支持向量回归（SVR）原理介绍

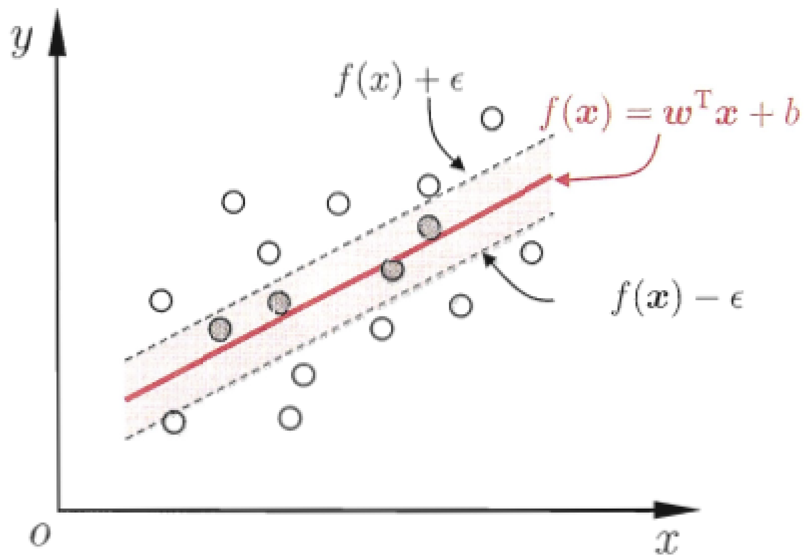
主要参考周志华老师的《机器学习》中对应内容。

考虑回归问题，给定训练样本$D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_m,y_m)\}$，我们希望得到如下回归模型：

$$f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b \tag{1}$$

使得$f(x)$和y尽可能接近。

传统的回归模型通常直接基于模型输出$f(x)$与真实输出y之间的差别来计算loss，当且仅当$f(x)$和y完全相同时，loss才为0。与此不同，SVR假设我们能容忍$f(x)$与y之间最多有ϵ的偏差，即仅当$f(x)$与y之间的差别绝对值大于ϵ时才计算loss。如下图所示，这相当于以$f(x)$为中心，构建了一个宽度为2ϵ的间隔带，若训练样本落入此间隔带，则认为是被预测正确的。



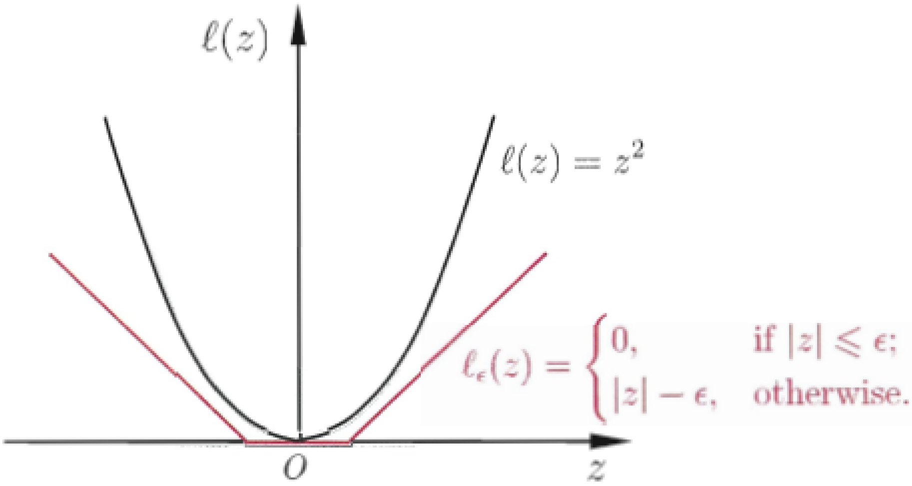
于是，SVR问题形式化为

$$\min_{\boldsymbol{w},b}\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{m}\ell_\epsilon\left(f\left(\boldsymbol{x}_i\right)-y_i\right) \tag{2}$$

