Support Vector Regression
using
Deflected Subgradient Methods

Elia Piccoli
Nicola Gugole

February 7, 2021

A project presented for the
Computational Mathematics for Learning and Data Analysis
course

University of Pisa
Artificial Intelligence
A.Y. 2020/2021

Contents

1 Introduction **2**

2 Dual Representation **3**

3 APPENDIX A **5**

Abstract

Project aim is developing the implementation of a model which follows an SVR-type approach including various different kernels. The implementation uses as optimization algorithm a dual approach with appropriate choices of the constraints to be dualized, where the Lagrangian Dual is solved by an algorithm of the class of deflected subgradient methods.

1 Introduction

SVR objective is predicting a uni-dimensional real-valued output y through the use of an *objective function* built by optimization using an ε-insensitive loss function. Another fundamental aspect about SVR is keeping the function *as flat as possible* through the tuning of a C parameter in order to avoid overfitting and generating a correct trade-off between accuracy and generalization.
The resulting function can be generically described as:

$$f(x) = wx + b \tag{1}$$

Keeping the above function *as flat as possible* is equivalent to an optimization problem formulated as having minimum $\|w\|$, or, for a more convenient mathematical derivation, minimum $\|w\|^2$, not changing the semantics of the problem.
This brings us to a convex minimization problem, which will be called *primal problem*:

$$\min_{w,\xi_i,\xi_i^*} \frac{1}{2}\|w\|^2 + C\sum_i(\xi_i + \xi_i^*) \tag{2}$$

Where ξ and ξ^* are called *slack variables*, used in conjunction with C to create a *regularization factor* and consequently a *penalty measure* to elements which are not part of the ε-tube. Slack variables allow the definition of constraints applicable to (2):

$$y_i - w^T\phi(x_i) - b \leq \varepsilon + \xi_i, \tag{3a}$$

$$b + w^T\phi(x_i) - y_i \leq \varepsilon + \xi_i, \tag{3b}$$

$$\xi_i, \xi_i^* \geq 0 \tag{3c}$$

x_i input, y_i output

2 Dual Representation

As expressed in the abstract, the implementation will follow a dual approach, which in SVR models is preferred due to the applicability and efficiency of the use of *kernels*. *Dual problem* formulation can be achieved defining the *Lagrangian* function:

$$
\begin{aligned}
\mathcal{L}(\alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|w\|^2 \\
&+ C\sum_{i=1}^{m}(\xi_i + \xi_i^*) \\
&+ \sum_{i=1}^{m}(\alpha_i(y_i - w^T\phi(x_i) - b - \varepsilon - \xi_i)) \\
&+ \sum_{i=1}^{m}(\alpha_i^*(w^T\phi(x_i) + b - y_i - \varepsilon - \xi_i^*)) \\
&- \sum_{i=1}^{m}(\mu_i\xi_i + \mu_i^*\xi_i^*)
\end{aligned}
\tag{4}
$$

From which the following optimization problem can be obtained (full derivation shown in 3):

$$
\begin{aligned}
\max_{\alpha_i,\alpha_i^*} -\frac{1}{2}\sum_i\sum_j(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j) \\
- \epsilon\sum_i(\alpha_i + \alpha_i^*) \\
+ \sum_i y_i(\alpha_i - \alpha_i^*)
\end{aligned}
\tag{5}
$$

With constraints:

$$\forall i\ \alpha_i, \alpha_i^* \geq 0 \qquad (KKT\ condition) \tag{6a}$$

$$\forall i\ \alpha_i, \alpha_i^* \in [0, C] \qquad (from\ derivation) \tag{6b}$$

$$\forall i\ \sum(\alpha_i - \alpha_i^*) = 0 \qquad (from\ derivation) \tag{6c}$$

$$\forall i\ \alpha_i\alpha_i^* = 0 \qquad (from\ model\ construction) \tag{6d}$$

At this point a reformulation of (5) is necessary to follow the task objective, which is solving the *Lagrangian Dual* maximization with a subgradient method, therefore requiring a *non-differentiable function.* Such function is achievable with a simple variable substitution:

$$\beta_i \longleftarrow (\alpha_i - \alpha_i^*)$$
$$|\beta_i| \longleftarrow (\alpha_i + \alpha_i^*)$$

Bringing the definitive dual problem definition:

$$\begin{aligned}\max_{\beta_i} & -\frac{1}{2}\sum_i\sum_j \beta_i\beta_j K(x_i, x_j) \\ & -\epsilon\sum_i|\beta_i| \\ & +\sum_i y_i\beta_i \\ & \textit{With the constraints} \qquad \begin{cases}\sum_i \beta_i = 0 \\ \beta_i \in [-C,\ C] \\ |\beta_i| \in [0,\ C]\end{cases}\end{aligned} \tag{7}$$

It is important to notice how the above formulation defines a convex non-differentiable problem which still maintains the *strong duality* property of SVR since it is possible to reshape (5) to a quadratic problem, as shown in APPENDIX B (qui la nostra prossima idea per l'appendice B era ripartire da (5) e mostrare come si può arrivare a formulare il problema quadratico tramite l'uso della matrice $Q = [K\ \ -K\ ; -K\ \ K]$, sarebbe sufficiente come dimostrazione?).
Having the *strong duality* property assures that the optimal solution of the dual problem coincides with the one of the primal problem.

