Introduction to Machine Learning

Fall 2023

University of Science and Technology of China

Lecturer: Jie Wang | Homework 6

Posted: Dec. 10, 2023 | Due: Dec. 24, 2023

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. Please show that Problem (2) admits an optimal solution.

4. Please show that Problems (1) and (2) share the same set of optimal solutions.

5. Let $(\mathbf{w}^*, b^*)$ be the optimal solution to Problem (2). Please show that

   (a) If the training set $\mathcal{D}$ is linearly separable, we have $\mathbf{w}^* \neq 0$;

   (b)If all samples in training set $\mathcal{D}$ are positive or negative, then $\mathbf{w}^*$ can be 0.

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 and there is at least one of the constraints holds as an equality at the optimum.

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

Solution: ■

Exercise 2: Discussions on Geometric Multiplier and Duality Gap

Consider the primal problem

$$\begin{aligned} \min_{\mathbf{x}} \ & f(\mathbf{x}) \\ \text{s.t. } & g_i(\mathbf{x}) \leq 0,\ i = 1, \cdots, m, \\ & h_i(\mathbf{x}) = 0, i = 1, \cdots, p, \\ & \mathbf{x} \in X. \end{aligned} \tag{3}$$

Let

$$S = \{(\mathbf{g}(\mathbf{x}), \mathbf{h}(\mathbf{x}), f(\mathbf{x})) : \mathbf{x} \in X\} \subset \mathbb{R}^{m+p+1}. \tag{4}$$

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 (3) 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 (3).
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 $\mathbf{x}^* \in \mathbf{argmin}_{\mathbf{x}\in X}\ L(\mathbf{x}, \lambda^*, \mu^*)$.

Exercise 3: 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

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \ & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^n \xi_i, \\ \text{s.t. } & y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i, i = 1, \ldots, n, \\ & \xi_i \geq 0,\ i = 1, \ldots, n, \end{aligned} \tag{5}$$

The corresponding dual problem is

$$\begin{aligned} \min_{\alpha} \ & \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j\rangle - \sum_{i=1}^n \alpha_i \\ \text{s.t. } & \sum_{i=1}^n \alpha_i y_i = 0, \\ & \alpha_i \in [0, C], i = 1, \ldots, n. \end{aligned} \tag{6}$$

1. Show that the problems (5) and (6) always admit optimal solutions.
2. Let $(\mathbf{w}^*, b^*)$ be the solution to (5) and α^* be the corresponding solution to (6).
   (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.
3. For a linearly separable data sample, shall we arrive at the same separating hyperplane by solving the problems in (2) and (6), respectively?

Solution: ■

Exercise 4: An Example of the Soft Margin SVM

Recall that the soft margin SVM takes the form of

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \ & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i, \\ \text{s.t. } & y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i, i = 1, \ldots, n, \\ & \xi_i \geq 0,\ i = 1, \ldots, n, \end{aligned} \tag{7}$$

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

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \ & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i^p, \\ \text{s.t. } & y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i, i = 1, \ldots, n, \\ & \xi_i \geq 0,\ i = 1, \ldots, n. \end{aligned} \tag{8}$$

   Please find the dual problem of (8) 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 (8).

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. (Optional) 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?

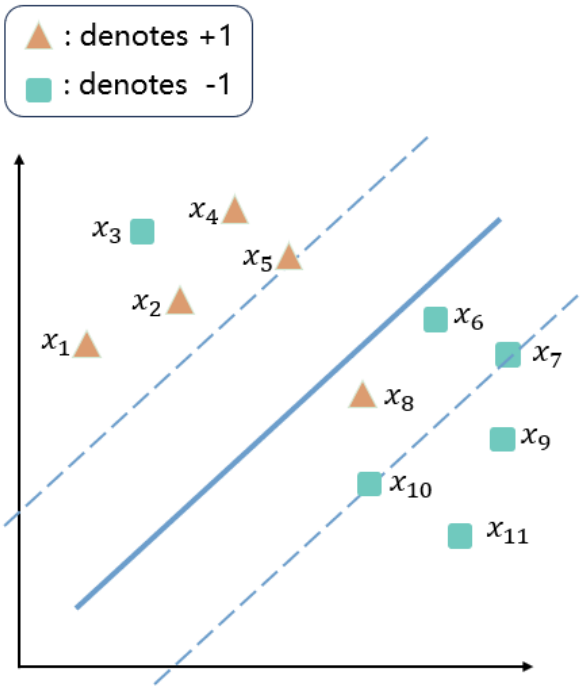

Figure 1: Classifying the data points using the soft margin SVM. H_0 is the separating hyperplane. H_1 and H_2 are the marginal hyperplanes.

