Exercise 1: SVM – Regression

For the data set $\mathcal{D} = \left(\left(\mathbf{x}^{(1)}, y^{(1)}\right), \dots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)\right)$ with $y^{(i)} \in \mathbb{R}$, assume that for a fixed $\epsilon > 0$ all observations are within the ϵ-tube around $f(\mathbf{x} \mid \theta_0, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{x} + \theta_0$ for any $(\theta_0, \boldsymbol{\theta})^\top \in \tilde{\Theta}$, i.e.,

$$y^{(i)} \in \left[f(\mathbf{x}^{(i)} \mid \theta_0, \boldsymbol{\theta}) - \epsilon,\ f(\mathbf{x}^{(i)} \mid \theta_0, \boldsymbol{\theta}) + \epsilon\right], \quad \forall i \in \{1, \dots, n\},\ \forall (\theta_0, \boldsymbol{\theta})^\top \in \tilde{\Theta},$$

where $\tilde{\Theta} \subset \mathbb{R}^{p+1}$ is some non-empty parameter subset. Let

$$d_\epsilon \left(f(\cdot \mid \theta_0, \boldsymbol{\theta}), \mathbf{x}^{(i)}\right) := \epsilon - |y^{(i)} - f(\mathbf{x}^{(i)} \mid \theta_0, \boldsymbol{\theta})| = \epsilon - |y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)} - \theta_0|$$

be the (signed) ϵ-distance of the prediction error. The maximal ϵ-distance of the prediction error of f to the whole data set $\mathcal{D}$ is

$$\gamma_\epsilon = \max_{i=1,\dots,n} \left\{ d_\epsilon \left(f(\cdot \mid \theta_0, \boldsymbol{\theta}), \mathbf{x}^{(i)}\right) \right\}.$$

(a) Let $(\theta_0, \boldsymbol{\theta})^\top \in \mathbb{R}^{p+1}$ be arbitrary. Which (type of) values does $d_\epsilon \left(f(\cdot \mid \theta_0, \boldsymbol{\theta}), \mathbf{x}^{(i)}\right)$ have if $y^{(i)}$ is

- within the ϵ-tube around $f(\mathbf{x}^{(i)} \mid \theta_0, \boldsymbol{\theta})$?
- not within the ϵ-tube around $f(\mathbf{x}^{(i)} \mid \theta_0, \boldsymbol{\theta})$?

What would be a desirable choice of the parameters $(\theta_0, \boldsymbol{\theta})^\top$ with respect to γ_ϵ? Is the choice of the parameters unique in general?

(b) Argue that

$$\begin{aligned} \min_{\boldsymbol{\theta}, \theta_0} \quad & \frac{1}{2}\|\boldsymbol{\theta}\|^2 \\ \text{s.t.} \quad & \epsilon - y^{(i)} + \boldsymbol{\theta}^\top \mathbf{x}^{(i)} + \theta_0 \geq 0 \quad \forall\, i \in \{1, \dots, n\} \\ \text{and} \quad & \epsilon + y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)} - \theta_0 \geq 0 \quad \forall\, i \in \{1, \dots, n\}. \end{aligned}$$

is a suitable optimization problem for the desired choice in (a).

(c) Derive the Lagrange function $L(\boldsymbol{\theta}, \theta_0, \boldsymbol{\alpha})$ of the optimization problem as well as its dual form.

(d) Find the stationary points of L. What can be inferred from the solution of the dual problem?

(e) Derive the "softened" version of the optimization problem in (b).

(f) Rewrite the "softened" version of the optimization problem in (b) as a regularized empirical risk minimization problem for a suitable loss function for regression.

Exercise 2: SVM – Optimization

Write your own stochastic subgradient descent routine to solve the soft-margin SVM in the primal formulation.

Hints:

- Use the regularized-empirical-risk-minimization formulation, i.e., an optimization criterion without constraints.
- No kernels, just a linear SVM.
- Compare your implementation with an existing implementation (e.g., `kernlab` in R). Are your results similar? Note that you might have to switch off the automatic data scaling in the already existing implementation.