Support Vector Machine - Corrections

Mai Quyen PHAM - ITI, IMT Atalantique

Exercise I Show that for the kernel $K_i(x, x')$ with $i \in \mathbb{N}_+$, holds:

1. **Scaling**: $cK_1(x, x')$ is a kernel with $c \in \mathbb{R}_+$
2. **Sum**: $K_1(x, x') + K_2(x, x')$ is a kernel
3. **Linear combination**: $\sum_{i=1}^{m} w_i K_i(x, x')$ is a kernel with $w \in \mathbb{R}_+^m$
4. **Product**: $K_1(x, x')K_2(x, x')$ is a kernel
5. **Power**: $[K_1(x, x')]^p$ is a kernel with $p \in \mathbb{N}_+$

Answer:

1. **Scaling**

$$\begin{aligned} cK_1(x, x') &= c\phi_1(x)^\top \phi_1(x') \\ &= \sqrt{c}\phi_1(x)^\top \sqrt{c}\phi_1(x') \\ &= \varphi(x)^\top \varphi(x') \end{aligned}$$

   where $\varphi(x) = \sqrt{c}\phi_1(x)$

2. **Sum**

$$\begin{aligned} K_1(x, x') + K_2(x, x') &= \phi_1(x)^\top \phi_1(x') + \phi_2(x)^\top \phi_2(x') \\ &= \varphi(x)^\top \varphi(x') \end{aligned}$$

   where $\varphi(x) = \begin{bmatrix} \phi_1(x) \\ \phi_2(x) \end{bmatrix}$

3. **Linear combination**

$$\begin{aligned} \sum_{i=1}^{m} w_i K_i(x, x') &= \sum_{i=1}^{m} \sqrt{w_i}\phi_i(x)^\top \sqrt{w_i}\phi_i(x') \\ &= \varphi(x)^\top \varphi(x') \end{aligned}$$

   where $\varphi(x) = \begin{bmatrix} \phi_1(x) \\ \phi_2(x) \\ \vdots \\ \phi_n(x) \end{bmatrix}$

4. **Product**

$$\begin{aligned} K_1(x, x')K_2(x, x') &= \phi_1(x)^\top \phi_1(x')\phi_2(x)^\top \phi_2(x') \\ &= \left(\sum_{i=1}^{m_1} \phi_1(x)_i \phi_1(x')_i \right) \left(\sum_{j=1}^{m_2} \phi_2(x)_j \phi_2(x')_j \right) \\ &= \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \phi_1(x)_i \phi_2(x)_j \phi_1(x')_i \phi_2(x')_j \\ &= \varphi(x)^\top \varphi(x') \end{aligned}$$

where $\varphi(x) = \begin{bmatrix} \phi_1(x)_1\phi_2(x)_1 \\ \vdots \\ \phi_1(x)_1\phi_2(x)_{m_2} \\ \vdots \\ \phi_1(x)_{m_1}\phi_2(x)_1 \\ \vdots \\ \phi_1(x)_{m_1}\phi_2(x)_{m_2} \end{bmatrix}$

5. **Power**: proof by induction i.e.
   - from 4. Product we have $K_1(x, x')^2$ is a kernel
   - $K_1(x, x')^p = K_1(x, x')^{p-1}K_1(x, x')$ is a kernel

Exercise II In contrast to ordinary least squares which has a cost function, called *ridge regression*

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(\theta^\top x^{(i)} - y^{(i)})^2$$

we can also add a term that penalizes large weights in θ as following

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(\theta^\top x^{(i)} - y^{(i)})^2 + \frac{\lambda}{2}\|\theta\|^2$$

where $\lambda > 0$ is a fixed constant, called regularization parameter.

1. Find a closed-form expression for the value of θ which minimizes the ridge regression cost function.
2. Suppose that we want to use kernels to implicitly represent our feature vectors in a high-dimensional space. Using a feature mapping ϕ, the ridge regression cost function becomes

   $$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(\theta^\top \phi(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2}\|\theta\|^2$$

   Making a prediction on a new input x_{new} would now be done by computing $\theta^\top\phi(x_{new})$. Show how we can use the kernel strick to obtain a closed form for the prediction one the new input without ever explicitly computing $\phi(x_{new})$. You may assume that the parameter vector θ can be expressed as a linear combination of the input feature vectors i.e. $\theta = \sum_{i=1}^{m}\alpha_i\phi(x^{(i)})$ for some set of parameters α_i.

