

Exercise of Supervised Learning: SVM Part 2

Yawei Li

`yawei.li@stat.uni-muenchen.de`

December 23, 2024

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 $\phi : \mathcal{X} \rightarrow \Phi$, where $\Phi \subset \mathbb{R}^d$. In the featureized SVM learning problem we are facing the following optimization problem:

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

where $C \geq 0$ is some constant.

(a) Argue that this is equivalent to the following ERM problem:

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

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

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

1(a): Rewriting the Optimization Target

Optimization target:

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

1(a): Comparison between Optimization Target and \mathcal{R}_{emp}

Optimization target:

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

Empirical risk:

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

Observation: Both contain $\frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta}$ and $1 - y^{(i)}(\boldsymbol{\theta}^\top \phi(\mathbf{x}^{(i)}) + \theta_0)$. Both contain $C \sum_{i=1}^n \dots$ penalty terms.

Next: Prove that $\zeta^{(i)}$ equals $\max(\dots)$ term.

1(a): Comparison between Optimization Target and \mathcal{R}_{emp}

Optimization target:

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

Empirical risk:

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

Observation: Both contain $\frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta}$ and $1 - y^{(i)}(\boldsymbol{\theta}^\top \phi(\mathbf{x}^{(i)}) + \theta_0)$. Both contain $C \sum_{i=1}^n \dots$ penalty terms.

Next: Prove that $\zeta^{(i)}$ equals $\max(\dots)$ term.

1(a): Comparison between Optimization Target and \mathcal{R}_{emp}

Optimization target:

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

Empirical risk:

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

Observation: Both contain $\frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta}$ and $1 - y^{(i)}(\boldsymbol{\theta}^\top \phi(\mathbf{x}^{(i)}) + \theta_0)$. Both contain $C \sum_{i=1}^n \dots$ penalty terms.

Next: Prove that $\zeta^{(i)}$ equals $\max(\dots)$ term.

1 (a): Prove $\zeta^{(i)}$ Equals $\max(\dots)$ Term

For each i , The constraints in the optimization problem:

$$\zeta^{(i)} \geq \underbrace{1 - y^{(i)} (\langle \boldsymbol{\theta}, \phi(\mathbf{x}^{(i)}) \rangle + \theta_0)}_{(i)}$$

$$\zeta^{(i)} \geq \underbrace{0}_{(ii)}$$

$\zeta^{(i)} \geq (i) \text{ or } (ii) \Rightarrow \zeta^{(i)} \geq \text{the larger term in (i) and (ii)} \Rightarrow \zeta^{(i)} \geq \max((i), (ii)).$

Therefore, the constraints translate to $\zeta^{(i)} \geq \max(1 - y^{(i)} (\boldsymbol{\theta}^\top \phi(\mathbf{x}^{(i)}) + \theta_0), 0)$

Note: Our target is to $\min_{\boldsymbol{\theta}, \theta_0, \zeta^{(i)}} \frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} + C \sum_{i=1}^n \zeta^{(i)}$, so smaller $\zeta^{(i)}$ are preferred. \rightsquigarrow Choose $\zeta^{(i)} = \max(\dots, 0)$

1 (a): Prove $\zeta^{(i)}$ Equals $\max(\dots)$ Term

For each i , The constraints in the optimization problem:

$$\zeta^{(i)} \geq \underbrace{1 - y^{(i)} (\langle \boldsymbol{\theta}, \phi(\mathbf{x}^{(i)}) \rangle + \theta_0)}_{(i)}$$

$$\zeta^{(i)} \geq \underbrace{0}_{(ii)}$$

$\zeta^{(i)} \geq (i) \text{ or } (ii) \Rightarrow \zeta^{(i)} \geq \text{the larger term in (i) and (ii)} \Rightarrow \zeta^{(i)} \geq \max((i), (ii)).$

Therefore, the constraints translate to $\zeta^{(i)} \geq \max(1 - y^{(i)} (\boldsymbol{\theta}^\top \phi(\mathbf{x}^{(i)}) + \theta_0), 0)$

Note: Our target is to $\min_{\boldsymbol{\theta}, \theta_0, \zeta^{(i)}} \frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} + C \sum_{i=1}^n \zeta^{(i)}$, so smaller $\zeta^{(i)}$ are preferred. \rightsquigarrow Choose $\zeta^{(i)} = \max(\dots, 0)$

