SVR

SVM Categorization

SVM

- SVM
  - SVC (Support Vector Classification)
  - SVR (Support Vector Regression)

Support Vector Regression (SVR)

Contents

❖ Linear Support Vector Regression
❖ Nonlinear Support Vector Regression
❖ Conclusion

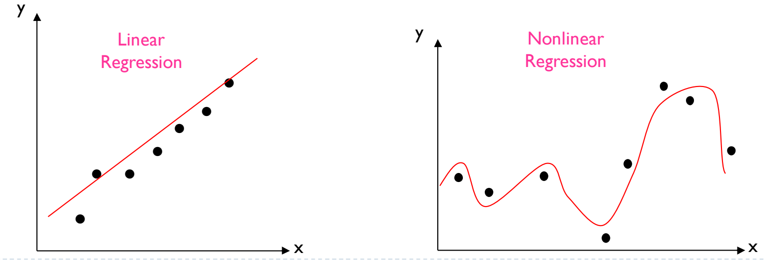
Regression

❖ Given training data:

$$\{ (x_1, y_1), \dots, (x_n, y_n) \}, \quad x_i \in \Re^d, y_i \in \Re$$

❖ Find function:

$$f : \Re^d \to \Re$$



Regression

- "*best function*" or "*best approximation*":

  the expected error on *unseen* data (x_{n+1}, y_{n+1}), … , (x_{n+k}, y_{n+k}) is minimal.

- As unseen data are not available
  - The goal is to minimize *Residual Sum of Squared* (*RSS*)

$$RSS = \sum_{i=1}^{n}\left(y_i^{expected} - y_i\right)^2 = \sum_{i=1}^{n}\left(f(x_i) - y_i\right)^2$$

  - The goal is to minimize *Root Mean Squared Error* (*RMS*)

$$RMS = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^{expected} - y_i\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - y_i\right)^2}$$

- Existing techniques to solve the classification task:
  - Classical (Linear) Regression
  - NN

Linear SVR

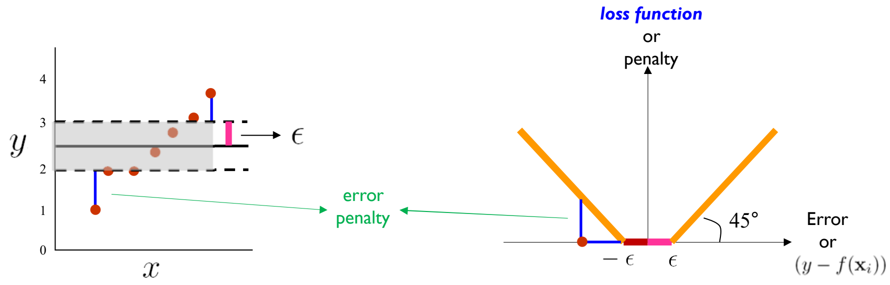
"Lazy case" (underfitting)

"Suspiciously smart case" (overfitting)

"Compromise case", SVR (good generalizability)

Linear SVR

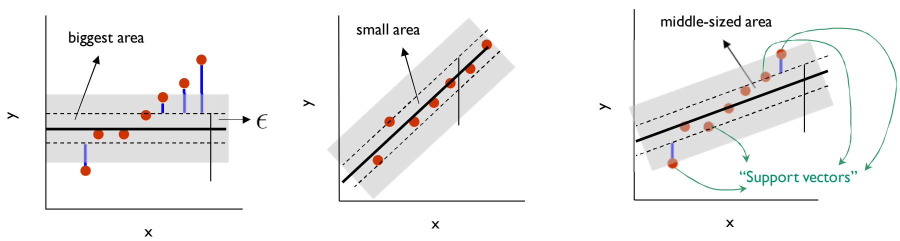
The ε-insensitive ***loss function***



$$y = w_1 x + b$$
$$w_1 = 0$$
$$b = 2.5$$

$$|y_i - f(\mathbf{x}_i)|_\epsilon \equiv \max\{0, |y_i - f(\mathbf{x}_i)| - \epsilon\}$$

