (XOR problem)
Then the set of primal optimal solutions is

$$\{(\mathbf{w}, b, \xi) : \mathbf{w} = \mathbf{0},\ \xi_1 = \xi_2 = 1 - t,\ \xi_3 = \xi_4 = 1 + t,\ b = t,\ t \in [-1, 1]\}.$$

Here b^* is not unique.

3. The Lagrange function of the primal problem is

$$L(\mathbf{w}, b, \xi, \alpha, \mu) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i(1 - \xi_i - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)) - \sum_{i=1}^{n} \mu_i \xi_i.$$

where $\alpha_i, \mu_i \ge 0$, $i = 1, \ldots, n$.

The optimal dual solution (α^*, μ^*) is also a geometric multiplier, i.e.,

$$\begin{aligned} f^* &= \inf_{\mathbf{w},b,\xi} L(\mathbf{w}, b, \xi, \alpha^*, \mu^*) \\ &= \inf_{\mathbf{w}} \left(\frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i^* y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \right) + \inf_b b \sum_{i=1}^n \alpha_i^* y_i + \inf_\xi \sum_{i=1}^n (C - \alpha_i^* - \mu_i^*) \xi_i. \end{aligned}$$

Then the first order optimal condition implies that $\nabla_{\mathbf{w}} L(\mathbf{w}^*, b^*, \xi^*, \alpha^*, \mu^*) = \mathbf{0}$, $\nabla_b L(\mathbf{w}^*, b^*, \xi^*, \alpha^*, \mu^*) = 0$, $\nabla_{\xi_i} L(\mathbf{w}^*, b^*, \xi^*, \alpha^*, \mu^*) = 0$, i.e.,

$$\begin{aligned} \mathbf{w}^* - \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i &= \mathbf{0}, \\ \sum_{i=1}^n \alpha_i^* y_i &= 0, \\ C - \alpha_i^* - \mu_i^* &= 0. \end{aligned}$$

Thus, we have

$$\mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i.$$

Suppose $\alpha_k^* \in (0, C)$, from the equation

$$C - \alpha_k^* - \mu_k^* = 0,$$

we have $\mu_k^* \in (0, C)$.

As the complementary slackness holds, we have

$$\begin{aligned} y_i(\langle \mathbf{w}^*, \mathbf{x}_k \rangle + b) &= 1 - \xi_k^*, \\ \xi_k^* &= 0. \end{aligned}$$

Thus

$$b^* = y_k - \langle \mathbf{w}^*, \mathbf{x}_k \rangle,$$

and

$$\xi_i^* = \max\{1 - y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*), 0\}.$$

4. Counterexample:

   Consider two samples $(x_1, y_1) = (1, 1)$ and $(x_2, y_2) = (-1, -1)$.

   For hard margin SVM, the optimal solution is $(w^*, b^*) = (1, 0)$.

   However, for soft margin SVM with $C = 0.01$, $(w, b, \xi_1, \xi_2) = (1, 0, 0, 0)$ is not the optimal solution.

■