1 (a): Prove $\zeta^{(i)}$ Equals $\max(\dots)$ Term

For each i , The constraints in the optimization problem:

$$\zeta^{(i)} \geq \underbrace{1 - y^{(i)} (\langle \boldsymbol{\theta}, \phi(\mathbf{x}^{(i)}) \rangle + \theta_0)}_{(i)}$$

$$\zeta^{(i)} \geq \underbrace{0}_{(ii)}$$

$\zeta^{(i)} \geq (i) \text{ or } (ii) \Rightarrow \zeta^{(i)} \geq \text{the larger term in (i) and (ii)} \Rightarrow \zeta^{(i)} \geq \max((i), (ii)).$

Therefore, the constraints translate to $\zeta^{(i)} \geq \max(1 - y^{(i)} (\boldsymbol{\theta}^\top \phi(\mathbf{x}^{(i)}) + \theta_0), 0)$

Note: Our target is to $\min_{\boldsymbol{\theta}, \theta_0, \zeta^{(i)}} \frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} + C \sum_{i=1}^n \zeta^{(i)}$, so smaller $\zeta^{(i)}$ are preferred. \rightsquigarrow Choose $\zeta^{(i)} = \max(\dots, 0)$

1 (a): Choose $\zeta^{(i)} = \max(\dots, 0)$

The optimization target

$$\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\theta} + c \sum_{i=1}^n \zeta^{(i)}$$

becomes

$$\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\theta} + c \sum_{i=1}^n \max \left(1 - y^{(i)} \left(\boldsymbol{\theta}^\top \phi(\mathbf{x}^{(i)}) + \theta_0 \right), 0 \right)$$

which is exactly \mathcal{R}_{emp} .



Exercise 1 (b)

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

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

where $\tilde{\boldsymbol{\theta}} := (\theta_0, \boldsymbol{\theta}^\top)^\top \in \mathbb{R}^{d+1}$, and $\psi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^{d+1}$ 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). **P.S.: Red colored text means the places different from the exercise. May be updated in the next version of the exercise.**

1(b): A Closer Look Into the Multiclass Hinge Loss

Goal: Prove that Multiclass Hinge Loss resolves to the Binary Hinge Loss (a) in the binary case.

► **Class label encoding:** Binary: $y^{(i)} \in \{-1, +1\}$. Multiclass: $y^{(i)} \in \{1, \dots, g\}$.

► **Penalty:**

► Binary: $C \sum_{i=1}^n \max(1 - y^{(i)} (\theta^\top \phi(\mathbf{x}^{(i)}) + \theta_0), 0)$.

► Multiclass: $C \sum_{i=1}^n \sum_{y \neq y^{(i)}} \max(1 + \theta^\top \psi(\mathbf{x}^{(i)}, y) - \theta^\top \psi(\mathbf{x}^{(i)}, y^{(i)}), 0)$.

► Align class label encoding: $y^{(i)} \in \{1, 2\} \rightsquigarrow y^{(i)} \in \{-1, 1\}$.

► $\sum_{y \neq y^{(i)}}$ means: $y^{(i)} = +1, y = -1$ or $y^{(i)} = -1, y = +1$.

► **Note** $\psi(\mathbf{x}^{(i)}, y^{(i)})$ takes both $\mathbf{x}^{(i)}$ and $y^{(i)}$ as arguments, while $\phi(\mathbf{x}^{(i)})$ only operates on $\mathbf{x}^{(i)}$.

► There is no θ_0 in Multiclass Hinge Loss. **How to deal with θ_0 ?**

1(b): A Closer Look Into the Multiclass Hinge Loss

Goal: Prove that Multiclass Hinge Loss resolves to the Binary Hinge Loss (a) in the binary case.

► **Class label encoding:** Binary: $y^{(i)} \in \{-1, +1\}$. Multiclass: $y^{(i)} \in \{1, \dots, g\}$.

► **Penalty:**

► Binary: $C \sum_{i=1}^n \max(1 - y^{(i)} (\boldsymbol{\theta}^\top \phi(\mathbf{x}^{(i)}) + \theta_0), 0)$.

► Multiclass: $C \sum_{i=1}^n \sum_{y \neq y^{(i)}} \max(1 + \boldsymbol{\theta}^\top \psi(\mathbf{x}^{(i)}, y) - \boldsymbol{\theta}^\top \psi(\mathbf{x}^{(i)}, y^{(i)}), 0)$.

