Machine Learning (DS4023) Assignment 3

Deadline: Nov. 11, 2024.

Problem 1: Hard-margin SVM. (18 pts)

You are given the following two sets of data points, each belonging to one of the two classes (class 1 and class -1):

- Class 1 (labeled as +1):
$$(1, 2), (2, 3)$$
- Class -1 (labeled as -1):
$$(2, 1), (3, 2)$$

Please find the optimal separating hyperplane using a linear SVM and derive the equation of the hyperplane. Assume the hard-margin SVM.

1. Write down the formulation of SVM, including the separation hyperplane, the constraints and the final optimization problem with parameters. **(4 pts)**
2. Write down the Lagrangian form for this problem using the parameters and Lagrange multipliers. Please also write out its dual form. **(10 pts)**
3. Assume that the Lagrangian multipliers $\boldsymbol{\alpha}_i$'s are all 0.5 and that the point $(1, 2)$ is a support vector for ease of calculation. Please calculate the values of weight vector $\boldsymbol{w}$ and bias b. Write out the explicit form of the hyperplane. **(4 pts)**

Solution

1. Formulation of SVM

For a linear SVM, the goal is to find a hyperplane that separates the two classes. The equation of the hyperplane is of the form: **(1 pt)**

$$\boldsymbol{w}^T \boldsymbol{x} + b = 0,$$

where $\boldsymbol{w}$ is the weight vector, $\boldsymbol{x}$ is the input vector, and b is the bias term. The constraints for correctly classified points are: **(1 pt)**

$$y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1, \quad \forall i$$

where y_i is the label of the data point $\boldsymbol{x}_i$. To find the optimal hyperplane, we minimize the following objective function: **(2 pts)**

$$\begin{aligned} \min \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{s.t.} \quad & y_i(w^T\boldsymbol{x}_i + b) \geq 1. \end{aligned}$$

2. Lagrangian form

The Lagrangian for the hard-margin SVM is:

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{4} \boldsymbol{\alpha}_i \left[y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1\right], \quad \textbf{(2 pts)}$$

where $\boldsymbol{\alpha}_i$'s ≥ 0 are the Lagrange multipliers. To minimize the Lagrangian with respect to the primal variables $\boldsymbol{w}$ and b, we take the partial derivative of $L(\boldsymbol{w}, b, \boldsymbol{\alpha})$ with respect to w and set it to zero:

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{4} \alpha_i y_i \boldsymbol{x}_i = 0, \quad \textbf{(1 pt)}$$

which leads to:

$$\boldsymbol{w} = \sum_{i=1}^{4} \boldsymbol{\alpha}_i y_i \boldsymbol{x}_i.$$

We then take the partial derivative of $L(\boldsymbol{w}, b, \boldsymbol{\alpha})$ with respect to b and set it to zero:

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{4} \boldsymbol{\alpha}_i y_i = 0. \quad \textbf{(1 pt)}$$

This gives us the constraints:

$$\sum_{i=1}^{4} \boldsymbol{\alpha}_i y_i = 0$$

Substituting the optimal values of $\boldsymbol{w} = \sum_{i=1}^{4} \boldsymbol{\alpha}_i y_i \boldsymbol{x}_i$ back, we have

$$||\boldsymbol{w}||^2 = \left(\sum_{i=1}^{4} \boldsymbol{\alpha}_i y_i \boldsymbol{x}_i\right)^T \left(\sum_{j=1}^{4} \boldsymbol{\alpha}_j y_j \boldsymbol{x}_j\right) = \sum_{i=1}^{4}\sum_{j=1}^{4} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j y_i y_j (\boldsymbol{x}_i^T \boldsymbol{x}_j) \quad \textbf{(2 pts)}$$

Thus, the Lagrangian becomes: **(2 pts)**

$$L(\alpha) = \sum_{i=1}^{4} \boldsymbol{\alpha}_i - \frac{1}{2}\sum_{i=1}^{4}\sum_{j=1}^{4} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j y_i y_j (\boldsymbol{x}_i^T \boldsymbol{x}_j)$$

The dual problem is to maximize the above expression for $L(\alpha)$, subject to the following constraints:

$$\sum_{i=1}^{4} \boldsymbol{\alpha}_i y_i = 0 \quad \text{and} \quad \boldsymbol{\alpha}_i \geq 0 \quad \forall i.$$

