Support Vector Regression

Max Welling
Department of Computer Science
University of Toronto
10 King's College Road
Toronto, M5S 3G5 Canada
welling@cs.toronto.edu

Abstract

This is a note to explain support vector regression. It is usefull to first read the ridge-regression and the SVM note.

1 SVR

In kernel ridge regression we have seen the final solution was not sparse in the variables $\boldsymbol{\alpha}$. We will now formulate a regression method that is sparse, i.e. it has the concept of support vectors that determine the solution.

The thing to notice is that the sparseness arose from complementary slackness conditions which in turn came from the fact that we had inequality constraints. In the SVM the penalty that was paid for being on the wrong side of the support plane was given by $C\sum_i \xi_i^k$ for positive integers k, where ξ_i is the orthogonal distance away from the support plane. Note that the term $||\mathbf{w}||^2$ was there to penalize large $\mathbf{w}$ and hence to regularize the solution. Importantly, there was *no* penalty if a data-case was on the right side of the plane. Because all these data-points do not have any effect on the final solution the $\boldsymbol{\alpha}$ was sparse. Here we do the same thing: we introduce a penalty for being to far away from predicted line $\mathbf{w}\Phi_i + b$, but once you are close enough, i.e. in some "epsilon-tube" around this line, there is no penalty. We thus expect that all the data-cases which lie inside the data-tube will have no impact on the final solution and hence have corresponding $\alpha_i = 0$. Using the analogy of springs: in the case of ridge-regression the springs were attached between the data-cases and the decision surface, hence every item had an impact on the position of this boundary through the force it exerted (recall that the surface was from "rubber" and pulled back because it was parameterized using a finite number of degrees of freedom or because it was regularized). For SVR there are only springs attached between data-cases outside the tube and these attach to the tube, not the decision boundary. Hence, data-items inside the tube have no impact on the final solution (or rather, changing their position slightly doesn't perturb the solution).

We introduce different constraints for violating the tube constraint from above and from below,

$$\begin{aligned} \text{minimize} - \mathbf{w}, \xi, \hat{\xi} \qquad & \frac{1}{2}||\mathbf{w}||^2 + \frac{C}{2}\sum_i (\xi_i^2 + \hat{\xi}_i^2) \\ \text{subject to} \qquad & \mathbf{w}^T\Phi_i + b - y_i \leq \varepsilon + \xi_i \;\; \forall i \\ & y_i - \mathbf{w}^T\Phi_i - b \leq \varepsilon + \hat{\xi}_i \;\; \forall i \end{aligned} \tag{1}$$

The primal Lagrangian becomes,

$$\mathcal{L}_P = \frac{1}{2}||\mathbf{w}||^2 + \frac{C}{2}\sum_i (\xi_i^2 + \hat{\xi}_i^2) + \sum_i \alpha_i(\mathbf{w}^T\Phi_i + b - y_i - \varepsilon - \xi_i) + \sum_i \hat{\alpha}_i(y_i - \mathbf{w}^T\Phi_i - b - \varepsilon - \hat{\xi}_i) \tag{2}$$

Remark I: We could have added the constraints that $\xi_i \geq 0$ and $\hat{\xi}_i \geq 0$. However, it is not hard to see that the final solution will have that requirement automatically and there is no sense in constraining the optimization to the optimal solution as well. To see this, imagine some ξ_i is negative, then, by setting $\xi_i = 0$ the cost is lower and non of the constraints is violated, so it is preferred. We also note due to the above reasoning we will always have at least one of the $\xi, \hat{\xi}$ zero, i.e. inside the tube both are zero, outside the tube one of them is zero. This means that at the solution we have $\xi\hat{\xi} = 0$.
Remark II: Note that we don't scale $\varepsilon = 1$ like in the SVM case. The reason is that $\{y_i\}$ now determines the scale of the problem, i.e. we have not over-parameterized the problem.

We now take the derivatives w.r.t. $\mathbf{w}$, b, ξ and $\hat{\xi}$ to find the following KKT conditions (there are more of course),

$$\mathbf{w} = \sum_i (\hat{\alpha}_i - \alpha_i)\Phi_i \tag{3}$$

$$\xi_i = \alpha_i/C \qquad \hat{\xi}_i = \hat{\alpha}_i/C \tag{4}$$

Plugging this back in and using that now we also have $\alpha_i\hat{\alpha}_i = 0$ we find the dual problem,

$$\begin{aligned} \text{maximize}_{\alpha,\hat{\alpha}} \qquad & -\frac{1}{2}\sum_{ij}(\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)(K_{ij} + \frac{1}{C}\delta_{ij}) + \sum_i (\hat{\alpha}_i - \alpha_i)y_i - \sum_i (\hat{\alpha}_i + \alpha_i)\varepsilon \\ \text{subject to} \qquad & \sum_i (\hat{\alpha}_i - \alpha_i) = 0 \\ & \alpha_i \geq 0, \quad \hat{\alpha}_i \geq 0 \qquad \forall i \end{aligned} \tag{5}$$

