Support vector regression

Sparse kernel methods

Francesco Corona

Outline

Support vector regression
Sparse kernel methods

Support vector regression

We now extend SVMs to regression problems while preserving sparseness

In simple linear regression, we minimised a regularised error function

$$\frac{1}{2}\sum_{n=1}^{N}\Big(y(\mathbf{x}_n) - t_n\Big)^2 + \frac{\lambda}{2}||\mathbf{w}||^2 \qquad (1)$$

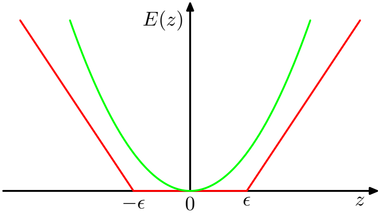
The quadratic error function is replaced by an ε-**insensitive error function**

- which gives zero error if the absolute difference between the prediction $y(\mathbf{x})$ and the target t is less than a $\varepsilon > 0$

Support vector regression (cont.)

An example of an ε-insensitive error function is given by

$$E_\varepsilon\Big(y(\mathbf{x} - t)\Big) = \begin{cases} 0, & \text{if } |y(\mathbf{x}) - t| < \varepsilon \\ |y(\mathbf{x}) - t| - \varepsilon, & \text{otherwise} \end{cases} \quad (2)$$



A linear cost associated with errors outside the insensitive region

Outside the insensitive region, the error increases linearly with distance

Support vector regression (cont.)

We thus minimise a regularised error function given by

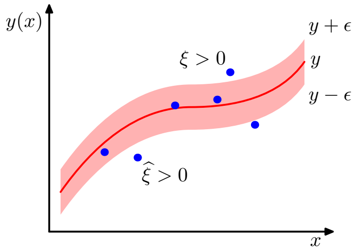
$$C \sum_{n=1}^{N} E_{\varepsilon}\Big(y(\mathbf{x}_n) - t_n\Big) + \frac{1}{2}||\mathbf{w}||^2 \tag{3}$$

C again denotes the (inverse) regularisation parameter

We can re-express the optimisation problem by introducing slack variables

- Two slack variables for each data point $\mathbf{x}_n$, $\xi_n \geq 0$ and $\hat{\xi}_n \geq 0$
- with $\xi_n > 0$, for points with $t_n > y(\mathbf{x}_n) + \varepsilon$
- with $\hat{\xi}_n > 0$, for points with $t_n < y(\mathbf{x}_n) - \varepsilon$

Support vector regression (cont.)

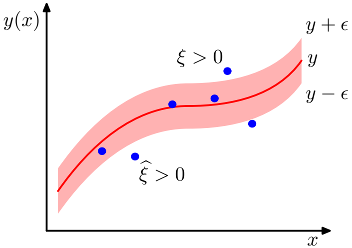


- $\xi_n > 0$ for points with $t_n > y(\mathbf{x}_n) + \varepsilon$
- $\hat{\xi}_n > 0$ for points with $t_n < y(\mathbf{x}_n) - \varepsilon$

Points inside the ε-insensitive region have $\xi_n = \hat{\xi}_n = 0$

Support vector regression (cont.)

Condition for a target point to lie inside the ε-tube is that $y_n - \varepsilon \leq t_n \leq y_n + \varepsilon$



By introducing the slack variables, we allow points to lie outside the tube

As long as the slacks are nonzero and

$$t_n \leq y(\mathbf{x}_n) + \varepsilon + \xi_n \qquad (4)$$

$$t_n \geq y(\mathbf{x}_n) - \varepsilon - \hat{\xi}_n \qquad (5)$$

As a result the error function for support vector regression can be written as

$$C \sum_{n=1}^{N} (\xi_n + \hat{\xi}_n) + \frac{1}{2}||\mathbf{w}||^2 \qquad (6)$$

to be minimised subject to the constraints $\xi_n, \hat{\xi}_n \geq 0$ and Equation 4 and 5

Support vector regression (cont.)

The corresponding Lagrange function with multipliers $a_n, \hat{a}_n \geq 0$ and $\mu_n, \hat{\mu}_n \geq 0$

$$L = C \sum_{n=1}^{N} (\xi_n + \hat{\xi}_n) + \frac{1}{2}||\mathbf{w}||^2 - \sum_{n=1}^{N} (\mu_n \xi_n + \hat{\mu}_n \hat{\xi}_n) - \sum_{n=1}^{N} a_n(\varepsilon + \xi_n + y_n - t_n) - \sum_{n=1}^{N} \hat{a}_n(\varepsilon + \hat{\xi}_n - y_n + t_n) \quad (7)$$

Support vector regression (cont.)