Solution: ■

Exercise 5: 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 **210 × 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

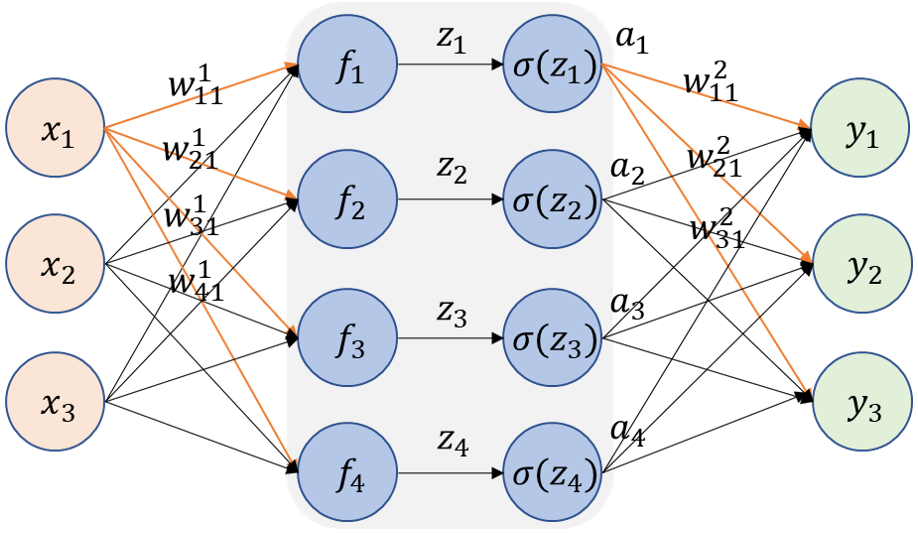

Figure 2: A neural network with a single hidden layer.

Solution:

■

Exercise 6: Exercises of Dual Problems

Consider the optimization problem

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

1. Give the feasible set, the optimal value, and the optimal solution $x^\star$.
2. Please write the dual problem of the primal problem, and verify that it is a concave maximization problem.
3. Please give the KKT conditions.
4. Please find the dual optimal value and dual optimal solution $\lambda^\star$. Does strong duality hold?
5. Consider the set $\mathcal{S} = \{(g(x), f(x)) | x \in \mathbb{R}\}$, please verify that the norm vector $(\lambda^\star, 1)$ supports $\mathcal{S}$ at $(g(x^\star), f(x^\star))$.

Solution: ■

Exercise 7: Some Network Layers, Linear Transformation and Gradient (Optional)

In this exercise, we explore several kinds of network layers in the view of linear transformation.

1. **1-dimensional convolutional layer.** Suppose we have an input $\mathbf{x} \in \mathbb{R}^n$ and filter $\mathbf{w} \in \mathbb{R}^k$ $(n > k)$. We can compute the convolution of $\mathbf{x} * \mathbf{w}$ as follows:

   - Take the convolutional filter $\mathbf{w}$ and align it with the beginning of $\mathbf{x}$. Take the dot product of $\mathbf{w}$ and the $\mathbf{x}[0 : k-1]$ and assign that as the first entry of the output.
   - Suppose we have stride s. Shift the filter down by s indices, and now take the dot product of $\mathbf{w}$ and $\mathbf{x}[s : k-1+s]$ and assign to the next entry of your output.
   - Repeat until we run out of entries in $\mathbf{x}$.

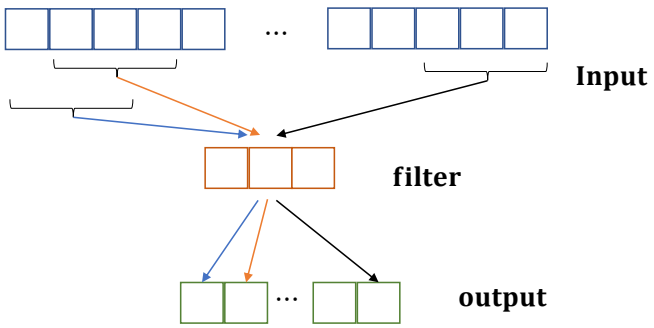

Figure 3: 1-dimensional convolutional layer.

   Now we set the stride s to be 1:

$$\mathbf{y} = \mathbf{x} * \mathbf{w} = \left(\sum_{i=1}^{k} w_i x_i, \sum_{i=1}^{k} w_i x_{i+1}, \cdots, \sum_{i=1}^{k} w_i x_{i+n-k} \right) \in \mathbb{R}^{n-k+1}.$$

   Is the 1-dimensional convolutional operation a linear transformation? If so, please find the transformation matrix, then write down the gradient with respective to $\mathbf{x}$.