Linear SVR



"Lazy case"
(underfitting)

"Suspiciously smart case"
(overfitting)

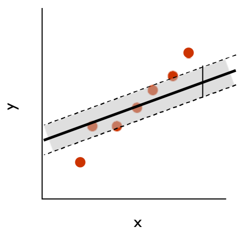
"Compromise case", SVR
(good generalizability)

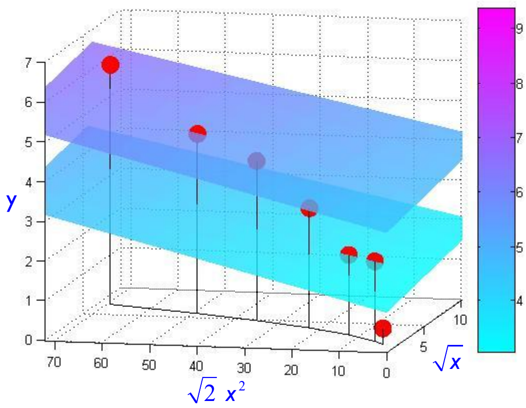
- The thinner the "tube", the more complex the model

Non-linear SVR

❖ Map the data into a *higher*-dimensional space:

$$x \rightarrow \Phi(x) = (\sqrt{x}, \sqrt{2}x^2)$$





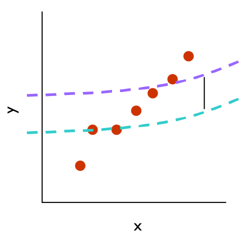
$$y = w_1 x + b$$

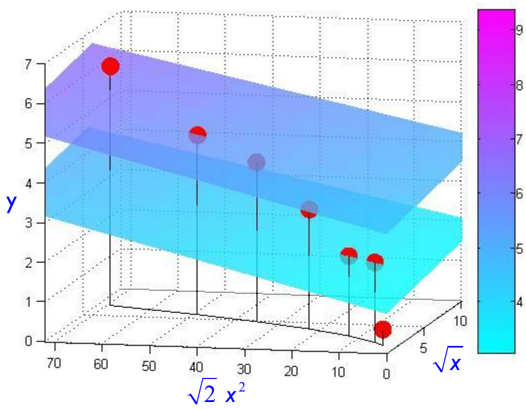
$$y = w_1\sqrt{x} + w_2\sqrt{2}x^2 + b$$

Non-linear SVR

❖ Map the data into a *higher*-dimensional space:

$$x \rightarrow \Phi(x) = (\sqrt{x}, \sqrt{2}x^2)$$





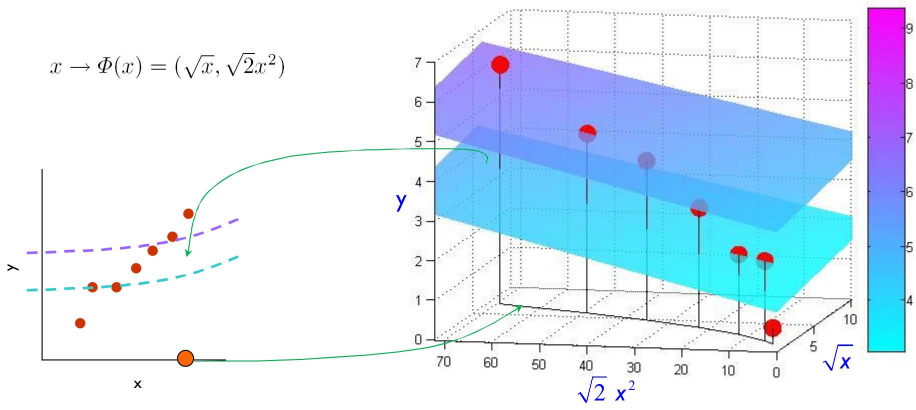
$$y = \mathbf{w}'\Phi(x) + b$$

$$y = w_1\sqrt{x} + w_2\sqrt{2}x^2 + b$$

Non-linear SVR

- Map the data into a *higher*-dimensional space:

