7.1.4 SVMs for regression

Extend SVM to regression problems while preserving the sparseness.

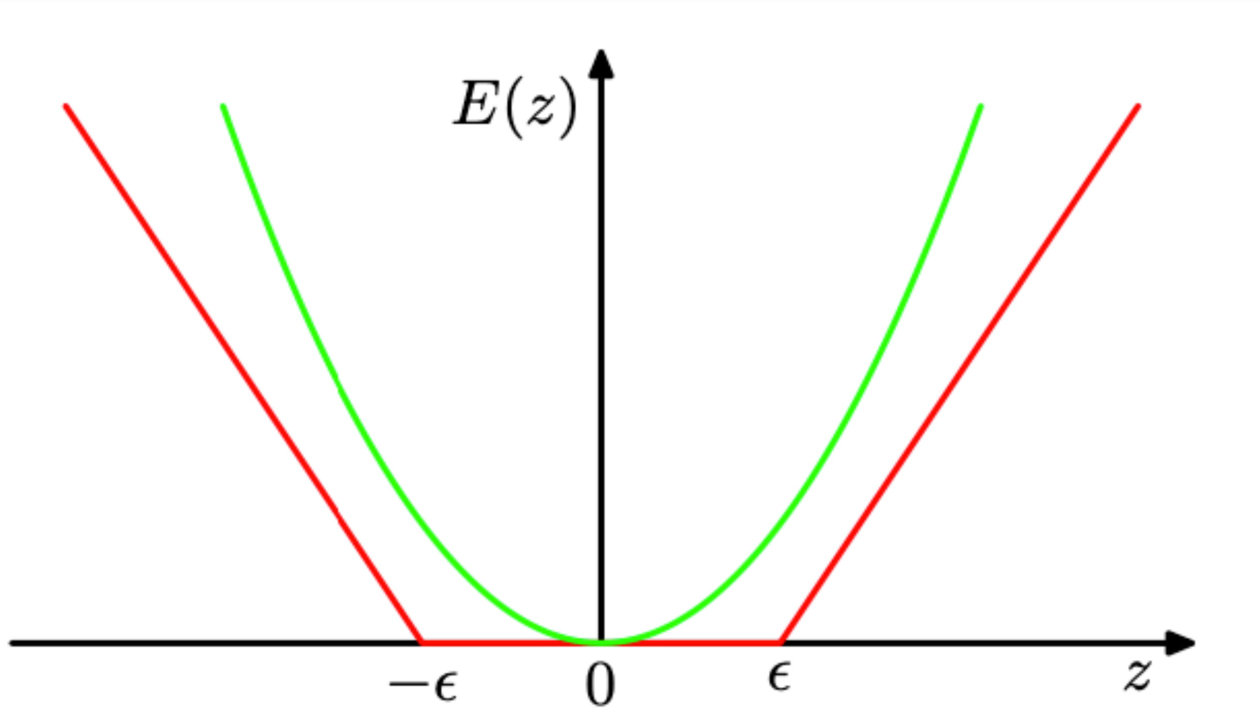
In simple linear regression, minimize a regularized error function : $E = \frac{1}{2}\sum_{n=1}^{N}\{y_n - t_n\}^2 + \frac{\lambda}{2}||w||^2$.

To obtain sparse solutions, the previous error function is replaced by an ϵ-insensitive error function,

Which gives zero if the absolute difference between the prediction $y(x)$ and the target t is less than ϵ.

A simple example is

$$E_\epsilon(y(x) - t) = 0 \qquad if\ |y(x) - t| < \epsilon;$$
$$|y(x) - t| - \epsilon, otherwise$$



