

Notes on Support Vector Regression

Masaru Okada

November 8, 2025

Abstract

This document provides an overview of Support Vector Regression, a regression technique, as a machine learning tip.

1 Linear Support Vector Regression

Let's start with simple Linear Support Vector Regression.

Linear Support Vector Regression deals only with linear models.

In this context, the linear prediction model is expressed as follows:

$$h_{\theta}(x) = \theta^T x + \theta_0$$

Here,

$$x = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$$

is the feature vector (training data), and

$$\theta = (\theta_0, \theta_1, \theta_2, \dots, \theta_n)^T \in \mathbb{R}^{n+1}$$

are the model's coefficients, known as the bias parameters; these are n -dimensional and $(n + 1)$ -dimensional vectors, respectively.

The solution from Support Vector Regression (the model, i.e., the bias parameters θ) is expressed as the solution to the (constrained) minimization problem for the loss function $J(\theta)$, given by:

$$J(\theta) = C \sum_{i=1}^n \left(\xi^{(i)} + \hat{\xi}^{(i)} \right) + \frac{1}{2} \sum_{j=1}^n \theta_j^2$$

The first term is called the margin violation loss function, and the second term is the L2 regularization loss function. Here, $\xi, \hat{\xi} \in \mathbb{R}^n$ are vectors.

However, the minimization problem for the loss function $J(\theta)$ in Support Vector Regression is subject to the following constraints:

$$y^{(i)} \leq h_{\theta}(x^{(i)}) + \varepsilon + \xi^{(i)}$$

$$y^{(i)} \geq h_{\theta}(x^{(i)}) - \varepsilon - \hat{\xi}^{(i)}$$

$$\xi^{(i)} \geq 0$$

$$\hat{\xi}^{(i)} \geq 0$$

Here, $y \in \mathbb{R}^n$ is the target data.

C and ε are 1-dimensional parameters tuned manually by humans. (Such parameters associated with machine learning models are called hyperparameters.)

In other words, (simply by rearranging the terms,)

$$\xi = y - (h_{\theta}(x) + \varepsilon)$$

$$\hat{\xi} = (h_{\theta}(x) - \varepsilon) - y$$

These represent the differences between the target data and the upper and lower bounds of the prediction model, respectively. The prediction is made by solving the minimization problem for the region defined by this ε width (which is called the 'tube').

1.1 Hyperparameter C

The error in the Support Vector Regression algorithm is represented by ξ and $\hat{\xi}$ on the upper and lower sides of the ε -tube, respectively. For example, the error ξ above the tube is the residual between the target y and the prediction model $h_{\theta}(x) + \varepsilon$. The larger the value of C , the greater the residual ξ when the target y is far from the tube, and thus the greater the influence of the margin violation loss function.

1.2 Hyperparameter ε

Using the method of Lagrange multipliers is effective for the constrained minimization problem of $J(\theta)$. Applying the method of Lagrange multipliers, the function L , known as the Lagrangian, can be written as follows:

$$L(a, \hat{a}) = -\frac{1}{2} \sum_{i,j} (a^{(i)} - \hat{a}^{(i)})(a^{(j)} - \hat{a}^{(j)})K(x^{(i)}, x^{(j)}) - \varepsilon \sum_i (a^{(i)} + \hat{a}^{(i)}) + \varepsilon \sum_i (a^{(i)} - \hat{a}^{(i)})y^{(i)}$$

Here, K is a function called the kernel function. In Linear Support Vector Regression,

$$K(x^{(i)}, x) = ((x^{(i)})^T x)$$

it becomes the inner product of the vectors. $a, \hat{a} \in \mathbb{R}^n$ represent the training data points outside the ε -tube (on the upper and lower sides, respectively), and are called support vectors. The model is obtained by finding the minimum value of this Lagrangian $L(a, \hat{a})$. From this, it is clear that the model is generated only from the training data outside the ε -tube (the support vectors).

From the method of multipliers, θ which minimizes $J(\theta)$ is expressed by the following equation:

$$\theta = \sum_i (a^{(i)} - \hat{a}^{(i)})x^{(i)}$$

Substituting this into the prediction model

$$h_{\theta}(x) = \theta^T x + \theta_0$$

yields:

$$h_{\theta}(x) = \sum_i (a^{(i)} - \hat{a}^{(i)})((x^{(i)})^T x) + \theta_0$$

2 Support Vector Regression

As we have seen, when the kernel function is an inner product, it is called Linear Support Vector Regression. Now, let's consider extending Support Vector Regression to non-linear cases. The inner product in the prediction model equation is rewritten using the kernel function K as follows:

$$h_{\theta}(x) = \sum_i (a^{(i)} - \hat{a}^{(i)})K(x^{(i)}, x) + \theta_0$$

Specific examples of kernel functions include:

Linear kernel:

$$K(x^{(i)}, x) = ((x^{(i)})^T x)$$

Gaussian kernel:

$$K(x^{(i)}, x) = \exp(-\gamma|x^{(i)} - x|^2) \quad \text{where} \quad \gamma = \frac{1}{2\sigma^2}$$

Here, σ^2 is the variance of the Gaussian.

Polynomial and sigmoid kernels are also used.

The functional form of the kernel is also a hyperparameter.

3 Hyperparameters for Time Series Trend Prediction

The above provides a rough explanation of Support Vector Regression.

As an example, the hyperparameters used in this case are as follows:

Table 1: Hyperparameters for time series trend prediction

C	ε	K	γ
100	7	Gaussian kernel	$(\dim x)^{-1}$

Although one would typically scan the parameter space to select the parameter set with the smallest error, this time they were fixed arbitrarily.

For a more serious attempt, prediction accuracy could likely be further improved by applying regression for each day of the week or each week, or by setting flags for weekends and holidays. However, this exploration was intentionally limited.

This is because it felt that a method different from Support Vector Regression might be more suitable for achieving higher accuracy more earnestly.

The next step is to try other methods, such as XGBoost and LightGBM, and then write another article providing a simple explanation, much like this one.