► Align class label encoding: $y^{(i)} \in \{1, 2\} \rightsquigarrow y^{(i)} \in \{-1, 1\}$.

► $\sum_{y \neq y^{(i)}}$ means: $y^{(i)} = +1, y = -1$ or $y^{(i)} = -1, y = +1$.

► **Note** $\psi(\mathbf{x}^{(i)}, y^{(i)})$ takes both $\mathbf{x}^{(i)}$ and $y^{(i)}$ as arguments, while $\phi(\mathbf{x}^{(i)})$ only operates on $\mathbf{x}^{(i)}$.

► There is no θ_0 in Multiclass Hinge Loss. **How to deal with θ_0 ?**

1(b): A Closer Look Into the Multiclass Hinge Loss

Goal: Prove that Multiclass Hinge Loss resolves to the Binary Hinge Loss (a) in the binary case.

- ▶ **Class label encoding:** Binary: $y^{(i)} \in \{-1, +1\}$. Multiclass: $y^{(i)} \in \{1, \dots, g\}$.
- ▶ **Penalty:**
 - ▶ Binary: $C \sum_{i=1}^n \max(1 - y^{(i)} (\boldsymbol{\theta}^\top \phi(\mathbf{x}^{(i)}) + \theta_0), 0)$.
 - ▶ Multiclass: $C \sum_{i=1}^n \sum_{y \neq y^{(i)}} \max(1 + \boldsymbol{\theta}^\top \psi(\mathbf{x}^{(i)}, y) - \boldsymbol{\theta}^\top \psi(\mathbf{x}^{(i)}, y^{(i)}), 0)$.
- ▶ Align class label encoding: $y^{(i)} \in \{1, 2\} \rightsquigarrow y^{(i)} \in \{-1, 1\}$.
- ▶ $\sum_{y \neq y^{(i)}}$ means: $y^{(i)} = +1, y = -1$ or $y^{(i)} = -1, y = +1$.
- ▶ **Note** $\psi(\mathbf{x}^{(i)}, y^{(i)})$ takes both $\mathbf{x}^{(i)}$ and $y^{(i)}$ as arguments, while $\phi(\mathbf{x}^{(i)})$ only operates on $\mathbf{x}^{(i)}$.
- ▶ There is no θ_0 in Multiclass Hinge Loss. **How to deal with θ_0 ?**

1(b): A Closer Look Into the Multiclass Hinge Loss

Goal: Prove that Multiclass Hinge Loss resolves to the Binary Hinge Loss (a) in the binary case.

- ▶ **Class label encoding:** Binary: $y^{(i)} \in \{-1, +1\}$. Multiclass: $y^{(i)} \in \{1, \dots, g\}$.
- ▶ **Penalty:**
 - ▶ Binary: $C \sum_{i=1}^n \max(1 - y^{(i)} (\boldsymbol{\theta}^\top \phi(\mathbf{x}^{(i)}) + \theta_0), 0)$.
 - ▶ Multiclass: $C \sum_{i=1}^n \sum_{y \neq y^{(i)}} \max(1 + \boldsymbol{\theta}^\top \psi(\mathbf{x}^{(i)}, y) - \boldsymbol{\theta}^\top \psi(\mathbf{x}^{(i)}, y^{(i)}), 0)$.
- ▶ Align class label encoding: $y^{(i)} \in \{1, 2\} \rightsquigarrow y^{(i)} \in \{-1, 1\}$.
- ▶ $\sum_{y \neq y^{(i)}}$ means: $y^{(i)} = +1, y = -1$ or $y^{(i)} = -1, y = +1$.
- ▶ **Note** $\psi(\mathbf{x}^{(i)}, y^{(i)})$ takes both $\mathbf{x}^{(i)}$ and $y^{(i)}$ as arguments, while $\phi(\mathbf{x}^{(i)})$ only operates on $\mathbf{x}^{(i)}$.
- ▶ There is no θ_0 in Multiclass Hinge Loss. **How to deal with θ_0 ?**

1(b): A Closer Look Into the Multiclass Hinge Loss

Goal: Prove that Multiclass Hinge Loss resolves to the Binary Hinge Loss (a) in the binary case.