$x \rightarrow \Phi(x) = (\sqrt{x}, \sqrt{2}x^2)$

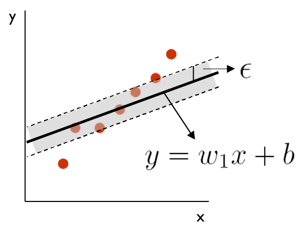


$y = \mathbf{w}'\Phi(x) + b$ $\qquad$ $y = w_1\sqrt{x} + w_2\sqrt{2}x^2 + b \pm \epsilon$

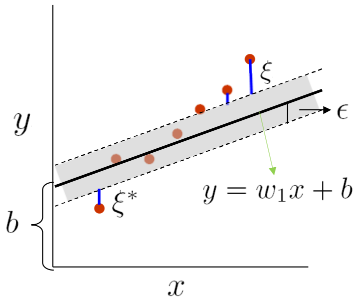
Linear SVR: derivation

- Given training data $\{x_i, y_i\}_{i=1}^{n}$
- Find: w_1 , b

  such that $y = w_1 x + b$ optimally describes the data:



Linear SVR: derivation



$\| w_1 \|$ vs. $\sum_i (\xi_i + \xi_i^*)$

Complexity — Sum of errors

$$\min_{w_1, b, \xi_i, \xi_i^*} \frac{1}{2} w_1^2 + C \sum_i (\xi_i + \xi_i^*)$$

Case I: $w_1 \downarrow$ ⟶ "tube"$\uparrow$ ⟶ complexity $\downarrow$ ⟶ $\sum_i (\xi_i + \xi_i^*) \uparrow$

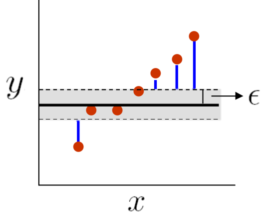
Case II: $w_1 \uparrow$ ⟶ "tube"$\downarrow$ ⟶ complexity $\uparrow$ ⟶ $\sum_i (\xi_i + \xi_i^*) \downarrow$

15

Linear SVR: derivation

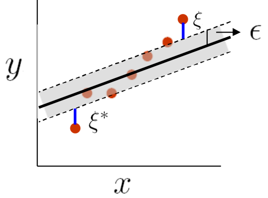
$$\min_{w_1, b, \xi_i, \xi_i^*} \frac{1}{2} w_1^2 + C \sum_i (\xi_i + \xi_i^*)$$

C is small



The role of C

C is big

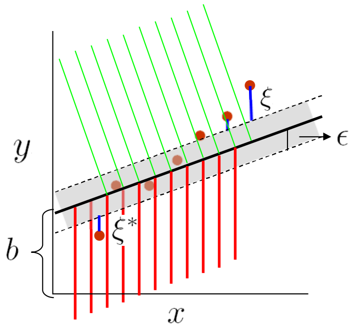


Case I: $w_1 \downarrow$ ⟶ "tube"$\uparrow$ ⟶ complexity $\downarrow$ ⟶ $\sum_i (\xi_i + \xi_i^*) \uparrow$

Case II: $w_1 \uparrow$ ⟶ "tube"$\downarrow$ ⟶ complexity $\uparrow$ ⟶ $\sum_i (\xi_i + \xi_i^*) \downarrow$

16

Linear SVR: derivation



$y = w_1 x + b$

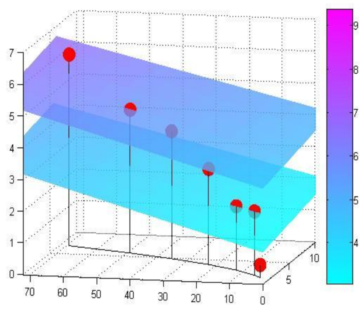
$$\min_{w_1, b, \xi_i, \xi_i^*} \frac{1}{2} w_1^2 + C \sum_i (\xi_i + \xi_i^*)$$

s.t.:

$$y_i - [(w_1 x_{i1}) + b] \leq \epsilon + \xi_i$$
$$[(w_1 x_{i1}) + b] - y_i \leq \epsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \geq 0 \quad i = 1, 2, \ldots, n$$

