Exercises for the course

Machine Learning 1

Winter semester 2023/24

Fachgebiet Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 13

We consider a class optimization problems of the type:

$$\min_{\theta} J(\theta) \quad \text{s.t.} \quad \forall_{i=1}^{m}: \; g_i(\theta) = 0 \quad \text{and} \quad \forall_{i=1}^{l}: \; h_i(\theta) \leq 0$$

For this class of problem, we can build the Lagrangian:

$$\mathcal{L}(\theta, \beta, \lambda) = J(\theta) + \sum_{i=1}^{m} \beta_i g_i(\theta) + \sum_{i=1}^{l} \lambda_i h_i(\theta).$$

where $(\beta_i)_i$ and $(\lambda_i)_i$ are the dual variables. According to the Karush-Kuhn-Tucker (KKT) conditions, it is necessary for a solution of this optimization problem that the following constraints are satisfied (in addition to the original constraints of the optimization problem):

$$\frac{\partial \mathcal{L}}{\partial \theta} = 0 \qquad \text{(stationarity)}$$

$$\forall_{i=1}^{l}: \; \lambda_i \geq 0 \qquad \text{(dual feasibility)}$$

$$\forall_{i=1}^{l}: \; \lambda_i h_i(\theta) = 0 \qquad \text{(complementary slackness)}$$

We will make use of these conditions to derive the dual form of the kernel ridge regression problem.

Exercise 1: Kernel Ridge Regression with Lagrange Multipliers (10 + 20 + 10 + 10 P)

Let $x_1, \ldots, x_N \in \mathbb{R}^d$ be a dataset with labels $y_1, \ldots, y_N \in \mathbb{R}$. Consider the regression model $f(x) = w^\top \phi(x)$ where $\phi \colon \mathbb{R}^d \to \mathbb{R}^h$ is a feature map and w is obtained by solving the constrained optimization problem

$$\min_{\xi, w} \sum_{i=1}^{N} \frac{1}{2} \xi_i^2 \quad \text{s.t.} \quad \forall_{i=1}^{N}: \; \xi_i = w^\top \phi(x_i) - y_i \quad \text{and} \quad \frac{1}{2}\|w\|^2 \leq C.$$

where equality constraints define the errors of the model, where the objective function penalizes these errors, and where the inequality constraint imposes a regularization on the parameters of the model.

(a) *Construct* the Lagrangian and *state* the KKT conditions for this problem *(Hint: rewrite the equality constraint as $\xi_i - w^\top \phi(x_i) + y_i = 0$.)*

(b) *Show* that the solution of the kernel regression problem above, expressed in terms of the dual variables $(\beta_i)_i$, and λ is given by:

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

where K is the kernel Gram matrix.

(c) *Express* the prediction $f(x) = w^\top \phi(x)$ in terms of the parameters of the dual.

(d) *Explain* how the new parameter λ can be related to the parameter C of the original formulation.

Exercise 2: Programming (50 P)

Download the programming files on ISIS and follow the instructions.

Exercise Sheet 13

We consider a class optimization problems of the type:

$$\min_{\theta} J(\theta) \quad \text{s.t.} \quad \forall_{i=1}^{m}: \; g_i(\theta) = 0 \quad \text{and} \quad \forall_{i=1}^{l}: \; h_i(\theta) \le 0$$

For this class of problem, we can build the Lagrangian:

$$\mathcal{L}(\theta, \beta, \lambda) = J(\theta) + \sum_{i=1}^{m} \beta_i g_i(\theta) + \sum_{i=1}^{l} \lambda_i h_i(\theta).$$

where $(\beta_i)_i$ and $(\lambda_i)_i$ are the dual variables. According to the Karush-Kuhn-Tucker (KKT) conditions, it is necessary for a solution of this optimization problem that the following constraints are satisfied (in addition to the original constraints of the optimization problem):

$$\frac{\partial \mathcal{L}}{\partial \theta} = 0 \qquad \text{(stationarity)}$$

$$\forall_{i=1}^{l}: \; \lambda_i \ge 0 \qquad \text{(dual feasibility)}$$

$$\forall_{i=1}^{l}: \; \lambda_i h_i(\theta) = 0 \qquad \text{(complementary slackness)}$$

We will make use of these conditions to derive the dual form of the kernel ridge regression problem.

Exercise 1: Kernel Ridge Regression with Lagrange Multipliers $(10 + 20 + 10 + 10 \text{ P})$

Let $x_1, \dots, x_N \in \mathbb{R}^d$ be a dataset with labels $y_1, \dots, y_N \in \mathbb{R}$. Consider the regression model $f(x) = w^\top \phi(x)$ where $\phi \colon \mathbb{R}^d \to \mathbb{R}^h$ is a feature map and w is obtained by solving the constrained optimization problem

$$\min_{\xi, w} \sum_{i=1}^{N} \frac{1}{2}\xi_i^2 \quad \text{s.t.} \quad \forall_{i=1}^{N}: \; \xi_i = w^\top \phi(x_i) - y_i \quad \text{and} \quad \frac{1}{2}\|w\|^2 \le C.$$

where equality constraints define the errors of the model, where the objective function penalizes these errors, and where the inequality constraint imposes a regularization on the parameters of the model.