- ▶ **Class label encoding:** Binary: $y^{(i)} \in \{-1, +1\}$. Multiclass: $y^{(i)} \in \{1, \dots, g\}$.
- ▶ **Penalty:**
 - ▶ Binary: $C \sum_{i=1}^n \max(1 - y^{(i)} (\boldsymbol{\theta}^\top \phi(\mathbf{x}^{(i)}) + \theta_0), 0)$.
 - ▶ Multiclass: $C \sum_{i=1}^n \sum_{y \neq y^{(i)}} \max(1 + \boldsymbol{\theta}^\top \psi(\mathbf{x}^{(i)}, y) - \boldsymbol{\theta}^\top \psi(\mathbf{x}^{(i)}, y^{(i)}), 0)$.
- ▶ Align class label encoding: $y^{(i)} \in \{1, 2\} \rightsquigarrow y^{(i)} \in \{-1, 1\}$.
- ▶ $\sum_{y \neq y^{(i)}}$ means: $y^{(i)} = +1, y = -1$ or $y^{(i)} = -1, y = +1$.
- ▶ **Note** $\psi(\mathbf{x}^{(i)}, y^{(i)})$ takes both $\mathbf{x}^{(i)}$ and $y^{(i)}$ as arguments, while $\phi(\mathbf{x}^{(i)})$ only operates on $\mathbf{x}^{(i)}$.
- ▶ There is no θ_0 in Multiclass Hinge Loss. How to deal with θ_0 ?

1(b): A Closer Look Into the Multiclass Hinge Loss

Goal: Prove that Multiclass Hinge Loss resolves to the Binary Hinge Loss (a) in the binary case.

- ▶ **Class label encoding:** Binary: $y^{(i)} \in \{-1, +1\}$. Multiclass: $y^{(i)} \in \{1, \dots, g\}$.
- ▶ **Penalty:**
 - ▶ Binary: $C \sum_{i=1}^n \max(1 - y^{(i)} (\boldsymbol{\theta}^\top \phi(\mathbf{x}^{(i)}) + \theta_0), 0)$.
 - ▶ Multiclass: $C \sum_{i=1}^n \sum_{y \neq y^{(i)}} \max(1 + \boldsymbol{\theta}^\top \psi(\mathbf{x}^{(i)}, y) - \boldsymbol{\theta}^\top \psi(\mathbf{x}^{(i)}, y^{(i)}), 0)$.
- ▶ Align class label encoding: $y^{(i)} \in \{1, 2\} \rightsquigarrow y^{(i)} \in \{-1, 1\}$.
- ▶ $\sum_{y \neq y^{(i)}}$ means: $y^{(i)} = +1, y = -1$ or $y^{(i)} = -1, y = +1$.
- ▶ **Note** $\psi(\mathbf{x}^{(i)}, y^{(i)})$ takes both $\mathbf{x}^{(i)}$ and $y^{(i)}$ as arguments, while $\phi(\mathbf{x}^{(i)})$ only operates on $\mathbf{x}^{(i)}$.
- ▶ There is no θ_0 in Multiclass Hinge Loss. **How to deal with θ_0 ?**

1(b): Define $\psi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^{d+1}$

1. Motivation: functional margin has form $y(\langle \phi(\mathbf{x}), \boldsymbol{\theta} \rangle + \theta_0)$. We need it into an inner product $\langle \cdot, \cdot \rangle$.
2. We need to merge θ_0 into the inner product. We can add a dummy feature 1 to $\phi(\mathbf{x})$, as $\langle \phi(\mathbf{x}), \boldsymbol{\theta} \rangle + \theta_0 = \langle (1, \phi(\mathbf{x}))^\top, (\theta_0, \boldsymbol{\theta})^\top \rangle$. Define $\tilde{\phi}(\mathbf{x}) = (1, \phi(\mathbf{x}))^\top$, and $\tilde{\boldsymbol{\theta}} = (\theta_0, \boldsymbol{\theta})^\top$.
3. We can merge the coefficient y into $\tilde{\phi}(\mathbf{x})$, obtaining $y\tilde{\phi}(\mathbf{x})$.
4. We have transformed $y(\langle \phi(\mathbf{x}), \boldsymbol{\theta} \rangle + \theta_0)$ into inner product $\langle y\tilde{\phi}(\mathbf{x}), \tilde{\boldsymbol{\theta}} \rangle$.
5. Multiply with a magic number $\frac{1}{2}$. Consider $\psi(\mathbf{x}, y) = \frac{1}{2}y\tilde{\phi}(\mathbf{x})$.

