Exercise of Supervised Learning: SVM Part 2

Yawei Li

yawei.li@stat.uni-muenchen.de

December 15, 2023

Exercise 1: Kernelized Multiclass SVM

For a data set $\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(n)}, y^{(n)})\}$ with $y^{(i)} \in \mathcal{Y} = \{+1, -1\}$, assume that we are provided with a suitable feature map $\phi : \mathcal{X} \to \Phi$, where $\Phi \subset \mathbb{R}^d$. In the featureized SVM learning problem we are facing the following optimization problem:

$$
\begin{aligned}
\min_{\boldsymbol{\theta}, \theta_0, \zeta^{(i)}} \quad & \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} + C\sum_{i=1}^{n} \zeta^{(i)} \\
\text{s.t.} \quad & y^{(i)}\left(\left\langle \boldsymbol{\theta}, \phi(\mathbf{x}^{(i)}) \right\rangle + \theta_0\right) \geq 1 - \zeta^{(i)} \qquad \forall i \in \{1, \dots, n\}, \\
\text{and} \quad & \zeta^{(i)} \geq 0 \qquad i \in \{1, \dots, n\},
\end{aligned}
$$

where $C \geq 0$ is some constant.
(a) Argue that this is equivalent to the following ERM problem:

$$
\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \frac{1}{2}||\boldsymbol{\theta}||^2 + C\sum_{i=1}^{n} \max(1 - y^{(i)}(\boldsymbol{\theta}^T\phi(\mathbf{x}^{(i)}) + \theta_0), 0).
$$

i.e., the regularized ERM problem for the hinge loss for the hypothesis space

$$
\mathcal{H} = \{f : \Phi \to \mathbb{R} \mid f(\boldsymbol{z}) = \boldsymbol{\theta}^T\boldsymbol{z} + \theta_0, \boldsymbol{\theta} \in \mathbb{R}^d, \theta_0 \in \mathbb{R}\}
$$

Solution to Exercise 1 (a)

1. Identify that: $\langle \boldsymbol{\theta}, \phi(\mathbf{x}^{(i)}) \rangle + \theta_0 = \boldsymbol{\theta}^T \phi(\mathbf{x}^{(i)}) + \theta_0$.
2. Check the conditions $y^{(i)}(\boldsymbol{\theta}^T \phi(\mathbf{x}^{(i)}) + \theta_0) \geq 1 - \zeta^{(i)}$ and $\zeta^{(i)} \geq 0$.
   - Case 1: if $y^{(i)}(\boldsymbol{\theta}^T \phi(\mathbf{x}^{(i)}) + \theta_0) \geq 1$, then $\zeta^{(i)} = 0$. $\triangleright$
   - Case 2: if $y^{(i)}(\boldsymbol{\theta}^T \phi(\mathbf{x}^{(i)}) + \theta_0) < 1$, then $\zeta^{(i)} = 1 - y^{(i)}(\boldsymbol{\theta}^T \phi(\mathbf{x}^{(i)}) + \theta_0) > 0$. $\triangleright$

   Combining both cases, we can write

$$\zeta^{(i)} = \max(0, 1 - y^{(i)}(\boldsymbol{\theta}^T \phi(\mathbf{x}^{(i)}) + \theta_0))$$

   Plug in to the primal problem we can prove that it is equivalent to the $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ using hinge loss.

Exercise 1 (b)

(b) Now assume we deal with a multiclass classification problem with a data set $\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)})\}$ such that $y^{(i)} \in \mathcal{Y} = \{1, \ldots, g\}$ for each $i \in \{1, \ldots, n\}$. In this case, we can derive a similar regularized ERM problem by using the multiclass hinge loss (see Exercse Sheet 4(b)):

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \frac{1}{2}||\boldsymbol{\theta}||^2 + C \sum_{i=1}^{n} \sum_{y \neq y^{(i)}} \max(1 + \boldsymbol{\theta}^T \psi(\mathbf{x}^{(i)}, y) - \boldsymbol{\theta}^T \psi(\mathbf{x}^{(i)}, y^{(i)}), 0),$$

where $\psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ is a suitable (multiclass) feature map. Specify a ψ such that this regularized multiclass ERM problem coincides with the regularized binary ERM problem in (a).