Answer:

1. Let $X = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(m)} \end{bmatrix} \in \mathbb{R}^{n\times m}$ be a matrix of with the rows $x^{(i)} \in \mathbb{R}^n$ and the label vector $y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix} \in \mathbb{R}^m$.
   The ridge regression cost function $J(\theta)$ can rewrite as

   $$J(\theta) = \frac{1}{2}\|X\theta - y\|^2 + \frac{\lambda}{2}\|\theta\|^2$$

   Then the gradient is

   $$\nabla J(\theta) = X^\top X\theta + \lambda\theta$$

   setting the gradient to 0 gives us

   $$\theta = (X^\top X + \lambda I)^{-1}X^\top y$$

2. Let Φ be the design matrix associated with the feature vectors $\phi(x^{(i)})$. Then from part (1) we have,

   $$\begin{aligned}\theta &= (\Phi^\top\Phi + \lambda I)^{-1}\Phi^\top y \\ &= \Phi^\top(\Phi\Phi^\top + \lambda I)^{-1}y \\ &= \Phi^\top(K + \lambda I)^{-1}y\end{aligned}$$

where K is the kernel matrix. To predict y_{new} we compute

$$\begin{aligned} y_{new} &= \theta^\top \phi(x_n ew) \\ &= y^\top (K + \lambda I)^{-1} \Phi \phi(x_{new}) \\ &= \sum_{i=1}^{m} \alpha_i K(x^{(i)}, x_{new}) \end{aligned}$$

where $\alpha = (K + \lambda I)^{-1} y$

Exercise III In class, we saw that if our data is not linearly separable, then we need to modify our support vector machine algorithm by introducing an error margin that must be minimized. Specifically, the formulation we have looked at is known as the ℓ_1 norm soft margin SVM. In this problem we will consider an alternative method, known as the ℓ_2 norm soft margin SVM. This new algorithm is given by the following optimization problem (notice that the slack penalties are now squared):

$$\begin{aligned} \min_{w \in \mathbb{R}, b \in \mathbb{R}, \xi} \quad & \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{n} \xi_i^2 \\ s.t. \quad & y_i(w^\top x_i + b) \geq 1 - \xi_i \ \forall i = 1, \ldots, n \end{aligned}$$

1. Notice that we have dropped the $\xi_i \geq 0$ constraint in the ℓ_2 problem. Show that these non-negativity constraints can be removed. That is, show that the optimal value of the objective will be the same whether or not these constraints are present.
2. What is the Lagrangian of the ℓ_2 soft margin SVM optimization problem?
3. Minimize the Lagrangian with respect to w, b, and ξ.
4. What is the dual of the ℓ_2 soft margin SVM optimization problem?

Answer:

1. Suppose that ξ^* is the optimal solution of the problem with $\xi_i^* < 0$. Let define a new vector $\bar{\xi} = [\xi_1^*, \ldots, \xi_{i-1}^*, 0, \xi_{i+1}^*, \ldots, \xi_n^*]^\top$. Then $\bar{\xi}$ satisfies the constraint of the problem, and the objective function would be lower. Therefore ξ^* with negative value could not be an optimal solution of the problem.
2. The Lagrangian of ℓ_2 soft margin SVM optimization problem is

$$L(w, b, \xi, \alpha) = \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{m} \xi_i^2 - \sum_{i=1}^{m} \alpha_i [y^{(i)}(w^\top x^{(i)} + b) - 1 + \xi_i]$$

   where $\alpha_i \geq 0$ for $i = 1, \ldots, m$.
3. Taking the gradient with respect to w, we get

$$w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0$$

   which gives us $w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$
   Taking the gradient with respect to b, we get

$$-\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

   Finally, taking the gradient with respect to ξ, we get

$$\alpha = C\xi.$$

4. The objective function for the dual problem is

$$\begin{aligned}
\ell(\alpha) &= \min_{w,b,\xi} \; L(w,b,\xi,\alpha) \\
&= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\alpha_i y^{(i)} x^{(i)}\right)^\top \alpha_j y^{(j)} x^{(j)} + \frac{1}{2}\sum_{j=1}^{m}\frac{\alpha_i}{\xi_i}\xi_i^2 - \sum_{i=1}^{m}\alpha_i\left[y^{(i)}\left(\sum_{j=1}^{m}\alpha_j y^{(j)} x^{(j)\top} x^{(i)} + b\right) - 1 + \xi_i\right] \\
&= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^{(i)} y^{(j)} x^{(i)\top} x^{(j)} - \frac{1}{2}\alpha^\top\xi + \sum_{i=1}^{m}\alpha_i \\
&= -\frac{1}{2C}\|\alpha\|^2 - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^{(i)} y^{(j)} x^{(i)\top} x^{(j)} + \sum_{i=1}^{m}\alpha_i
\end{aligned}$$

Then the dual formulation of this problem is

$$\begin{aligned}
&\max_{\alpha} \; \ell(\alpha) \\
&\text{s.t. } \alpha_i \geq 0, \; \forall i = 1,\ldots,m \\
&\alpha^\top y = 0
\end{aligned}$$