Using the model $y(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b$ and setting derivatives to zero, we get

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{n=1}^{N} (a_n - \hat{a}_n)\phi(\mathbf{x}_n) \tag{8}$$

$$\frac{\partial L}{\partial b} = 0 \quad \Rightarrow \quad \sum_{n=1}^{N} (a_n - \hat{a}_n) = 0 \tag{9}$$

$$\frac{\partial L}{\partial \xi_n} = 0 \quad \Rightarrow \quad a_n + \mu_n = C \tag{10}$$

$$\frac{\partial L}{\partial \hat{\xi}_n} = 0 \quad \Rightarrow \quad \hat{a}_n + \hat{\mu}_n = C \tag{11}$$

Support vector regression (cont.)

Eliminating $\mathbf{w}$, b, ξ_n and $\hat{\xi}_n$ from the Lagrangian and introducing the kernel $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}')$, the dual maximisation problem wrt $\{a_n\}$ and $\{\hat{a}_n\}$ is

$$\tilde{L}(\mathbf{a}, \tilde{\mathbf{a}}) = \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} (a_n - \hat{a}_n)(a_m - \hat{a}_m) k(\mathbf{x}_n, \mathbf{x}_m) - \varepsilon \sum_{n=1}^{N} (a_n + \hat{a}_n) + \sum_{n=1}^{N} (a_n - \hat{a}_n) t_n \quad (12)$$

subject to constraints $a_n, \hat{a}_n \geq 0$ (Lagrange multipliers), the box constraints

$$0 \leq a_n \leq C \quad (13)$$

$$0 \leq \hat{a}_n \leq C \quad (14)$$

(from $\mu_n, \hat{\mu}_n \geq 0$ together with $a_n + \mu_n = C$ and $\hat{a}_n + \hat{\mu}_n = C$), plus

$$\sum_{n=1}^{N} (a_n - \hat{a}_n) = 0$$

Support vector regression (cont.)

Predictions for new input points are obtained in terms of the kernel function, again by substituting $\mathbf{w} = \sum_{n=1}^{N} (a_n - \hat{a}_n)\phi(\mathbf{x}_n)$ into $y(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b$

$$y(\mathbf{x}) = \sum_{n=1}^{N} (a_n - \hat{a}_n) k(\mathbf{x}, \mathbf{x}_n) + b \tag{15}$$

Support vector regression (cont.)

The corresponding Karush-Kuhn-Tucker conditons

$$a_n(\varepsilon + \xi_n + y_n - t_n) = 0 \quad (16)$$
$$\hat{a}_n(\varepsilon + \hat{\xi}_n - y_n + t_n) = 0 \quad (17)$$
$$(C - a_n)\xi_n = 0 \quad (18)$$
$$(C - \hat{a}_n)\hat{\xi}_n = 0 \quad (19)$$

- a_n can be nonzero only if $(\varepsilon + \xi_n + y_n - t_n) = 0$, the point must lie on or above the upper boundary of the ε-tube where $\xi_n \geq 0$
- $\hat{a}_n$ can be nonzero only if $(\varepsilon + \hat{\xi}_n - y_n + t_n) = 0$, the point must lie on or below the lower boundary of the ε-tube where $\hat{\xi}_n \geq 0$

- Constraints $(\varepsilon + \xi_n + y_n - t_n) = 0$ and $(\varepsilon + \hat{\xi}_n - y_n + t_n = 0)$ are incompatible because $\xi_n, \hat{\xi}_n$ are both nonnegative and ε is strictly positive, so for every point $\mathbf{x}_n$ either a_n or $\hat{a}_n$ or both must be zero

Support vector regression (cont.)

The support vectors are points $\mathbf{x}_n$ that contribute to prediction, and thus they must be those for which either $a_n \neq 0$ or $\hat{a}_n \neq 0$

- They lie on the boundary of the ε-tube or outside the tube
- Points inside the tube are those for which $a_n = \hat{a}_n = 0$

Parameter b can be found considering a point for which $0 < a_n < C$ (thus with $\xi_n = 0$) and for which $\varepsilon + y_n - t_n = 0$, by using the model $y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b$

$$\begin{aligned} b &= t_n - \varepsilon - \mathbf{w}^T \phi(\mathbf{x}_n) \\ &= t_n - \varepsilon - \sum_{m=1}^{N} (a_m - \hat{a}_m) k(\mathbf{x}_n, \mathbf{x}_m) \end{aligned} \tag{20}$$

A result can be found by considering a point for which $0 < \hat{a}_n < C$ ($\hat{\xi}_n = 0$)

Support vector regression (cont.)

An application of support vector regression (Gaussian kernels) to the sine data