Exercises for
Pattern Recognition
Peter Fischer, Shiyang Hu
Assignment 9, 16./19.12.2013

SOLUTION

General Information:

Exercises (1 SWS):	Tue 12:15 – 13:45 (0.154-115) and Fri 08:15 – 09:45 (0.151-115)
Certificate:	Oral exam at the end of the semester
Contact:	peter.fischer@fau.de, shiyang.hu@fau.de

Support Vector Regression

Exercise 1 In the lecture, you learn how an SVM can be used for classification. In this exercise, we consider *Support Vector Regression* (SVR). Let $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}, \boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}$ be a set of observations. The task for regression is to predict y_i from $\boldsymbol{x}_i$ according to the linear regression function:

$$F(\boldsymbol{x}) = \boldsymbol{\alpha}^T \boldsymbol{x} + \alpha_0, \tag{1}$$

for a weight vector $\boldsymbol{\alpha} \in \mathbb{R}^d$ and the bias $\alpha_0 \in \mathbb{R}$. The intuition behind SVR is to penalize only deviations that are larger than ϵ.

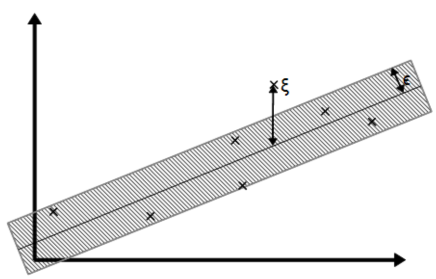

Figure 1: ϵ-tube of the SVR

The primal optimization problem for SVR is given by the following inequality-constraint minimization:

$$\begin{aligned}
\boldsymbol{\alpha}^* &= \operatorname*{argmin}_{\boldsymbol{\alpha}} \frac{1}{2}||\boldsymbol{\alpha}||^2 + C \sum_i (\xi_i + \hat{\xi}_i) \text{ , s.t.} \\
y_i &\leq \left(\boldsymbol{\alpha}^T \boldsymbol{x}_i + \alpha_0\right) + \epsilon + \xi_i \\
y_i &\geq \left(\boldsymbol{\alpha}^T \boldsymbol{x}_i + \alpha_0\right) - \epsilon - \hat{\xi}_i \\
\xi_i, \hat{\xi}_i &\geq 0
\end{aligned}$$

Here, ξ_i, $\hat{\xi}_i$ are slack variables (see also SVM classification) and ϵ specifies uncertainty of the regression function.

(a) Write down the Lagrangian L of the primal optimization problem using Lagrange multipliers $\lambda_i,\ \hat{\lambda}_i,\ \mu_i,\ \hat{\mu}_i$.
Hint: bring the constraints to the standard form $f_i\left(\boldsymbol{x}\right) \leq 0$
First, reformulate the inequality constraints to standard form $f_i(\boldsymbol{\alpha}) \leq 0$

$$
\begin{aligned}
y_i - \left(\boldsymbol{\alpha}^T \boldsymbol{x}_i + \alpha_0\right) - \epsilon - \xi_i \leq 0 \quad &\rightarrow f_i\left(\boldsymbol{\alpha}\right) \leq 0 \\
-y_i + \left(\boldsymbol{\alpha}^T \boldsymbol{x}_i + \alpha_0\right) - \epsilon - \hat{\xi}_i \leq 0 \quad &\rightarrow \hat{f}_i\left(\boldsymbol{\alpha}\right) \leq 0 \\
-\xi_i \leq 0 \\
-\hat{\xi}_i \leq 0
\end{aligned}
$$

no equality constraints

Summarize into one equation:

$$
\begin{aligned}
L\left(\boldsymbol{\alpha}, \alpha_0, \boldsymbol{\xi}, \hat{\boldsymbol{\xi}}, \boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\right) = &\frac{1}{2}||\boldsymbol{\alpha}||^2 + C\sum_i (\xi_i + \hat{\xi}_i) + \\
&\sum_i \lambda_i \left(y_i - \boldsymbol{\alpha}^T \boldsymbol{x}_i - \alpha_0 - \epsilon - \xi_i\right) + \\
&\sum_i \hat{\lambda}_i \left(-y_i + \boldsymbol{\alpha}^T \boldsymbol{x}_i + \alpha_0 - \epsilon - \hat{\xi}_i\right) + \\
&\sum_i \left(-\mu_i \xi_i - \hat{\mu}_i \hat{\xi}_i\right)
\end{aligned}
$$

(b) Write down the Karush-Kuhn-Tucker (KKT) conditions for the primal optimization problem given above.

(a) Primal constraints: see above

(b) Dual constraints: $\forall i:\ \lambda_i,\ \hat{\lambda}_i,\ \mu_i,\ \hat{\mu}_i \geq 0$

(c) Complementary slackness:
$\forall i:\ \lambda_i f_i\left(\boldsymbol{\alpha}\right) = 0,\ \hat{\lambda}_i \hat{f}_i\left(\boldsymbol{\alpha}\right) = 0,\ \mu_i \xi_i = 0,\ \hat{\mu}_i \hat{\xi}_i = 0$