Solution to 1 (b)

1. Consider $\psi(\mathbf{x}, y) = \frac{1}{2} y \tilde{\phi}(\mathbf{x})$, where $\tilde{\phi}(\mathbf{x}) = (1, \phi(\mathbf{x}^{(i)}))^T$, and $\tilde{\boldsymbol{\theta}} = (\theta_0, \boldsymbol{\theta})^T$.
2. Then, for $y \neq y^{(i)}$, it follows that

$$
\begin{aligned}
&1 + \tilde{\boldsymbol{\theta}}^T \psi(\mathbf{x}^{(i)}, y) - \boldsymbol{\theta}^T \psi(\mathbf{x}^{(i)}, y^{(i)}) \\
&= 1 + \frac{1}{2} y \tilde{\boldsymbol{\theta}}^T \tilde{\phi}(\mathbf{x}^{(i)}) - \frac{1}{2} y^{(i)} \tilde{\boldsymbol{\theta}}^T \tilde{\phi}(\mathbf{x}^{(i)}) \\
&= 1 + \frac{1}{2} \left(y - y^{(i)} \right) \tilde{\boldsymbol{\theta}}^T \tilde{\phi}(\mathbf{x}^{(i)}) \\
&= \begin{cases} 1 + \tilde{\boldsymbol{\theta}}^T \tilde{\phi}(\mathbf{x}^{(i)}), & \text{if } y^{(i)} = -1 \\ 1 - \tilde{\boldsymbol{\theta}}^T \tilde{\phi}(\mathbf{x}^{(i)}), & \text{if } y^{(i)} = +1 \end{cases} \\
&= 1 - y^{(i)} \tilde{\boldsymbol{\theta}}^T \tilde{\phi}(\mathbf{x}^{(i)}).
\end{aligned}
$$

Solution to 1 (b): Continued

3. Thus,

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \frac{1}{2}||\boldsymbol{\theta}||^2 + C\sum_{i=1}^{n}\sum_{y\neq y^{(i)}} \max(1 + \tilde{\boldsymbol{\theta}}^T\psi(\mathbf{x}^{(i)}, y) - \tilde{\boldsymbol{\theta}}^T\psi(\mathbf{x}^{(i)}, y^{(i)}), 0)$$

$$= \frac{1}{2}||\boldsymbol{\theta}||^2 + C\sum_{i=1}^{n} \max(1 - y^{(i)}\tilde{\boldsymbol{\theta}}^T\tilde{\phi}(\mathbf{x}^{(i)}), 0)$$

$$= \frac{1}{2}||\boldsymbol{\theta}||^2 + C\sum_{i=1}^{n} \max(1 - y^{(i)}(\boldsymbol{\theta}^T\phi(\mathbf{x}^{(i)}) + \theta_0), 0).$$

Exercise 1 (c)

(c) Show that the regularized multiclass ERM problem in (b) can be written in the kernelized form:

$$\frac{1}{2}\boldsymbol{\beta}^T \boldsymbol{K} \boldsymbol{\beta} + C \sum_{i=1}^{n} \sum_{y \neq y^{(i)}} \max(1 + (\boldsymbol{K}\beta)_{(i-i)g+y} - (\boldsymbol{K}\boldsymbol{\beta})_{(i-1)g+y^{(i)}}), 0),$$

where $\boldsymbol{\beta} \in \mathbb{R}^{ng}$ and $\boldsymbol{K} = \mathbf{X}\mathbf{X}^T$ for $\mathbf{X} \in \mathbb{R}^{ng \times d}$ with row entries $\psi(\mathbf{x}^{(i)}, y)^T$ for $i = i, \ldots, n$, $y = 1, \ldots, g$, i.e.,

$$\mathbf{X} = \begin{pmatrix} \psi(\mathbf{x}^{(1)}, 1)^T \\ \psi(\mathbf{x}^{(1)}, 2)^T \\ \vdots \\ \psi(\mathbf{x}^{(1)}, g)^T \\ \psi(\mathbf{x}^{(2)}, 1)^T \\ \vdots \\ \psi(\mathbf{x}^{(n)}, g)^T \end{pmatrix}.$$

Here, $(\boldsymbol{K}\boldsymbol{\beta})_{(i-1)g+y}$ denotes the $((i-1)g+y)$-th entry of the vector $\boldsymbol{K}\beta$. *Hint:* The representation theorems tells us that for the solution $\boldsymbol{\theta}^*$ of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ it holds that $\boldsymbol{\theta}^* \in \mathrm{span}\{(\psi(\mathbf{x}^{(i)}, y))_{i=1,\ldots,n, y=1,\ldots,g}\}$

Solution to Exercise 1 (c)

$\boldsymbol{\theta}^* \in \mathrm{span}\{(\psi(\mathbf{x}^{(i)}, y))_{i=1,\ldots,n, y=1,\ldots,g}\}$ means that $\boldsymbol{\theta}$ is a linear combination of the spanning bases, i.e. $\boldsymbol{\theta} = \mathbf{X}^T \boldsymbol{\beta}$ for $\boldsymbol{\beta} \in \mathbb{R}^{ng}$ and

$$\mathbf{X} = \begin{pmatrix} \psi(\mathbf{x}^{(1)}, 1)^T \\ \psi(\mathbf{x}^{(1)}, 2)^T \\ \vdots \\ \psi(\mathbf{x}^{(1)}, g)^T \\ \psi(\mathbf{x}^{(2)}, 1)^T \\ \vdots \\ \psi(\mathbf{x}^{(n)}, g)^T \end{pmatrix}.$$