17

Non-linear SVR: derivation



$y = \mathbf{w}' \Phi(x) + b$

$$\min_{w_1, b, \xi_i, \xi_i^*} \frac{w_1^2 + w_2^2}{2} + C \sum_i (\xi_i + \xi_i^*)$$

s.t.:

$$y_i - (\mathbf{w}' \phi(x_{i1})) - b \leq \epsilon + \xi_i$$
$$(\mathbf{w}' \phi(x_{i1})) + b - y_i \leq \epsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \geq 0 \quad i = 1, 2, \ldots, n$$

18

Non-linear SVR: derivation

$$\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2} \parallel \mathbf{w} \parallel^2 + C \sum_i (\xi_i + \xi_i^*)$$

$$\left.\begin{array}{lll} \text{s.t.:} & y_i - (\mathbf{w}'\phi(\mathbf{x}_i)) - b \leq \epsilon + \xi_i & \\ & (\mathbf{w}'\phi(\mathbf{x}_i))) + b - y_i \leq \epsilon + \xi_i^* & \\ & \xi_i, \xi_i^* \geq 0 \qquad i = 1, 2, \ldots, n & \end{array}\right\}$$ These are $g_i(x) \leq 0$

$$L := \frac{1}{2} \parallel \mathbf{w} \parallel^2 + C \sum_i (\xi_i + \xi_i^*) - \sum_i (\eta_i \xi_i + \eta_i^* \xi_i^*)$$
$$- \sum_i \alpha_i (\epsilon + \xi_i - y_i + \mathbf{w}'\phi(\mathbf{x}_i) + b) - \sum_i \alpha_i^* (\epsilon + \xi_i^* + y_i - \mathbf{w}'\phi(\mathbf{x}_i)) - b)$$

- Since all constraints are inequality constraints, Lagrangian multiplier should satisfy KKT condition, i.e. $\alpha_i, \alpha_i^*, \xi_i, \xi_i^* \geq 0$
- Saddle point of *L* has to be found:
  - **min** with respect to $\mathbf{w}, b, \xi_i, \xi_i^*$
  - max with respect to $\alpha_i, \alpha_i^*, \eta_i, \eta_i^*$

19

Non-linear SVR: derivation

$$L := \frac{1}{2} \parallel \mathbf{w} \parallel^2 + C \sum_i (\xi_i + \xi_i^*) - \sum_i (\eta_i \xi_i + \eta_i^* \xi_i^*)$$
$$- \sum_i \alpha_i (\epsilon + \xi_i - y_i + \mathbf{w}'\phi(\mathbf{x}_i) + b) - \sum_i \alpha_i^* (\epsilon + \xi_i^* + y_i - \mathbf{w}'\phi(\mathbf{x}_i)) - b)$$

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_i (\alpha_i - \alpha_i^*) \phi(\mathbf{x}_i) = 0$$

…

regression hyperplane in ϕ space:
$f(\mathbf{x}) = \mathbf{w}'\phi(\mathbf{x}) + b$

$\Rightarrow$

$$f(\mathbf{x}) = \sum_i (\alpha_i - \alpha_i^*)(\phi(\mathbf{x}_i)'\phi(\mathbf{x})) + b$$
$$f(\mathbf{x}) = \sum_i (\alpha_i - \alpha_i^*) k(\mathbf{x}_i, \mathbf{x}) + b$$

20

Strengths and Weaknesses of SVR

Strengths of SVR:

- No local minima
- It scales relatively well to high dimensional data
- Trade-off between complexity and error can be controlled explicitly via *C* and *epsilon*
- Overfitting is avoided (for any fixed *C* and *epsilon*)
- Robustness of the results
- "Huber (1964) demonstrated that the best cost function over the worst model over any pdf of y given x is the linear cost function. Therefore, if the pdf p(y|x) is unknown the best cost function is the linear penalization over the errors" (Perez-Cruz *et al*., 2003)

Weaknesses of SVR:

- What is the best trade-off parameter *C* and best *epsilon*?
- What is a *good* transformation of the original space