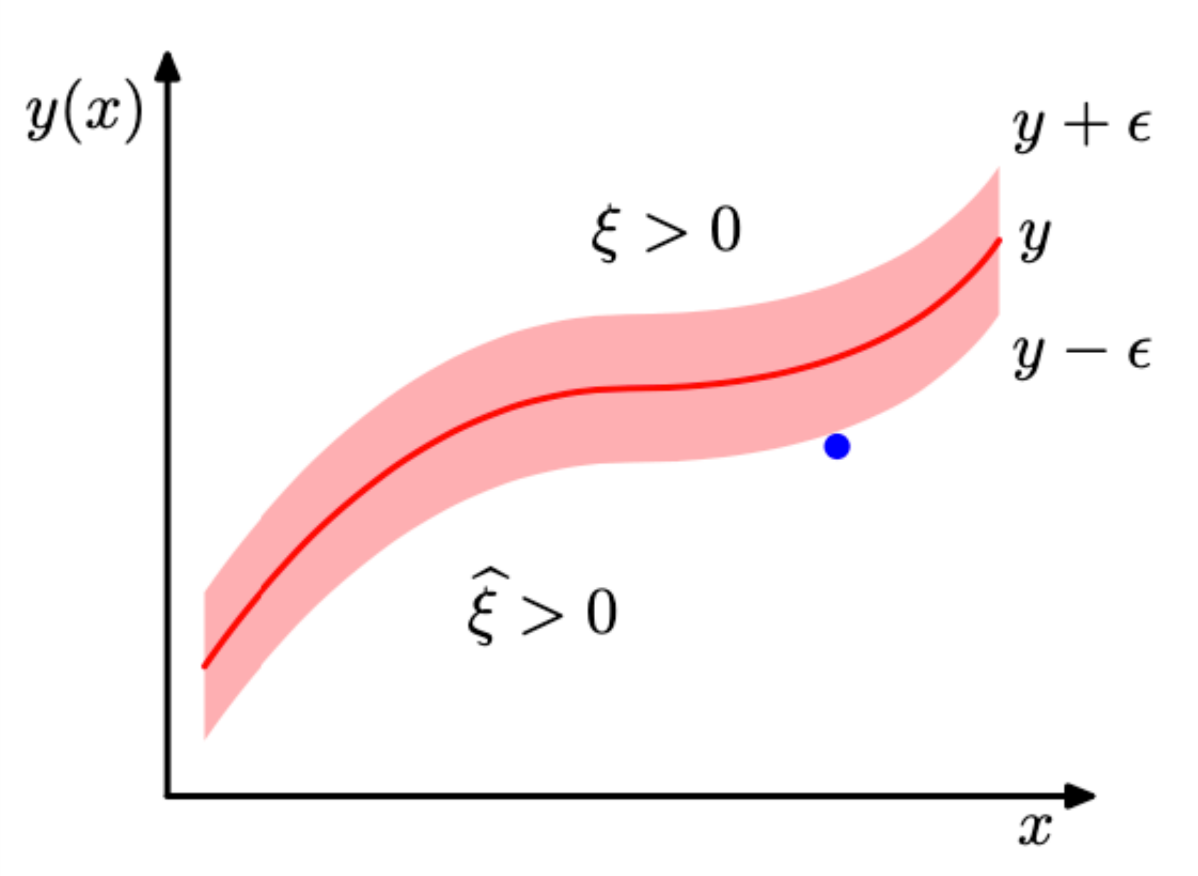
GOAL : Minimize a regularized error function $C\sum_{n=1}^{N} E_\epsilon(y(x_n) - t_n) + \frac{1}{2}||w||^2$, C : 정규화 term

Introduce slack variables again.

For each data point x_n,

we now need two slack variables $\xi_n \geq 0$ and $\hat{\xi}_n \geq 0$,

where $\xi_n > 0$ corresponds to a point for which $t_n > y(x_n) + \epsilon$, and $\hat{\xi}_n > 0$ corresponds to a point for which $t_n < y(x_n) - \epsilon$,



The condition for a target point to lie inside the ϵ-tube is that $y_n - \epsilon \leq t_n \leq y_n + \epsilon$. (Target을 여기에 넣고 싶어)

Introducing the slack variables allows points to lie outside the tube provided the slack variables are nonzero,

Slack variables을 도입함으로서 포인트들이 관 바깥에도 존재할 수 있게 된다. 이 경우 slack variables $\neq 0$ (Inequality)

$$y(x_n) - \epsilon - \hat{\xi}_n \leq t_n \leq y(x_n) + \epsilon + \xi_n$$

SV regression error fnt : $C \sum_{n=1}^{N} (\xi_n + \hat{\xi}_n) + \frac{1}{2} ||w||^2$

subject to the constraint $\xi_n \geq 0 \; and \; \hat{\xi}_n \geq 0$ plus. $y(x_n) - \epsilon - \hat{\xi}_n \leq t_n \leq y(x_n) + \epsilon + \xi_n$

By using Lagrange multipliers $a_n \geq 0$, $\hat{a}_n \geq 0$, $\mu_n \geq 0$, and $\hat{\mu}_n \geq 0$, Lagrangian

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

using $y(x) = w^{\top} \phi(x) + b$, set the derivatives of the Lagrangian with respect to $w, b, \xi_n, and \; \hat{\xi}_n$ to zero

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{n=1}^{N}(a_n - \hat{a}_n)\phi(x_n)$$
$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{n=1}^{N}(a_n - \hat{a}_n) = 0$$
$$\frac{\partial L}{\partial \xi_n} = 0 \Rightarrow a_n + \mu_n = C$$
$$\frac{\partial L}{\partial \hat{\xi}_n} = 0 \Rightarrow \hat{a}_n + \hat{\mu}_n = C$$

Eliminate the corresponding variables from the Lagrangian $\rightarrow$ the dual problem involves maximizing

With respect to $\{a_n\}$ and $\{\hat{a}_n\}$, where we have introduced the kernel $k(x, x') = \phi(x)^{\top}\phi(x')$.

