**Machine Intelligence 1** **Exercise Sheet 9**
WS 2016/17, Obermayer/Böhmer due: 12.01.2017

# Support Vector Regression

## Exercise T9.1: Regression with SVM (tutorial)

In support vector regression problems, we assume that we are given a training data set

$$\{(\underline{\mathbf{x}}^{(\alpha)}, y_T^{(\alpha)})\}, \quad \alpha \in \{1, \ldots, p\}, \qquad \underline{\mathbf{x}} \in \mathbb{R}^N, \quad y_T \in \mathbb{R},$$

and want to fit the linear regression function

$$\hat{y}(\boldsymbol{x}) \quad = \quad \underline{\mathbf{w}}^\top \underline{\mathbf{x}} + b\,.$$

(a) What is the $\epsilon$-insensitive cost function for regression?

(b) Derive the primal problem of the $\epsilon$-support vector regression ($\epsilon$-SVR).

(c) The optimal $\epsilon$-parameter depends linearly on the noise level in the data, which is unknown. Derive the primal problem for the $\nu$-SVR, which adjusts $\epsilon$ as a primal parameter.

(d) Derive the Lagrangian of the $\nu$-SVR.

**Solution**

(a) The $\epsilon$-insensitive cost function

$$e(\underline{\mathbf{x}}, y_T) \quad = \quad \max\left(0, |\hat{y}_{(\underline{\mathbf{x}})} - y_T| - \epsilon\right).$$

(b) $\epsilon$-SVR has the following primal problem:

$$\min_{\underline{\mathbf{w}}, b, \xi_\alpha, \xi_\alpha^*} \tfrac{1}{2}\|\underline{\mathbf{w}}\|^2 + C\left(\frac{1}{p}\sum_{\alpha=1}^{p}(\xi_\alpha + \xi_\alpha^*)\right)$$

s.t.

$$\begin{aligned}
(\underline{\mathbf{w}}^\top \underline{\mathbf{x}}^{(\alpha)} + b) - y_T^{(\alpha)} \quad &\leq \quad \epsilon + \xi_\alpha \\
y_T^{(\alpha)} - (\underline{\mathbf{w}}^\top \underline{\mathbf{x}}^{(\alpha)} + b) \quad &\leq \quad \epsilon + \xi_\alpha^* \\
\xi_\alpha, \xi_\alpha^* \quad &\geq \quad 0,
\end{aligned}$$

where $\xi_\alpha, \xi_\alpha^*$ are slack variables, $\boldsymbol{w}, b, \xi_\alpha, \xi_\alpha^*$ are called the primal variables and where the constant $C > 0$ determines the trade-off between the 'flatness' of $\hat{y}$ and the amount up to which deviations larger than $\epsilon$ are tolerated.

(c) The optimal $\epsilon$-parameter linearly depends on the noise level in the data, which is unknown. There exists, however, a method to automatically adjust $\epsilon$, and at the same time have a predetermined fraction of support vectors: The so-called **$\boldsymbol{\nu}$-SVR** allows the $\epsilon$-tube width to automatically adapt to the data. In contrast to $\epsilon$-support vector regression, $\epsilon$ becomes a variable of the primal optimization problem, which now includes an extra term which attempts to minimize $\epsilon$. Introducing a fixed parameter $\nu > 0$ (which was shown to provide a lower bound on the fraction of support vectors), the **primal problem** of the $\nu$-SVR is:

$$\min_{\underline{\mathbf{w}}, b, \xi_\alpha, \xi^*_\alpha, \epsilon} \tfrac{1}{2}\|\underline{\mathbf{w}}\|^2 + C\left(\nu\epsilon + \frac{1}{p}\sum_{\alpha=1}^{p}(\xi_\alpha + \xi^*_\alpha)\right)$$

s.t. $\forall \alpha \in \{1, \ldots, p\}$ :

$$\begin{aligned}
(\underline{\mathbf{w}}^\top \underline{\mathbf{x}}^{(\alpha)} + b) - y_T^{(\alpha)} &\leq \epsilon + \xi_\alpha \\
y_T^{(\alpha)} - (\underline{\mathbf{w}}^\top \underline{\mathbf{x}}^{(\alpha)} + b) &\leq \epsilon + \xi^*_\alpha \\
\xi_\alpha, \xi^*_\alpha, \epsilon &\geq 0
\end{aligned}$$

where $\xi_\alpha, \xi^*_\alpha$ are slack variables and $\boldsymbol{w}, b, \xi_\alpha, \xi^*_\alpha, \epsilon$ are the primal variables.