Our next target: Prove that in the binary case:

$$\sum_{y \neq y^{(i)}} \max(1 + \tilde{\boldsymbol{\theta}}^\top \psi(\mathbf{x}^{(i)}, y) - \tilde{\boldsymbol{\theta}}^\top \psi(\mathbf{x}^{(i)}, y^{(i)}), 0)$$

is equivalent to

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

1(b): Define $\psi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^{d+1}$

1. Motivation: functional margin has form $y(\langle \phi(\mathbf{x}), \boldsymbol{\theta} \rangle + \theta_0)$. We need it into an inner product $\langle \cdot, \cdot \rangle$.
2. We need to merge θ_0 into the inner product. We can add a dummy feature 1 to $\phi(\mathbf{x})$, as $\langle \phi(\mathbf{x}), \boldsymbol{\theta} \rangle + \theta_0 = \langle (1, \phi(\mathbf{x}))^\top, (\theta_0, \boldsymbol{\theta})^\top \rangle$. Define $\tilde{\phi}(\mathbf{x}) = (1, \phi(\mathbf{x}))^\top$, and $\tilde{\boldsymbol{\theta}} = (\theta_0, \boldsymbol{\theta})^\top$.
3. We can merge the coefficient y into $\tilde{\phi}(\mathbf{x})$, obtaining $y\tilde{\phi}(\mathbf{x})$.
4. We have transformed $y(\langle \phi(\mathbf{x}), \boldsymbol{\theta} \rangle + \theta_0)$ into inner product $\langle y\tilde{\phi}(\mathbf{x}), \tilde{\boldsymbol{\theta}} \rangle$.
5. Multiply with a magic number $\frac{1}{2}$. Consider $\psi(\mathbf{x}, y) = \frac{1}{2}y\tilde{\phi}(\mathbf{x})$.

Our next target: Prove that in the binary case:

$$\sum_{y \neq y^{(i)}} \max(1 + \tilde{\boldsymbol{\theta}}^\top \psi(\mathbf{x}^{(i)}, y) - \tilde{\boldsymbol{\theta}}^\top \psi(\mathbf{x}^{(i)}, y^{(i)}), 0)$$

is equivalent to

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

1(b): Define $\psi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^{d+1}$

1. Motivation: functional margin has form $y(\langle \phi(\mathbf{x}), \boldsymbol{\theta} \rangle + \theta_0)$. We need it into an inner product $\langle \cdot, \cdot \rangle$.
2. We need to merge θ_0 into the inner product. We can add a dummy feature 1 to $\phi(\mathbf{x})$, as $\langle \phi(\mathbf{x}), \boldsymbol{\theta} \rangle + \theta_0 = \langle (1, \phi(\mathbf{x}))^\top, (\theta_0, \boldsymbol{\theta})^\top \rangle$. Define $\tilde{\phi}(\mathbf{x}) = (1, \phi(\mathbf{x}))^\top$, and $\tilde{\boldsymbol{\theta}} = (\theta_0, \boldsymbol{\theta})^\top$.
3. We can merge the coefficient y into $\tilde{\phi}(\mathbf{x})$, obtaining $y\tilde{\phi}(\mathbf{x})$.
4. We have transformed $y(\langle \phi(\mathbf{x}), \boldsymbol{\theta} \rangle + \theta_0)$ into inner product $\langle y\tilde{\phi}(\mathbf{x}), \tilde{\boldsymbol{\theta}} \rangle$.
5. Multiply with a magic number $\frac{1}{2}$. Consider $\psi(\mathbf{x}, y) = \frac{1}{2}y\tilde{\phi}(\mathbf{x})$.

Our next target: Prove that in the binary case:

$$\sum_{y \neq y^{(i)}} \max(1 + \tilde{\boldsymbol{\theta}}^\top \psi(\mathbf{x}^{(i)}, y) - \tilde{\boldsymbol{\theta}}^\top \psi(\mathbf{x}^{(i)}, y^{(i)}), 0)$$