2. 1×1 **convolutional layer.** Convolutional operations are linear transformations. We study a simple case, 1×1 convolutional operation, in this question. Suppose a convolutional layer takes as inputs the RGB $3 \times 28 \times 28$ images $\mathbf{X} = (x_{ijk}) \in \mathbb{R}^{3 \times 28 \times 28}$. Suppose that the convolutional layer has three $3 \times 1 \times 1$ filters where the i^{th} filter is denoted by $\mathbf{w}_i \in \mathbb{R}^3$. We set stride $= 1$ and padding $= 0$.

   Specifically, we denote the output by $\mathbf{Y} = (y_{ijk}) \in \mathbb{R}^{3 \times 28 \times 28}$, then

$$y_{ijk} = \sum_{t=1}^{3} w_{it} x_{tjk}, \quad i \in \{1, 2, 3\}, j, k \in \{1, \cdots, 28\}.$$

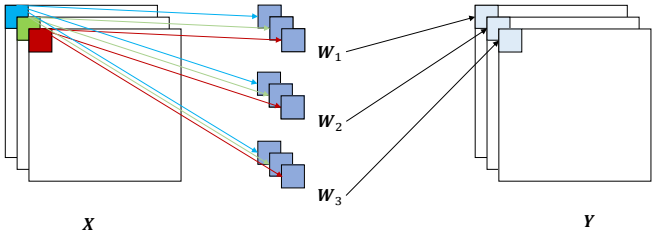

Figure 4: 1×1 convolutional layer.

Now we flatten the output $\mathbf{Y}$ to attain a $3 \times 28 \times 28$-dimensional vector,

$$\mathbf{y} = (y_{1,1,1}, y_{1,1,2}, \cdots, y_{1,1,28}, y_{1,2,1}, y_{1,2,2}, \cdots, y_{1,28,28}, y_{2,1,1}, y_{2,1,2}, \cdots, y_{3,28,28}).$$

We can also flatten $\mathbf{X}$ to attain a $3 \times 28 \times 28$-dimensional vector $\mathbf{x}$.

(a) Is the 1×1 convolutional operation a linear transformation? If so, Please find the transformation matrix.

(b) Please show that the 1×1 convolutional operation is invertible if and only if the matrix $(\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3)$ is invertible.
Hint: let $A = (a_{ij})_{m \times m}$, $B \in \mathbb{R}^{n \times n}$, then the $mn \times mn$ matrix

$$\begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1m}B \\ a_{21}B & a_{22}B & \cdots & a_{2m}B \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mm}B \end{pmatrix}$$

is called the Kronecker product of A and B, denoted by $A \otimes B$. Furthermore, $\det(A \otimes B) = (\det(A))^n (\det(B))^m$.

(c) Suppose $\mathbf{x}$ is sampled from a standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$, please find the density function of $\mathbf{y}$ if the 1×1 convolutional operation is invertible.

3. **Pooling layer.** We know that average pooling and overlapping pooling are linear transformations, but not the max pooling.

(a) Suppose an average pooling layer has window size 2×2 and stride 2, as illustrated in the following picture. The pooling layer takes as inputs the 4×4 matrices. Please find the transformation matrix of the average pooling layer.

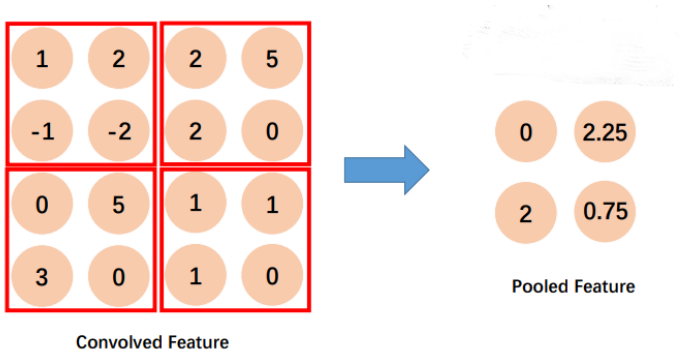

Figure 5: average pooling.

(b) Max pooling is generally not linear transformation. Consider the following example we studied in this course. The max pooling layer has window size 2×2 and stride 2, as illustrated in the following picture. The pooling layer takes as inputs the 4×4 matrices. Please find the subgradient of the max pooling operation. Then give an explanation of the "gradient" we studied in our course.

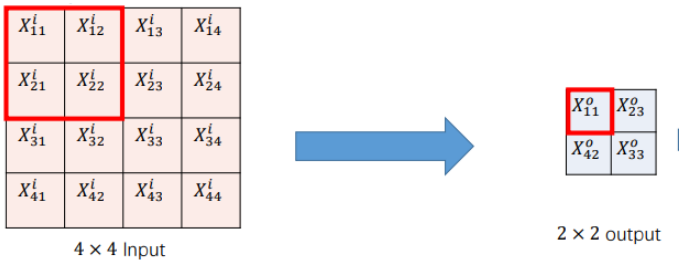

Figure 6: max pooling.

Solution: ■