Thus, the dual formulation of the SVM is: **(2 pts)**

$$\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{4} \boldsymbol{\alpha}_i - \frac{1}{2} \sum_{i=1}^{4} \sum_{j=1}^{4} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j y_i y_j (\boldsymbol{x}_i^T \boldsymbol{x}_j)$$

subject to:

$$\sum_{i=1}^{4} \boldsymbol{\alpha}_i y_i = 0 \quad \text{and} \quad \boldsymbol{\alpha}_i \geq 0 \quad \forall i.$$

3. Values of $\boldsymbol{w}$ and b.

Assume the solution given for $\boldsymbol{\alpha}$ is $(0.5, 0.5, 0.5, 0.5)$. The weight vector $\boldsymbol{w}$ is computed as:

$$\boldsymbol{w} = \sum_{i=1}^{4} \boldsymbol{\alpha}_i y_i \boldsymbol{x}_i \quad \textbf{(1 pt)}$$

Substituting the values of α_i, y_i, and x_i:

$$\boldsymbol{w} = 0.5 \times (1, 2) + 0.5 \times (2, 3) + 0.5 \times (-2, -1) + 0.5 \times (-3, -2) \quad \textbf{(1 pt)}$$
$$\boldsymbol{w} = (-1, 1).$$

To compute the bias b, we use one of the support vectors $\boldsymbol{x}_1 = (1, 2)$ with $y_1 = +1$:

$$y_1(\boldsymbol{w}^T \boldsymbol{x}_1 + b) = 1 \Rightarrow 1 \times ((-1 \times 1) + (1 \times 2) + b) = 1 \quad \textbf{(1 pt)}$$
$$\Rightarrow b = 0$$

Thus, the final hyperplane equation is:

$$\boldsymbol{w}^T \boldsymbol{x} + b = 0 \Rightarrow \boldsymbol{x}_1 = \boldsymbol{x}_2 \quad \textbf{(1 pt)}$$

Problem 2: Soft-margin SVM. (20 pts)

Suppose we have the data points $\boldsymbol{x} \in \mathbb{R}^{n \times d}$ with corresponding labels $\boldsymbol{y} \in \mathbb{R}^n$. We want to use a soft-margin SVM to classify these data points with a regularization parameter $C = 1$.

1. Write down the formulation of soft-margin SVM for this problem using $\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{y}, b$ and $\boldsymbol{\xi}$. Write out explicitly their dimensions. **(3 pts)**
2. Write down the Lagrangian form and derive the dual for the problem. Write down the detailed derivation steps. **(12 pts)**
3. Obtain the decision boundary. **(3 pts)**
4. Explain why $\boldsymbol{\xi}$ disappears in the dual. **(2 pts)**

Solution

1. The primal problem

The primal optimization problem for a soft margin SVM is:

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i \quad \textbf{(1 pt)}$$

subject to the constraints:

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i = 1, \ldots, N \quad \textbf{(1 pt)}$$

where:

- $\boldsymbol{w} \in \mathbb{R}^d$ is the weight vector, **(0.5 pt)**
- $b \in \mathbb{R}$ is the bias term, **(0.5 pt)**
- ξ_i are slack variables allowing classification errors,
- C is the regularization parameter.

2. The Lagrangian and the Dual

The Lagrangian for the primal problem is:

$$L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left[y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - (1 - \xi_i)\right] - \sum_{i=1}^{N}\mu_i\xi_i \quad \textbf{(2 pts)}$$

where: **(1 pt)**

- $\alpha_i \geq 0,\ \forall i$ are Lagrange multipliers for the margin constraints,
- $\mu_i \geq 0,\ \forall i$ are Lagrange multipliers for the non-negativity of slack variables.

We minimize the Lagrangian with respect to the primal variables $\mathbf{w}$, b, and ξ_i.