$\tilde{L}(a, \hat{a}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \hat{a}_n)(a_m - \hat{a}_m)k(x_n, x_m) - \epsilon\sum_{n=1}^{N}(a_n + \hat{a}_n) + \sum_{n=1}^{N}(a_n - \hat{a}_n)t_n$

Again, this is a constrained maximization, and to find the constraints that

$a_n \geq 0$ and $\hat{a}_n \geq 0$ are required because Lagrange multipliers. Also, $\mu_n \geq 0\ and\ \hat{\mu}_n \geq 0$ together

$a_n \leq C\ and\ \hat{a}_n \leq C$, we have the box constraints $0 \leq a_n \leq C,\ \ 0 \leq \hat{a}_n \leq C$ with $\sum_{n=1}^{N}(a_n - \hat{a}_n) = 0$

네번째 중 첫번에 y대입, predictions for new inputs can be made $y(x) = \sum_{n=1}^{N}(a_n - \hat{a}_n)k(x, x_n) + b$

which is again expressed in terms of the kernel function.

The corresponding KKT conditions are given by

$$a_n(\epsilon + \xi_n + y_n - t_n) = 0$$
$$\hat{a}_n(\epsilon + \hat{\xi}_n - y_n + t_n) = 0$$
$$(C - a_n)\xi_n = 0$$
$$(C - \hat{a}_n)\hat{\xi}_n = 0$$

From these we can obtain several useful results.

$a_n \neq 0$ only if $\epsilon + \xi_n + y_n - t_n = 0$, such points lies either on($\xi_n = 0$) or above the upper boundary ($\xi_n > 0$).

$\hat{a}_n \neq 0$ implies $\epsilon + \hat{\xi}_n - y_n + t_n = 0$, such points lie either on($\xi_n = 0$) or below the lower boundary($\xi_n < 0$)

Furthermore, the two constraints $\epsilon + \xi_n + y_n - t_n = 0$ and $\epsilon + \hat{\xi}_n - y_n + t_n = 0$ are incompatible,

By adding them together and $\xi_n, \hat{\xi}_n \neq 0$ and $\epsilon > 0 \Rightarrow$ for every x_n, either $a_n \; or \; \hat{a}_n$ (or both) must be 0.

The support vectors are those data points that contribute to predictions given by $y(x)$

i.e., which either $a_n \neq 0 \; or \; \hat{a}_n \neq 0$.

Support Vector lie on the boundary of the ϵ-tube or outside the tube.

All points within the tube have $a_n = \hat{a}_n = 0$.

Now, in $y(x) = \sum_{n=1}^{N} (a_n - \hat{a}_n) k(x, x_n) + b$ last terms are those that involve the support vectors. (다 죽었으니까)

The parameter b can be found by considering a data point for which $0 < a_n < C$,

4개중 3번째 must have $\xi_n = 0$, and 4개중 1번째 must therefore satisfy $\epsilon + y_n - t_n = 0$.

Using $y(x) = w^\top \phi(x) + b$ and solving for b, we obtain $b = t_n - \epsilon - w^\top \phi(x_n) = t_n - \epsilon - \sum_{m-1}^{N} (a_m - \hat{a}_m) k(x_n, x_m)$.

We can obtain an analogous result by considering a point for which $0 < \hat{a}_n < C$.

(In practice, it is better to average over all such estimates of b.)

분류 문제와 마찬가지로 복잡도를 조절하는 매개변수를 좀 더 직관적으로 해석할 수 있는 대안적인 방식이 존재한다.

There is an alternative formulation of the SVM for regression which has a more intuitive interpretation.

In particular, fixed width ϵ is replaced by ν which bounds the fraction of points lying outside the tube.