is equivalent to

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

1(b): Define $\psi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^{d+1}$

1. Motivation: functional margin has form $y(\langle \phi(\mathbf{x}), \boldsymbol{\theta} \rangle + \theta_0)$. We need it into an inner product $\langle \cdot, \cdot \rangle$.
2. We need to merge θ_0 into the inner product. We can add a dummy feature 1 to $\phi(\mathbf{x})$, as $\langle \phi(\mathbf{x}), \boldsymbol{\theta} \rangle + \theta_0 = \langle (1, \phi(\mathbf{x}))^\top, (\theta_0, \boldsymbol{\theta})^\top \rangle$. Define $\tilde{\phi}(\mathbf{x}) = (1, \phi(\mathbf{x}))^\top$, and $\tilde{\boldsymbol{\theta}} = (\theta_0, \boldsymbol{\theta})^\top$.
3. We can merge the coefficient y into $\tilde{\phi}(\mathbf{x})$, obtaining $y\tilde{\phi}(\mathbf{x})$.
4. We have transformed $y(\langle \phi(\mathbf{x}), \boldsymbol{\theta} \rangle + \theta_0)$ into inner product $\langle y\tilde{\phi}(\mathbf{x}), \tilde{\boldsymbol{\theta}} \rangle$.
5. Multiply with a magic number $\frac{1}{2}$. Consider $\psi(\mathbf{x}, y) = \frac{1}{2}y\tilde{\phi}(\mathbf{x})$.

Our next target: Prove that in the binary case:

$$\sum_{y \neq y^{(i)}} \max(1 + \tilde{\boldsymbol{\theta}}^\top \psi(\mathbf{x}^{(i)}, y) - \tilde{\boldsymbol{\theta}}^\top \psi(\mathbf{x}^{(i)}, y^{(i)}), 0)$$

is equivalent to

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

1(b): Define $\psi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^{d+1}$

1. Motivation: functional margin has form $y(\langle \phi(\mathbf{x}), \boldsymbol{\theta} \rangle + \theta_0)$. We need it into an inner product $\langle \cdot, \cdot \rangle$.
2. We need to merge θ_0 into the inner product. We can add a dummy feature 1 to $\phi(\mathbf{x})$, as $\langle \phi(\mathbf{x}), \boldsymbol{\theta} \rangle + \theta_0 = \langle (1, \phi(\mathbf{x}))^\top, (\theta_0, \boldsymbol{\theta})^\top \rangle$. Define $\tilde{\phi}(\mathbf{x}) = (1, \phi(\mathbf{x}))^\top$, and $\tilde{\boldsymbol{\theta}} = (\theta_0, \boldsymbol{\theta})^\top$.
3. We can merge the coefficient y into $\tilde{\phi}(\mathbf{x})$, obtaining $y\tilde{\phi}(\mathbf{x})$.
4. We have transformed $y(\langle \phi(\mathbf{x}), \boldsymbol{\theta} \rangle + \theta_0)$ into inner product $\langle y\tilde{\phi}(\mathbf{x}), \tilde{\boldsymbol{\theta}} \rangle$.
5. Multiply with a magic number $\frac{1}{2}$. Consider $\psi(\mathbf{x}, y) = \frac{1}{2}y\tilde{\phi}(\mathbf{x})$.

Our next target: Prove that in the binary case:

$$\sum_{y \neq y^{(i)}} \max(1 + \tilde{\boldsymbol{\theta}}^\top \psi(\mathbf{x}^{(i)}, y) - \tilde{\boldsymbol{\theta}}^\top \psi(\mathbf{x}^{(i)}, y^{(i)}), 0)$$

is equivalent to

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

1(b): Define $\psi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^{d+1}$

1. Motivation: functional margin has form $y(\langle \phi(\mathbf{x}), \boldsymbol{\theta} \rangle + \theta_0)$. We need it into an inner product $\langle \cdot, \cdot \rangle$.
2. We need to merge θ_0 into the inner product. We can add a dummy feature 1 to $\phi(\mathbf{x})$, as $\langle \phi(\mathbf{x}), \boldsymbol{\theta} \rangle + \theta_0 = \langle (1, \phi(\mathbf{x}))^\top, (\theta_0, \boldsymbol{\theta})^\top \rangle$. Define $\tilde{\phi}(\mathbf{x}) = (1, \phi(\mathbf{x}))^\top$, and $\tilde{\boldsymbol{\theta}} = (\theta_0, \boldsymbol{\theta})^\top$.
3. We can merge the coefficient y into $\tilde{\phi}(\mathbf{x})$, obtaining $y\tilde{\phi}(\mathbf{x})$.
4. We have transformed $y(\langle \phi(\mathbf{x}), \boldsymbol{\theta} \rangle + \theta_0)$ into inner product $\langle y\tilde{\phi}(\mathbf{x}), \tilde{\boldsymbol{\theta}} \rangle$.
5. Multiply with a magic number $\frac{1}{2}$. Consider $\psi(\mathbf{x}, y) = \frac{1}{2}y\tilde{\phi}(\mathbf{x})$.

