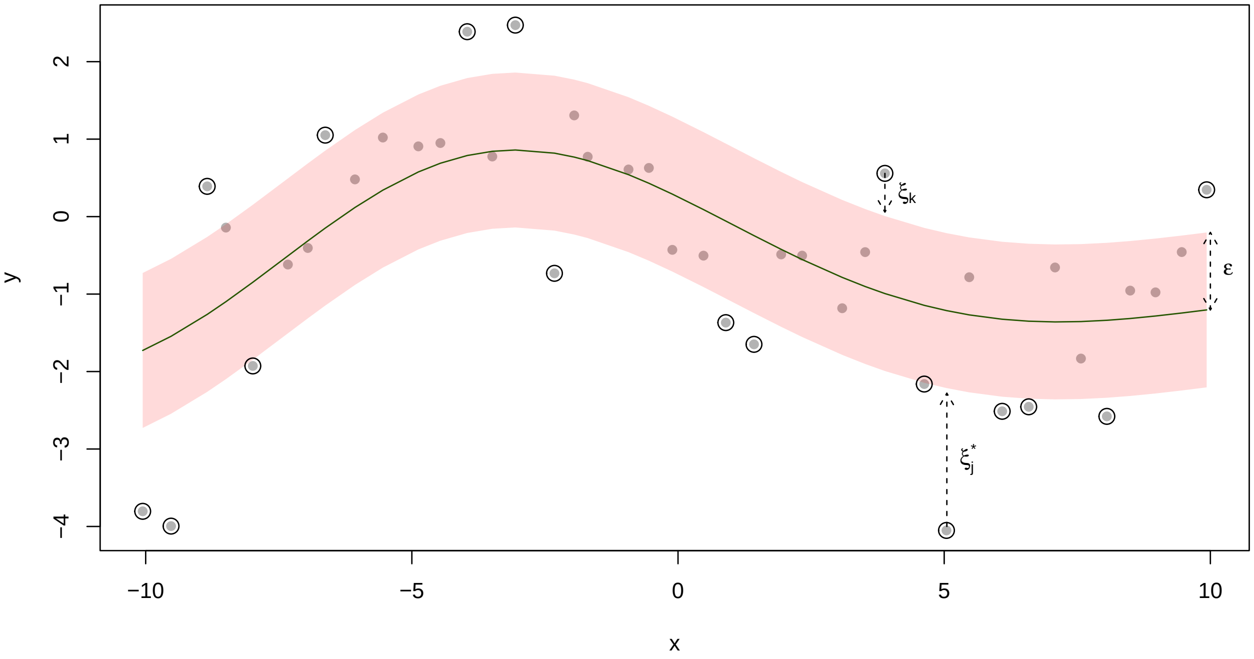
Support Vector Regression - Mathematical Overview

Luisa Kalkert

2025-12-22

Support Vector Machines for Regression



Support Vector Regression expands support Vector Machines from discrete classification to a continuos regression model.

For the regression task, instead of fitting a hyperplane, we are trying to fit a (multidimensional) function $f(x)$ to the data. Similar as in the classification task, we draw a margin around the function. However, now the margin ε is fixed at the start. (Note that ε here, is not the slack variable as in Hasties and Tibshirani, but rather defines the width of the margin. Also, we don't aim to optimize the margin anymore, but instead fix it before the computation starts). We want to find a function f(x) that is as flat as possible, while fitting all points into the required margin. Again, we allow for some slack using slack variables. A wider margin decreases the risk of overfitting, while a smaller margin captures more intricacies of the data.

Mathematically speaking, want to find a function $f(x)$ with $|f(x_i) - y_i| \leq \varepsilon$ for all i. To avoid overfitting, we also introduce the constraint of making the function as flat as possible. "Flatness" in this context is a measure of how sensitive the function is to change. For a linear function $f(x) = x^\top \beta + b$ increasing this flatness can be expressed through minimizing the norm of the function gradient: $\|\beta\|$. Minimizing the expression $\frac{1}{2}\|\beta\|^2$ leads to the same optimum, while allowing for more elegant mathematical solutions, and is therefore used for computation.

Again, we want to allow for some slack, so as in the SVM for classification, we are introducing slack variables ξ_i and ξ_i^*. These slack variables allow for some room for margin violations. The optimization problem we want to solve is to minimize:

$$\frac{1}{2}\|\beta\|^2 + K\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

subject to:

$$y_i - f(x_i) \leq \varepsilon + \xi_i,$$
$$f(x_i) - y_i \leq \varepsilon + \xi_i^*$$

and

$$\xi_i > 0, \xi_i^* > 0 \quad \forall i$$

where K is fixed and β, ξ and ξ^* are variable. The factor K is introduced as cost for regularization purposes. It is choosen a priori to determine how strong the tradeoff between flatness and overfitting should be. A larger K leads to a stronger impact of the term, therefore penalizing stronger margin violations more heavily and resulting in a less flat function curve (higher bias, lower variance). Conversly, a smaller K leads to a flatter curve (lower bias, higher variance).

As in the SVM for classification, this can be generalized to non-linear functions by transforming the features using kernels.

TODO: ADD the references here!!!