Exercises for the course

Machine Learning 1

Winter semester 2021/22

Abteilung 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 9

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 $x_1, \dots, x_N \in \mathbb{R}^d$ 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

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

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.

ML 9

1

a) The Lagrangian becomes

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

We now verify the 3 Karush-Kuhn-Tucker (KKT) conditions:

1) stationarity: $\frac{\partial \mathcal{L}}{\partial w} = - \sum_{i=1}^{N} \beta_i \phi(x_i) + \lambda w = 0$

This condition holds if $\lambda w = \sum_{i=1}^{N} \beta_i \phi(x_i)$

and $\frac{\partial \mathcal{L}}{\partial \xi_i} = \sum_{i=1}^{N} (\xi_i + \beta_i) = 0$

This condition is verified if $\xi_i = - \beta_i$ for all i.

2) Dual feasibility: $\forall_{i=1}^{N} : \beta_i \geq 0$

This condition must hold.

3) Complementary slackness:

Further, $\beta \left(\frac{1}{2} \|w\|^2 - c \right) = 0$ must hold for the KKT conditions.

b) We use the constraint

$$\xi_i = w^\top \phi(x_i) - y_i$$

$$\Leftrightarrow \lambda \xi_i = \lambda w^\top \phi(x_i) - \lambda y_i$$

We use $\xi_i = -\beta_i$ and $\lambda w = \sum_j \beta_j \phi(x_j)$ from exercise a):

$$-\lambda \beta_i = \sum_j \beta_j \phi(x_j)^\top \phi(x_i) - \lambda y_i$$

$$\Leftrightarrow -\lambda \mathbb{1} \beta = \sum_j \beta_j K_{ij} - \lambda y_i = K\beta - \lambda y$$

$$\Leftrightarrow \lambda y = K\beta + \lambda \mathbb{1} \beta = (K + \lambda \mathbb{1})\beta$$

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

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

From exercise b) we know

$$\lambda w = \sum_j \beta_j \phi(x_j) \iff w^T = \frac{1}{\lambda} \sum_j \beta_j \phi^T(x_j)$$

Therefore,

$$w^T \phi(x) = \frac{1}{\lambda} \sum_j \beta_j \phi^T(x_j) \cdot \phi(x) .$$

In b) we found $\beta = (K + \lambda \mathbb{1})^{-1} \lambda y$ which we insert:

$$w^T \phi(x) = \sum_j \frac{1}{\cancel{\lambda}} (K + \lambda \mathbb{1})^{-1} \cancel{\lambda} y_j \phi^T(x_j) \phi(x)$$

$$= (K + \lambda \mathbb{1})^{-1} y \, k(X, x)$$

$$= k(x, X) (K + \lambda \mathbb{1})^{-1} \cdot y$$

d) <u>Explain how the new parameter λ can be related to the parameter C of the original formulation.</u>

We saw in part a):

$\lambda(\frac{1}{2}\|w\|^2 - C) = 0$ $\rightarrow$ either λ or $\frac{1}{2}\|w\|^2 - C$ must be 0.

$\lambda = 0$ if $\frac{1}{2}\|w\|^2 \neq C$

Case 1: $\frac{1}{2}\|w\|^2 < C$ or $\frac{1}{2}\|w\|^2 > C \rightarrow \lambda = 0$

This corresponds to unregularized kernel regression

Case 2: $\frac{1}{2}\|w\|^2 = C$

λ must be larger or equal to 0.

This corresponds to kernel ridge regression.

From a) we know $\lambda w = \sum_i \beta_i \phi(x_i)$

$\Leftrightarrow w = \frac{1}{\lambda} \beta_i \phi(x_i) \Leftrightarrow \|w\|^2 = \frac{1}{\lambda^2} \beta^T K \beta$

Therefore, $C = \frac{1}{2}\|w\|^2 = \frac{1}{2}\frac{1}{\lambda^2}\beta^T K \beta$.

Using $\beta = (K + \lambda \mathbb{1})^{-1} \lambda y$ from b) this becomes:

$$C = \frac{1}{2\cancel{\lambda^2}} y^T \cancel{\lambda} (K + \lambda \mathbb{1})^{-1} K (K + \lambda \mathbb{1})^{-1} \cancel{\lambda} y$$

$$= \frac{1}{2} y^T (K + \lambda \mathbb{1})^{-1} K (K + \lambda \mathbb{1})^{-1} y$$