From the complementary slackness conditions we can read the sparseness of the solution out:

$$\alpha_i(\mathbf{w}^T\Phi_i + b - y_i - \varepsilon - \xi_i) = 0 \tag{6}$$

$$\hat{\alpha}_i(y_i - \mathbf{w}^T\Phi_i - b - \varepsilon - \hat{\xi}_i) = 0 \tag{7}$$

$$\xi_i\hat{\xi}_i = 0, \quad \alpha_i\hat{\alpha}_i = 0 \tag{8}$$

where we added the last conditions by hand (they don't seem to directly follow from the formulation). Now we clearly see that if a case is above the tube $\hat{\xi}_i$ will take on its smallest possible value in order to make the constraints satisfied $\hat{\xi}_i = y_i - \mathbf{w}^T\Phi_i - b - \varepsilon$. This implies that $\hat{\alpha}_i$ will take on a positive value and the farther outside the tube the larger the $\hat{\alpha}_i$ (you can think of it as a compensating force). Note that in this case $\alpha_i = 0$. A similar story goes if $\xi_i > 0$ and $\alpha_i > 0$. If a data-case is inside the tube the $\alpha_i, \hat{\alpha}_i$ are necessarily zero, and hence we obtain sparseness.

We now change variables to make this optimization problem look more similar to the SVM and ridge-regression case. Introduce $\beta_i = \hat{\alpha}_i - \alpha_i$ and use $\hat{\alpha}_i \alpha_i = 0$ to write $\hat{\alpha}_i + \alpha_i = |\beta_i|$,

$$\begin{array}{ll} \text{maximize}_{\boldsymbol{\beta}} & -\frac{1}{2}\sum_{ij}\beta_i\beta_j(K_{ij} + \frac{1}{C}\delta_{ij}) + \sum_i \beta_i y_i - \sum_i |\beta_i|\varepsilon \\ \text{subject to} & \sum_i \beta_i = 0 \end{array} \tag{9}$$

where the constraint comes from the fact that we included a bias term[1] b.

From the slackness conditions we can also find a value for b (similar to the SVM case). Also, as usual, the prediction of new data-case is given by,

$$y = \mathbf{w}^T \Phi(\mathbf{x}) + b = \sum_i \beta_i K(\mathbf{x}_i, \mathbf{x}) + b \tag{10}$$

It is an interesting exercise for the reader to work her way through the case where the penalty is linear instead of quadratic, i.e.

$$\begin{array}{lll} \text{minimize}_{\mathbf{w},\xi,\hat{\xi}} & \frac{1}{2}||\mathbf{w}||^2 + C\sum_i(\xi_i + \hat{\xi}_i) & \\ \text{subject to} & \mathbf{w}^T\Phi_i + b - y_i \leq \varepsilon + \xi_i \ \ \forall i & \\ & y_i - \mathbf{w}^T\Phi_i - b \leq \varepsilon + \hat{\xi}_i \ \ \forall i & (11) \\ & \xi_i \geq 0, \quad \hat{\xi}_i \geq 0 \ \ \forall i & (12) \end{array}$$

leading to the dual problem,

$$\begin{array}{lll} \text{maximize}_{\boldsymbol{\beta}} & -\frac{1}{2}\sum_{ij}\beta_i\beta_j K_{ij} + \sum_i \beta_i y_i - \sum_i |\beta_i|\varepsilon & \\ \text{subject to} & \sum_i \beta_i = 0 & (13) \\ & -C \leq \beta_i \leq +C \ \ \forall i & (14) \end{array}$$

where we note that the quadratic penalty on the size of $\boldsymbol{\beta}$ is replaced by a box constraint, as is to be expected in switching from L_2 norm to L_1 norm.

Final remark: Let's remind ourselves that the quadratic programs that we have derived are convex optimization problems which have a unique optimal solution which can be found efficiently using numerical methods. This is often claimed as great progress w.r.t. the old neural networks days which were plagued by many local optima.

[1] Note by the way that we could not use the trick we used in ridge-regression by defining a constant feature $\phi_0 = 1$ and $b = w_0$. The reason is that the objective does not depend on b.