Our next target: Prove that in the binary case:

$$\sum_{y \neq y^{(i)}} \max(1 + \tilde{\boldsymbol{\theta}}^\top \psi(\mathbf{x}^{(i)}, y) - \tilde{\boldsymbol{\theta}}^\top \psi(\mathbf{x}^{(i)}, y^{(i)}), 0)$$

is equivalent to

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

1 (b): We need to reach $\max(1 - y^{(i)}(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\mathbf{x}^{(i)}) + \theta_0), 0)$

1. In the binary case, $\sum_{y \neq y^{(i)}} \max(1 + \tilde{\boldsymbol{\theta}}^\top \boldsymbol{\psi}(\mathbf{x}^{(i)}, y) - \tilde{\boldsymbol{\theta}}^\top \boldsymbol{\psi}(\mathbf{x}^{(i)}, y^{(i)}), 0)$ has **only one term**.
2. The only term corresponds to $y^{(i)} = +1, y = -1$ or $y^{(i)} = -1, y = +1$.

Therefore,

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

1 (b): We need to reach $\max(1 - y^{(i)}(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\mathbf{x}^{(i)}) + \theta_0), 0)$

1. In the binary case, $\sum_{y \neq y^{(i)}} \max(1 + \tilde{\boldsymbol{\theta}}^\top \boldsymbol{\psi}(\mathbf{x}^{(i)}, y) - \tilde{\boldsymbol{\theta}}^\top \boldsymbol{\psi}(\mathbf{x}^{(i)}, y^{(i)}), 0)$ has **only one term**.
2. The only term corresponds to $y^{(i)} = +1, y = -1$ or $y^{(i)} = -1, y = +1$.

Therefore,

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

1 (b): We need to reach $\max(1 - y^{(i)}(\boldsymbol{\theta}^\top \phi(\mathbf{x}^{(i)}) + \theta_0), 0)$

1. In the binary case, $\sum_{y \neq y^{(i)}} \max(1 + \tilde{\boldsymbol{\theta}}^\top \psi(\mathbf{x}^{(i)}, y) - \tilde{\boldsymbol{\theta}}^\top \psi(\mathbf{x}^{(i)}, y^{(i)}), 0)$ has **only one term**.
2. The only term corresponds to $y^{(i)} = +1, y = -1$ or $y^{(i)} = -1, y = +1$.

Therefore,

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

Solution to 1 (b): Continued

Thus,

$$\begin{aligned}\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) &= \frac{1}{2} \|\boldsymbol{\theta}\|^2 + C \sum_{i=1}^n \sum_{y \neq y^{(i)}} \max(1 + \tilde{\boldsymbol{\theta}}^\top \psi(\mathbf{x}^{(i)}, y) - \tilde{\boldsymbol{\theta}}^\top \psi(\mathbf{x}^{(i)}, y^{(i)}), 0) \\ &= \frac{1}{2} \|\boldsymbol{\theta}\|^2 + C \sum_{i=1}^n \max(1 - y^{(i)} \tilde{\boldsymbol{\theta}}^\top \tilde{\phi}(\mathbf{x}^{(i)}), 0) \\ &= \frac{1}{2} \|\boldsymbol{\theta}\|^2 + C \sum_{i=1}^n \max(1 - y^{(i)} (\boldsymbol{\theta}^\top \phi(\mathbf{x}^{(i)}) + \theta_0), 0).\end{aligned}$$

Exercise 1 (c)

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

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

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

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

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

1 (c): Express $\|\boldsymbol{\theta}\|_2^2$ with \mathbf{K} and $\boldsymbol{\beta}$

$\boldsymbol{\theta}^* \in \text{span}\{(\psi(\mathbf{x}^{(i)}, y))_{i=1, \dots, n, y=1, \dots, g}\}$ means that $\boldsymbol{\theta}$ is a linear combination of the spanning bases,

i.e. $\boldsymbol{\theta} = \mathbf{X}^\top \boldsymbol{\beta}$ for $\boldsymbol{\beta} \in \mathbb{R}^{ng}$ and