(d) This corresponds to the following Lagrangian

$$\begin{aligned}
L = \tfrac{1}{2}|\underline{\mathbf{w}}|^2 + C\left(\nu\epsilon + \tfrac{1}{p}\sum_{\alpha=1}^{p}(\xi_\alpha + \xi^*_\alpha)\right) - \sum_{\alpha=1}^{p}\lambda_\alpha\{\xi_\alpha + \epsilon + y_T^{(\alpha)} - \underline{\mathbf{w}}^\top \underline{\mathbf{x}}^{(\alpha)} - b\} \\
- \sum_{\alpha=1}^{p}\lambda^*_\alpha\{\xi^*_\alpha + \epsilon - y_T^{(\alpha)} + \underline{\mathbf{w}}^\top \underline{\mathbf{x}}^{(\alpha)} + b\} - \sum_{\alpha=1}^{p}\eta_\alpha\xi_\alpha - \sum_{\alpha=1}^{p}\eta^*_\alpha\xi^*_\alpha - \delta\epsilon
\end{aligned}$$

The Lagrange multipliers $\lambda_\alpha, \lambda^*_\alpha, \eta_\alpha, \eta^*_\alpha, \delta$ must all be $\geq 0$ (since they correspond to inequality constraints) and are called the dual variables.

## Exercise H9.1: The dual problem of the $\nu$-SVR (homework, 5 points)

In this exercise you will derive the dual problem of the $\nu$-SVR.

(a) (2 points) Calculate the derivatives of the Lagrangian with respect to the primal variables.

(b) (3 points) By setting the derivatives from a) to zero and using the results to eliminate the primal variables from the Lagrangian show that the *dual problem* takes the following form:

$$\max_{\lambda_\alpha, \lambda^*_\alpha} -\tfrac{1}{2}\sum_{\alpha,\beta=1}^{p}(\lambda^*_\alpha - \lambda_\alpha)(\lambda^*_\beta - \lambda_\beta)(\underline{\mathbf{x}}^{(\alpha)})^T\underline{\mathbf{x}}^{(\beta)} + \sum_{\alpha=1}^{p}(\lambda^*_\alpha - \lambda_\alpha)y_T^{(\alpha)}$$

s.t. $\forall \alpha \in \{1, \ldots, p\}$ :

$$0 \leq \lambda_\alpha \leq \tfrac{C}{p}\,, \qquad 0 \leq \lambda^*_\alpha \leq \tfrac{C}{p}\,, \qquad \sum_{\alpha=1}^{p}(\lambda_\alpha - \lambda^*_\alpha) = 0\,, \qquad \sum_{\alpha=1}^{p}(\lambda_\alpha + \lambda^*_\alpha) \leq \nu C\,.$$

## Exercise H9.2: Regression with the $\nu$-SVR (homework, 5 points)

In this exercise you will apply $\nu$-SVR from a software package of your choice (e.g. `libsvm`) to the data set used in exercise sheet 5. The training set `TrainingRidge.csv` and the validation set `ValidationRidge-Y.csv` can be found on ISIS. Do **not** center, whiten or expand the data before training.

(a) (2 points) Train the $\nu$-SVR on the training set with the standard parameters of your library ("out of the box"). Plot the output of the validation set as an image plot (where colors represent the output values). Compare the plot with the true labels.

(b) (2 points) Perform a 10-fold nested cross-validation with a $\nu$-SVR and parameters $\nu = 0.5$ and $C \in 2^i,\ i \in \{-2, \ldots, 12\}$. Use a Gaussian RBF kernel with $\gamma \in 2^j,\ j \in \{-12, \ldots, 0\}$. Plot the resulting mean MSE over the folds as an image plot.

(c) (1 point) Find the best parameter combination $C$ and $\gamma$. Train the entire training set with these parameters and plot the output of the validation set as an image plot. Compare the plot with the true labels and the results from (a).

**Total 10 points.**