3 APPENDIX A

Define the Lagrangian function

$$
\begin{aligned}
\mathcal{L} = \frac{1}{2}\|w\|^2 + C\sum_i(\xi_i + \xi_i^*) &+ \sum_i \alpha_i(y_i - w\phi_i - b - \varepsilon - \xi_i) \\
&+ \sum_i \alpha_i(-y_i + w\phi_i - b - \varepsilon - \xi_i^*) \\
&- \sum_i \mu_i\xi_i \\
&- \sum_i \mu_i^*\xi_i^*
\end{aligned} \tag{8}
$$

$$\textit{where} \quad \forall_i \; \xi_i\xi_i^* \geq 0$$

Variables of the two definition of the problem:

Primal problem	$w,\ b,\ \xi_i,\ \xi_i^*$
Dual Problem	$\alpha_i,\ \alpha_i^*,\ \mu_i,\ \mu_i^*$

Next step is try to simplify the definition of the Lagrangian wrt the problem that needs to be solved. Since the objective is to find the *minimum* the developments proceeds imposing this condition.

$$\frac{\partial\mathcal{L}}{\partial w} = 0 \quad \longrightarrow \quad w + \sum_i \alpha_i(-\phi_i) + \sum_i \alpha_i^*\phi_i = 0 \tag{9a}$$

$$\frac{\partial\mathcal{L}}{\partial b} = 0 \quad \longrightarrow \quad \sum_i -\alpha_i + \sum_i \alpha_i^* = 0 \tag{9b}$$

$$\frac{\partial\mathcal{L}}{\partial \xi_i} = 0 \quad \longrightarrow \quad C - \alpha_i - \mu_i = 0 \tag{9c}$$

$$\frac{\partial\mathcal{L}}{\partial \xi_i^*} = 0 \quad \longrightarrow \quad C - \alpha_i^* - \mu_i^* = 0 \tag{9d}$$

From (9a) the definition of w can be derived

$$w = \sum_i(\alpha_i - \alpha_i^*)\phi_i \tag{10}$$

From (9b) the first constraint on the Lagrangian variables is obtained

$$\sum_i(\alpha_i^* - \alpha_i) = 0 \tag{11}$$

While from (9c)/(9d) with some further development the second constraint on the Lagrangian variables can be defined

$$
\begin{aligned}
&\alpha_i,\ \alpha_i^*,\ \mu_i,\ \mu_i^* \ \geq\ 0 \quad \forall_i \\
&C = \alpha_i + \mu_i \quad \longrightarrow \quad \alpha_i = C - \mu_i \\
&\Longrightarrow \quad \alpha_i\ \in\ [0,\ C] \\
&\textit{and equivalently} \quad \alpha_i^*\ \in\ [0,\ C]
\end{aligned}
$$

Simplify (8) using the substitution (10)

$$
\begin{aligned}
\mathcal{L}\ =& \frac{1}{2}\sum_i\sum_j(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\phi_i\phi_j \\
&-\sum_i\sum_j(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\phi_i\phi_j \\
&+\sum_i(\alpha_i-\alpha_i^*)y_i+\sum_i(\alpha_i-\alpha_i^*)b-\sum_i(\alpha_i+\alpha_i^*)\varepsilon \\
&+\sum_i\alpha_i(-\xi_i)+\sum_i\alpha_i^*(-\xi_i^*) \\
&-\sum_i\mu_i\xi_i-\sum_i\mu_i^*\xi_i^* \\
&+C\sum_i\xi_i+\xi_i^*
\end{aligned}
$$

Apply condition (11) and (9c) to simplify some terms and obtain the final formulation

$$
\begin{aligned}
\mathcal{L}(\alpha,\alpha^*)\ =&\ -\frac{1}{2}\sum_i\sum_j(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\phi_i\phi_j \\
&+\sum_i(\alpha_i-\alpha_i^*)y_i \\
&-\sum_i(\alpha_i+\alpha_i^*)\varepsilon \\
&With\ the\ constraints \qquad \begin{cases}\sum_i(\alpha_i^*-\alpha_i)=0\\ \alpha_i\in[0,\ C]\\ \alpha_i^*\in[0,\ C]\end{cases}
\end{aligned}
$$