(a) *Construct* the Lagrangian and *state* the KKT conditions for this problem *(Hint: rewrite the equality constraint as $\xi_i - w^\top \phi(x_i) + y_i = 0$.)*

$$\mathcal{L}(\xi, \beta, \lambda, w) = \sum_{i=1}^{N} \frac{1}{2}\xi_i^2 + \sum_{i=1}^{N} \beta_i(\xi_i - w^T \Phi(x_i) + y_i) + \lambda\left(\frac{1}{2}\|w\|^2 - C\right)$$

1) stationarity

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = \xi_i + \beta_i = 0$$

$$\frac{\partial \mathcal{L}}{\partial w} = \sum_{i=1}^{N} -\beta_i \Phi(x_i) + \lambda w$$

$$\Rightarrow \sum_{i=1}^{N} \beta_i \Phi(x_i) = \lambda w$$

2) dual feasibility $\quad \lambda \geq 0$

3) complementary slackness $\quad \lambda\left(\frac{1}{2}\|w\|^2 - C\right) = 0$

(b) *Show* that the solution of the kernel regression problem above, expressed in terms of the dual variables $(\beta_i)_i$, and λ is given by:

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

where K is the kernel Gram matrix.

$$\begin{cases} \forall_{i=1}^{N}: \xi_i = w^T \Phi(x_i) - y_i \\ \xi_i = -\beta_i \quad \text{from a)} \end{cases} \Rightarrow -\beta_i = w^T \Phi(x_i) - y_i$$

from (a) we have: $\sum_{i=1}^{N} \beta_i \Phi(x_i) = \lambda w \quad \Rightarrow w = \frac{1}{\lambda} \sum_{i=1}^{N} \beta_i \Phi(x_i)$

$$-\beta_i = w^T \Phi(x_i) - y_i$$

$$\Rightarrow \quad -\beta_i = \frac{1}{\lambda} \sum_{j=1}^{N} \beta_j \underbrace{\Phi(x_j)^T \Phi(x_i)}_{K_{ji}} - y_i$$

$$\Rightarrow \quad -\beta_i = \frac{1}{\lambda} \sum_{j=1}^{N} \beta_j \cdot K_{ji} - y_i$$

$$\Rightarrow \quad -\beta = \frac{1}{\lambda} \beta \cdot K - y$$

$$\Rightarrow \quad \left(\frac{1}{\lambda} \cdot K + I\right) \beta = y$$

$$\Rightarrow \quad (K + \lambda \cdot I) \beta = \lambda y$$

$$\Rightarrow \quad \beta = (K + \lambda I)^{-1} \lambda \cdot y$$

(c) *Express* the prediction $f(x) = w^\top \phi(x)$ in terms of the parameters of the dual.

$$\begin{cases} w = \frac{1}{\lambda} \sum_{i=1}^{N} \beta_i \Phi(x_i) \\ f(x) = w^T \Phi(x) \end{cases}$$

$\Rightarrow f(x) = \frac{1}{\lambda} \sum_{i=1}^{N} \beta_i \Phi(x_i)^T \Phi(x)$ (from b we have $\beta = (K + \lambda I)^{-1} \lambda \cdot y$)

$$= \frac{1}{\cancel{\lambda}} \cdot \sum_{i=1}^{N} \left[(K + \lambda I)^{-1} \cancel{\lambda} y\right]_i \underbrace{\Phi(x_i) \Phi(x)}_{k(x_i, x)}$$

$$= \sum_{i=1}^{N} \left[(K + \lambda I)^{-1} \cdot y\right]_i \cdot k(x_i, x)$$

$$= (k + \lambda I)^{-1} \cdot k(x_i, x) \cdot y$$

(d) *Explain* how the new parameter λ can be related to the parameter C of the original formulation.

from a) we have $\lambda\left(\frac{1}{2} \|w\|^2 - C\right) = 0$, therefore we have 2 cases

<u>Case 1</u> $\lambda = 0$ and $\frac{1}{2} \|w\|^2 < C$

That means the constraint is not used

<u>Case 2</u> $\lambda \neq 0$ and $\frac{1}{2} \|w\|^2 = C$

That means the ridge regularization does exist

$$C = \frac{1}{2} \|w\|^2 = \frac{1}{2} w^T \cdot w$$

$$= \frac{1}{2\lambda^2} \left[\sum_{i=1}^{N} \beta_i \Phi(x_i)\right]^T \cdot \sum_{i=1}^{N} \beta_i \Phi(x_i)$$

$$= \frac{1}{2\lambda^2} \cdot \left(\Phi(x)^T \cdot \beta \right)^T \cdot \Phi(x)^T \beta$$

$$= \frac{1}{2\lambda^2} \quad \beta^T \Phi(x) \Phi(x)^T \beta$$

$$= \frac{1}{2\lambda^2} \quad \beta^T K \beta$$

(from b) we have $\beta = (K + \lambda I)^{-1} \lambda \cdot y$)

$$= \frac{1}{2\lambda^2} \cdot \left[(K+\lambda I)^{-1} \cdot \lambda \cdot y \right]^T K \, (K+\lambda I)^{-1} \lambda y$$

$$= \frac{1}{2} \cdot \quad y^T \cdot \left[(K+\lambda I)^{-1} \right]^T \cdot K \cdot (K+\lambda I)^{-1} y$$

If we choose a small C, that means $\|w\|$ cannot be too large. and λ must be larger

If C is large, we can let $\|w\|$ grow as large, the λ becomes smaller. Specifically, if C is large enough,

the λ will led to 0, which means there will be no Regularization. (case 1)