여기까지 하고 넘어가도 될듯

This involves maximizing $\tilde{L}(a, \hat{a}) = -\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} (a_n - \hat{a}_n)(a_m - \hat{a}_m) k(x_n, x_m) + \sum_{n=1}^{N} (a_n - \hat{a}_n) t_n$

Subject to the constraints

$$0 \leq a_n \leq \frac{C}{N}$$

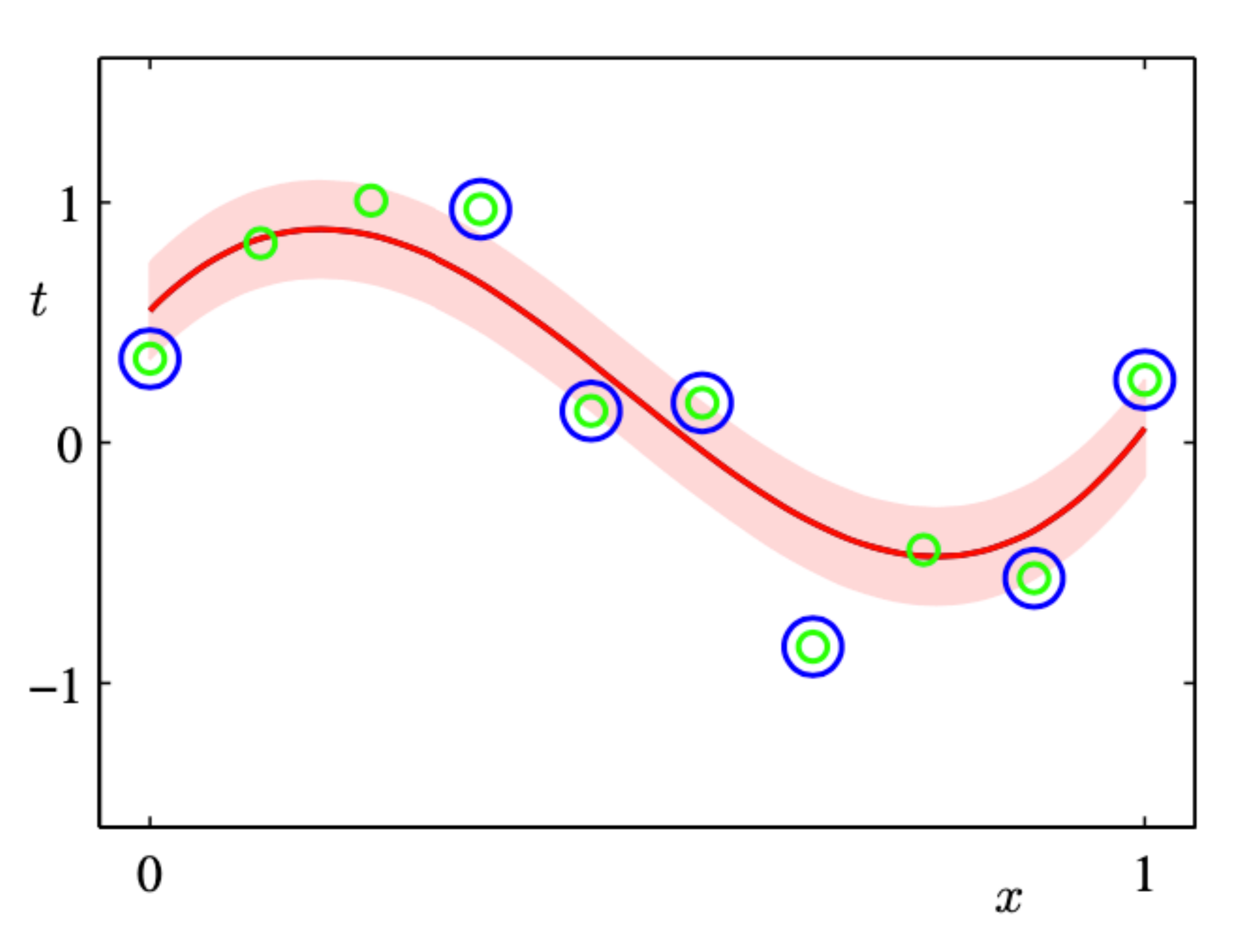
$$0 \leq \hat{a}_n \leq \frac{C}{N}$$

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

$$\sum_{n=1}^{N} (a_n + \hat{a}_n) \leq \nu C.$$

It can be shown that there are at most νN data points falling outside the insensitive tube, while at least νN data points are support vectors and so lie either on the tube or outside it.

The use of a support vector machine to solve a regression problem is illustrated using the sinusoidal data set in Figure.



Here the parameters ν and C have been chosen by hand.

In practice, their values would typically be determined by cross-validation