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

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

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

1 (c): Express $\|\boldsymbol{\theta}\|_2^2$ with \mathbf{K} and $\boldsymbol{\beta}$

$\boldsymbol{\theta}^* \in \text{span}\{(\psi(\mathbf{x}^{(i)}, y))_{i=1, \dots, n, y=1, \dots, g}\}$ means that $\boldsymbol{\theta}$ is a linear combination of the spanning bases,

i.e. $\boldsymbol{\theta} = \mathbf{X}^\top \boldsymbol{\beta}$ for $\boldsymbol{\beta} \in \mathbb{R}^{ng}$ and

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

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

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

1 (c): Express $\|\boldsymbol{\theta}\|_2^2$ with \mathbf{K} and $\boldsymbol{\beta}$

$\boldsymbol{\theta}^* \in \text{span}\{(\psi(\mathbf{x}^{(i)}, y))_{i=1, \dots, n, y=1, \dots, g}\}$ means that $\boldsymbol{\theta}$ is a linear combination of the spanning bases,

i.e. $\boldsymbol{\theta} = \mathbf{X}^\top \boldsymbol{\beta}$ for $\boldsymbol{\beta} \in \mathbb{R}^{ng}$ and

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

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

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

1 (c): Express $\theta^\top \psi(\mathbf{x}^{(i)}, y)$ with \mathbf{K} and β

Furthermore,

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

Note that the result is a scalar.

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

Therefore,

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

1 (c): Express $\theta^\top \psi(\mathbf{x}^{(i)}, y)$ with \mathbf{K} and β

Furthermore,

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

Note that the result is a scalar.

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

Therefore,

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

1 (c): Express $\theta^\top \psi(\mathbf{x}^{(i)}, y)$ with \mathbf{K} and β

Furthermore,

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

Note that the result is a scalar.

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

Therefore,

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

1 (c): Express $\theta^\top \psi(\mathbf{x}^{(i)}, y)$ with \mathbf{K} and β

Furthermore,

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

Note that the result is a scalar.

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

Therefore,

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

1 (c): Express $\theta^\top \psi(\mathbf{x}^{(i)}, y)$ with \mathbf{K} and β

Furthermore,

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

Note that the result is a scalar.

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

Therefore,

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

Exercise 2: Kernel Trick

The polynomial kernel is defined as

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

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)

$$k(x, \tilde{x}) = (x^\top \tilde{x} + b)^2 = \left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^\top \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix} + b \right)^2$$

Solution to 2 (a)

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

Solution to 2 (a)

$$\begin{aligned}k(x, \tilde{x}) &= (x^\top \tilde{x} + b)^2 = \left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^\top \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix} + b \right)^2 \\&= (x_1 \tilde{x}_1 + x_2 \tilde{x}_2 + b)^2 \\&= x_1^2 \tilde{x}_1^2 + 2x_1 \tilde{x}_1 x_2 \tilde{x}_2 + x_2^2 \tilde{x}_2^2 + 2bx_1 \tilde{x}_1 + 2bx_2 \tilde{x}_2 + b^2\end{aligned}$$

Solution to 2 (a)

$$\begin{aligned}k(x, \tilde{x}) &= (x^\top \tilde{x} + b)^2 = \left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^\top \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix} + b \right)^2 \\&= (x_1 \tilde{x}_1 + x_2 \tilde{x}_2 + b)^2 \\&= x_1^2 \tilde{x}_1^2 + 2x_1 \tilde{x}_1 x_2 \tilde{x}_2 + x_2^2 \tilde{x}_2^2 + 2bx_1 \tilde{x}_1 + 2bx_2 \tilde{x}_2 + b^2 \\&= \left\langle \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1 x_2 \\ x_2^2 \\ \sqrt{2}bx_1 \\ \sqrt{2}bx_2 \\ b \end{pmatrix}, \begin{pmatrix} \tilde{x}_1^2 \\ \sqrt{2}\tilde{x}_1 \tilde{x}_2 \\ \tilde{x}_2^2 \\ \sqrt{2}b\tilde{x}_1 \\ \sqrt{2}b\tilde{x}_2 \\ b \end{pmatrix} \right\rangle \\&= \langle \phi(x), \phi(\tilde{x}) \rangle\end{aligned}$$

Exercise 2 (b)

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

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

Exercise 2 (b)

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

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