Derivative with respect to $\boldsymbol{w}$:

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{N}\alpha_i y_i \boldsymbol{x}_i = 0 \quad \Rightarrow \quad \boldsymbol{w} = \sum_{i=1}^{N}\alpha_i y_i \boldsymbol{x}_i \quad \textbf{(1 pt)}$$

Derivative with respect to b:

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{N}\alpha_i y_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{N}\alpha_i y_i = 0 \quad \textbf{(1 pt)}$$

Derivative with respect to ξ_i:

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \quad \Rightarrow \quad 0 \leq \alpha_i \leq C \quad \textbf{(1 pt)}$$

Substituting $\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$ into the Lagrangian to eliminate $\boldsymbol{w}$, we get:

$$\frac{1}{2}\|\boldsymbol{w}\|^2 = \frac{1}{2}\left(\sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i\right)^T \left(\sum_{j=1}^{N} \alpha_j y_j \boldsymbol{x}_j\right) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \quad \textbf{(2 pts)}$$

Thus, the Lagrangian simplifies to:

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \quad \textbf{(1 pt)}$$

Dual Problem Formulation:

The dual optimization problem is to maximize the Lagrangian with respect to α_i, subject to the constraints derived earlier:

$$\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \quad \textbf{(2 pts)}$$

subject to:

$$\sum_{i=1}^{N} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad \forall i = 1, \ldots, N \quad \textbf{(1 pt)}$$

3. The Decision Boundary

Once the dual problem is solved, the weight vector $\boldsymbol{w}$ and bias term b can be computed as:

$$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i \quad \textbf{(1 pt)}$$

The bias b is computed from any support vector $\boldsymbol{x}_i$ that lies on the margin:

$$b = y_i - \boldsymbol{w}^T \boldsymbol{x}_i \quad \textbf{(1 pt)}$$

The final decision function for classifying a new point $\boldsymbol{x}$ is:

$$f(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^T \boldsymbol{x} + b) \quad \textbf{(1 pt)}$$

4. The reason that ξ disappears.

In the primal formulation, the slack variables ξ_i are introduced to penalize violations of the margin constraints, allowing the SVM to handle non-linearly separable data by introducing classification errors. However, when we derive the dual problem by minimizing the

Lagrangian with respect to $\boldsymbol{w}$, b, and ξ_i, the slack variables disappear from the final dual problem. The key step occurs when we take the partial derivative of the Lagrangian with respect to ξ_i, yielding:

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \mu_i = 0. \quad \textbf{(1 pt)}$$

This implies that the Lagrange multiplier α_i, which corresponds to the margin constraint violation, is bounded by $\alpha_i \leq C$. Thus, in the dual problem, the effect of the slack variables is implicitly captured by α_i **(1 pt)**, and there is no need for ξ_i to appear explicitly. The dual formulation captures the balance between maximizing the margin and allowing for misclassifications through the constraint $0 \leq \alpha_i \leq C$.

Problem 3: Kernel SVM. (17 pts)

Consider the following 2D dataset with four training points:

$$\mathbf{x}_1 = (1, 2), \quad y_1 = 1$$

$$\mathbf{x}_2 = (2, 3), \quad y_2 = 1$$

$$\mathbf{x}_3 = (3, 1), \quad y_3 = -1$$

$$\mathbf{x}_4 = (4, 3), \quad y_4 = -1$$

We want to use the **polynomial kernel** $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^\top \mathbf{x}_j + 1)^2$ to classify these points with a soft-margin SVM. The regularization parameter $C = 1$.

1. Compute the kernel matrix K. **(6 pts)**
2. Set up the dual optimization problem. You can use the results from Problem 2. **(4 pts)**
3. Suppose the Lagrange multipliers α's are

   $$\alpha_1 = 0.0182, \quad \alpha_2 = 0.0068, \quad \alpha_3 = 0.0250, \quad \alpha_4 = 0,$$

   and $\boldsymbol{x}_3$ is a support vector. Please compute the bias term b. **(2 pts)**
4. Classify a new point $\mathbf{x}_5 = (2, 1)$ using the learned kernel SVM model. **(5 pts)**

Solution

1. Compute the Kernel Matrix

The kernel matrix K is computed using the polynomial kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^\top \mathbf{x}_j + 1)^2 \quad \textbf{(1 pt)}$$