So for $\boldsymbol{K} = \mathbf{X}\mathbf{X}^T$, we obtain

$$||\boldsymbol{\theta}||^2 = \boldsymbol{\theta}^T \boldsymbol{\theta} = (\mathbf{X}^T \boldsymbol{\beta})^T \mathbf{X}^T \boldsymbol{\beta} = \boldsymbol{\beta}^T \boldsymbol{K} \boldsymbol{\beta}$$

Solution to Exercise 1 (c): Continued

Furthermore,

$$\boldsymbol{\theta}^T\psi(\mathbf{x}^{(i)}, y) - \boldsymbol{\theta}^T\psi(\mathbf{x}^{(i)}, y^{(i)}) = \boldsymbol{\beta}^T\mathbf{X}\psi(\mathbf{x}^{(i)}, y) - \beta^T\mathbf{X}\psi(\mathbf{x}^{(i)}, y^{(i)})$$

Note that the result is a scalar.

- Recall that $\boldsymbol{K} = \mathbf{X}\mathbf{X}^T$.
- $\psi(\mathbf{x}^{(i)}, y)$ is the $((i-1)g + y)$-th row of $\mathbf{X}$. (Similar argument for $\psi(\mathbf{x}^{(i)}, y^{(i)})$)
- So, $\mathbf{X}\psi(\mathbf{x}^{(i)}, y)$ is the $((i-1)g + y)$-th row/column of $\boldsymbol{K} = \mathbf{X}\mathbf{X}^T$ (symmetric).
- So, the inner product $\boldsymbol{\beta}^T(\mathbf{X}\psi(\mathbf{x}^{(i)}, y))$ is equivalent to: first compute $\boldsymbol{K}\boldsymbol{\beta}$, and then retrieve the entry in the $((i-1)g + y)$-th row.

Therefore,

$$\begin{aligned}\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) &= \frac{1}{2}||\boldsymbol{\theta}||^2 + C\sum_{i=1}^{n}\sum_{y\neq y^{(i)}} \max(1 + \boldsymbol{\theta}^T\psi(\mathbf{x}^{(i)}, y) - \boldsymbol{\theta}^T\psi(\mathbf{x}^{(i)}, y^{(i)}), 0) \\ &= \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{K}\boldsymbol{\beta} + \sum_{i=1}^{n}\sum_{y\neq y^{(i)}} \max(1 + (\boldsymbol{K}\boldsymbol{\beta})_{(i-1)g+y} - (\boldsymbol{K}\boldsymbol{\beta})_{(i-1)g+y^{(i)}}), 0)\end{aligned}$$

Exercise 2: Kernel Trick

The polynomial kernel is defined as

$$k(x, \tilde{x}) = (x^T \tilde{x} + b)^d.$$

Furthermore, assume that $x \in \mathbb{R}^2$ and $d = 2$. (a) Derive the explicit feature map ϕ taking into account that the following equation holds:

$$k(x, \tilde{x}) = \langle \phi(x), \phi(\tilde{x}) \rangle$$

Solution to 2 (a)

$$
\begin{aligned}
k(x, \tilde{x}) &= \left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix} + b \right)^2 \\
&= (x_1 \tilde{x}_1 + x_2 \tilde{x}_2 + b)^2 \\
&= x_1^2 \tilde{x}_1^2 + 2 x_1 \tilde{x}_1 x_2 \tilde{x}_2 + x_2^2 \tilde{x}_2^2 + 2 b x_1 \tilde{x}_1 + 2 b x_2 \tilde{x}_2 + b^2 \\
&= \left\langle \begin{pmatrix} x_1^2 \\ \sqrt{2} x_1 x_2 \\ x_2^2 \\ \sqrt{2b} x_1 \\ \sqrt{2b} x_2 \\ b \end{pmatrix}, \begin{pmatrix} \tilde{x}_1^2 \\ \sqrt{2} \tilde{x}_1 \tilde{x}_2 \\ \tilde{x}_2^2 \\ \sqrt{2b} \tilde{x}_1 \\ \sqrt{2b} \tilde{x}_2 \\ b \end{pmatrix} \right\rangle \\
&= \langle \phi(x), \phi(\tilde{x}) \rangle
\end{aligned}
$$

Exercise 2 (b)

(b) Descrie the main differences between the kernel method and the explicit feature map.

Solution to Exercise 2 (b)

Using the kernel method reduces the compuational costs of computing the scalar product in the higher-dimensional features space after calculating the feature map.