(d) Gradient of Lagrangian is zero: $\nabla L\left(\boldsymbol{\alpha}, \alpha_0, \boldsymbol{\xi}, \hat{\boldsymbol{\xi}}, \boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\right) = 0$
The gradient of the Lagrangian is:

$$
\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{\alpha}} &= \boldsymbol{\alpha} - \sum_i \lambda_i \boldsymbol{x}_i + \sum_i \hat{\lambda}_i \boldsymbol{x}_i = 0 \Rightarrow \boldsymbol{\alpha} = \sum_i \left(\lambda_i - \hat{\lambda}_i\right) \boldsymbol{x}_i && (2) \\
\frac{\partial L}{\partial \alpha_0} &= -\sum_i \lambda_i + \sum_i \hat{\lambda}_i = 0 \Rightarrow \sum_i \left(\lambda_i - \hat{\lambda}_i\right) = 0 && (3) \\
\frac{\partial L}{\partial \xi_i} &= C - \lambda_i - \mu_i = 0 \Rightarrow \lambda_i + \mu_i = C && (4) \\
\frac{\partial L}{\partial \hat{\xi}_i} &= C - \hat{\lambda}_i - \hat{\mu}_i = 0 \Rightarrow \hat{\lambda}_i + \hat{\mu}_i = C && (5)
\end{aligned}
$$

(c) Derive the dual optimization problem. To derive the dual optimization problem, you have to eliminate $\boldsymbol{\alpha}$, $\boldsymbol{\xi}$, and $\hat{\boldsymbol{\xi}}$ from L using the gradient of L.

Preliminary solution:

$$\begin{aligned}L\left(\boldsymbol{\alpha},\alpha_0,\boldsymbol{\xi},\hat{\boldsymbol{\xi}},\boldsymbol{\lambda},\hat{\boldsymbol{\lambda}},\boldsymbol{\mu},\hat{\boldsymbol{\mu}}\right)=&\frac{1}{2}||\boldsymbol{\alpha}||^2+C\sum_i(\xi_i+\hat{\xi}_i)+\sum_i\left(-\mu_i\xi_i-\hat{\mu}_i\hat{\xi}_i\right)+\\&\sum_i\lambda_i\left(y_i-\boldsymbol{\alpha}^T\boldsymbol{x}_i-\alpha_0-\epsilon-\xi_i\right)+\\&\sum_i\hat{\lambda}_i\left(-y_i+\boldsymbol{\alpha}^T\boldsymbol{x}_i+\alpha_0-\epsilon-\hat{\xi}_i\right)\end{aligned}$$

For $\boldsymbol{\alpha}$, a direct replacement was found in equation (2). ξ_i are eliminated after collecting all the terms in the Lagrangian, because $C\sum_i\xi_i-\sum_i(\lambda_i+\mu_i)\xi_i=0$ using (4). The same is true for $\hat{\xi}_i$. α_0 is eliminated, because $\sum_i\left(\hat{\lambda}_i-\lambda_i\right)\alpha_0=0$ using (3). The resulting dual optimization problem is:

$$\begin{aligned}\tilde{L}\left(\boldsymbol{\lambda},\hat{\boldsymbol{\lambda}}\right)=&\overbrace{\frac{1}{2}\sum_i\sum_j\left(\lambda_i-\hat{\lambda}_i\right)\left(\lambda_j-\hat{\lambda}_j\right)\boldsymbol{x}_i^T\boldsymbol{x}_j}^{\frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{\alpha}}-\\&\sum_i\lambda_i\left[\sum_j\left(\lambda_j-\hat{\lambda}_j\right)\boldsymbol{x}_j^T\right]\boldsymbol{x_i}+\sum_i\hat{\lambda}_i\left[\sum_j\left(\lambda_j-\hat{\lambda}_j\right)\boldsymbol{x}_j^T\right]\boldsymbol{x_i}-\\&\epsilon\sum_i\left(\lambda_i+\hat{\lambda}_i\right)+\sum_i\left(\lambda_i-\hat{\lambda}_i\right)y_i\\=&-\frac{1}{2}\sum_i\sum_j\left(\lambda_i-\hat{\lambda}_i\right)\left(\lambda_j-\hat{\lambda}_j\right)\boldsymbol{x}_i^T\boldsymbol{x}_j-\\&\epsilon\sum_i\left(\lambda_i+\hat{\lambda}_i\right)+\sum_i\left(\lambda_i-\hat{\lambda}_i\right)y_i\end{aligned}\tag{6}$$

The dual optimization is also constrained. λ_i and $\hat{\lambda}_i$ are Lagrangian multipliers and therefore non-negative. In addition, they are constrained by (4) and (5). Equation (3) is also a constraint.

$$0\leq\lambda_i,\ \hat{\lambda}_i\leq C\tag{7}$$

$$\sum_i\left(\lambda_i-\hat{\lambda}_i\right)=0\tag{8}$$

(d) Which property must be fulfilled for *support vectors* in SVR?
Hint: replace $\boldsymbol{\alpha}$ in Equation (1).

Support vectors are the only ones necessary to compute the regression. Therefore, they are included in (1).

$$\begin{aligned}F(\boldsymbol{x})&=\boldsymbol{\alpha}^T\boldsymbol{x}+\alpha_0\\&=\sum_{i=1}^{N}\left(\lambda_i-\hat{\lambda}_i\right)\boldsymbol{x}_i^T\boldsymbol{x}+\alpha_0\end{aligned}$$

For support vectors, the term in the sum is not $0\Rightarrow\lambda_i-\hat{\lambda}_i\neq 0$. Due to the complementary slackness condition, this is the case only for points that are on the boundary or outside of the ϵ-tube. This is a sparse solution, because most points are close to the estimated output of the regression.