The kernel values are computed as follows: **(4 pts, 0.4pt each)**

$$k(\mathbf{x}_1, \mathbf{x}_1) = (1 \cdot 1 + 2 \cdot 2 + 1)^2 = 6^2 = 36$$

$$\begin{aligned}
k(\mathbf{x}_1, \mathbf{x}_2) &= (1 \cdot 2 + 2 \cdot 3 + 1)^2 = 9^2 = 81 \\
k(\mathbf{x}_1, \mathbf{x}_3) &= (1 \cdot 3 + 2 \cdot 1 + 1)^2 = 6^2 = 36 \\
k(\mathbf{x}_1, \mathbf{x}_4) &= (1 \cdot 4 + 2 \cdot 3 + 1)^2 = 11^2 = 121 \\
k(\mathbf{x}_2, \mathbf{x}_2) &= (2 \cdot 2 + 3 \cdot 3 + 1)^2 = 14^2 = 196 \\
k(\mathbf{x}_2, \mathbf{x}_3) &= (2 \cdot 3 + 3 \cdot 1 + 1)^2 = 10^2 = 100 \\
k(\mathbf{x}_2, \mathbf{x}_4) &= (2 \cdot 4 + 3 \cdot 3 + 1)^2 = 18^2 = 324 \\
k(\mathbf{x}_3, \mathbf{x}_3) &= (3 \cdot 3 + 1 \cdot 1 + 1)^2 = 11^2 = 121 \\
k(\mathbf{x}_3, \mathbf{x}_4) &= (3 \cdot 4 + 1 \cdot 3 + 1)^2 = 16^2 = 256 \\
k(\mathbf{x}_4, \mathbf{x}_4) &= (4 \cdot 4 + 3 \cdot 3 + 1)^2 = 26^2 = 676
\end{aligned}$$

Thus, the kernel matrix K is: **(1 pt)**

$$K = \begin{bmatrix} 36 & 81 & 36 & 121 \\ 81 & 196 & 100 & 324 \\ 36 & 100 & 121 & 256 \\ 121 & 324 & 256 & 676 \end{bmatrix}$$

2. Solve the Dual Optimization Problem

The dual optimization problem is:

$$\max_{\alpha} \sum_{i=1}^{4} \alpha_i - \frac{1}{2} \sum_{i=1}^{4} \sum_{j=1}^{4} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j), \quad \textbf{(2 pts)}$$

subject to the constraints:

$$\sum_{i=1}^{4} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C = 1 \quad \textbf{(2 pts)}$$

3. Compute the Bias Term b

Assume that optimized Lagrange multipliers are

$$\alpha_1 = 0.0182, \quad \alpha_2 = 0.0068, \quad \alpha_3 = 0.0250, \quad \alpha_4 = 0,$$

and $\boldsymbol{x}_3$ is a support vector. The bias term b is computed using $\alpha_3 = 0.0250$. With the kernel values for $k(\mathbf{x}_i, \mathbf{x}_3)$, we have: **(1 pts)**

$$k(\mathbf{x}_1, \mathbf{x}_3) = 36, \quad k(\mathbf{x}_2, \mathbf{x}_3) = 100, \quad k(\mathbf{x}_3, \mathbf{x}_3) = 121, \quad k(\mathbf{x}_4, \mathbf{x}_3) = 256$$

Substituting these values into the equation for b:

$$-1\left(0.0182 \cdot 1 \cdot 36 + 0.0068 \cdot 1 \cdot 100 + 0.0250 \cdot (-1) \cdot 121 + 0 \cdot (-1) \cdot 256 + b\right) = 1$$

$$b = 0.6898 \quad \textbf{(1 pt)}$$

Thus, the bias term is $b = 0.6898$.

4. Classify the New Point $\mathbf{x}_5 = (2, 1)$

We will classify the new point $\mathbf{x}_5 = (2, 1)$ using the decision function:

$$f(\mathbf{x}_5) = \sum_{i=1}^{4} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}_5) + b \quad \textbf{(2 pts)}$$

First, compute the kernel values:

$$k(\mathbf{x}_1, \mathbf{x}_5) = 25, \quad k(\mathbf{x}_2, \mathbf{x}_5) = 64, \quad k(\mathbf{x}_3, \mathbf{x}_5) = 64, \quad k(\mathbf{x}_4, \mathbf{x}_5) = 144 \quad \textbf{(1 pt)}$$

Now, compute the decision function:

$$f(\mathbf{x}_5) = 0.0182 \cdot 1 \cdot 25 + 0.0068 \cdot 1 \cdot 64 + 0.0250 \cdot (-1) \cdot 64 + 0 \cdot (-1) \cdot 144 + 0.6898$$

$$f(\mathbf{x}_5) = 0.455 + 0.4352 - 1.6 + 0.6898 = -0.02 \quad \textbf{(1 pt)}$$

Since $f(\mathbf{x}_5) < 0$, the point $\mathbf{x}_5 = (2, 1)$ is classified as -1. **(1 pt)**

Problem 4: Programming (45 pts)

Complete the jupyter notebook attached on programming for ensemble learning and SVM. Submit the completed file.