其中C为正则化系数，ℓ_ϵ是ϵ-不敏感损失函数：

$$\ell_\epsilon(z)=\begin{cases}0, & \text{if } |z|\leqslant\epsilon\\ |z|-\epsilon, & \text{otherwise}\end{cases} \tag{3}$$

其函数图像为：



样本点落在$\epsilon-$间隔带中的条件为：

$$f(x_i) - \epsilon \leq y_i \leq f(x_i) + \epsilon \tag{4}$$

我们引入松弛变量ξ_i和$\hat{\xi}_i$（间隔带两侧的松弛程度不同），条件（4）变成：

$$f(x_i) - \epsilon - \hat{\xi}_i \leq y_i \leq f(x_i) + \epsilon + \xi_i \tag{5}$$

于是优化问题可以重写为：

$$\begin{aligned} \min_{w, b, \xi_i, \hat{\xi}_i} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{m} \left(\xi_i + \hat{\xi}_i\right) \\ \text{s.t.} \quad & f\left(x_i\right) - y_i \leqslant \epsilon + \xi_i, \\ & y_i - f\left(\boldsymbol{x}_i\right) \leqslant \epsilon + \hat{\xi}_i, \\ & \xi_i \geqslant 0, \hat{\xi}_i \geqslant 0, i = 1, 2, \ldots, m. \end{aligned} \tag{6}$$

引入拉格朗日乘子$\mu_i \geqslant 0, \hat{\mu}_i \geqslant 0, \alpha_i \geqslant 0, \hat{\alpha}_i \geqslant 0$，构造拉格朗日函数为：

$$\begin{aligned} & L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\xi}, \hat{\boldsymbol{\xi}}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) \\ & = \frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{m} \left(\xi_i + \hat{\xi}_i\right) - \sum_{i=1}^{m} \mu_i \xi_i - \sum_{i=1}^{m} \hat{\mu}_i \hat{\xi}_i \\ & + \sum_{i=1}^{m} \alpha_i \left(f\left(\boldsymbol{x}_i\right) - y_i - \epsilon - \xi_i\right) + \sum_{i=1}^{m} \hat{\alpha}_i \left(y_i - f\left(\boldsymbol{x}_i\right) - \boldsymbol{\epsilon} - \hat{\xi}_i\right) \end{aligned} \tag{7}$$

令$L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\xi}, \hat{\boldsymbol{\xi}}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$对$\boldsymbol{w}, b, \xi_i$和$\hat{\xi}_i$的偏导为零，可得：

$$\boldsymbol{w} = \sum_{i=1}^{m} \left(\hat{\alpha}_i - \alpha_i\right) \boldsymbol{x}_i \tag{8}$$

$$0 = \sum_{i=1}^{m} \left(\hat{\alpha}_i - \alpha_i\right) \tag{9}$$

$$C = \alpha_i + \mu_i \tag{10}$$

$$C = \hat{\alpha}_i + \hat{\mu}_i \tag{11}$$

将式（8）~（11）代入式（7），即可得到SVR的对偶问题

$$
\begin{aligned}
\max_{\alpha, \hat{\alpha}} & \sum_{i=1}^{m} y_i\left(\hat{\alpha}_i-\alpha_i\right)-\epsilon\left(\hat{\alpha}_i+\alpha_i\right) \\
& -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m}\left(\hat{\alpha}_i-\alpha_i\right)\left(\hat{\alpha}_j-\alpha_j\right) \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j \\
\text { s.t. } & \sum_{i=1}^{m}\left(\hat{\alpha}_i-\alpha_i\right)=0 \\
& 0 \leqslant \alpha_i, \hat{\alpha}_i \leqslant C
\end{aligned} \tag{12}
$$

上述过程满足KKT条件，有

$$
\left\{\begin{array}{l}
\alpha_i\left(f\left(\boldsymbol{x}_i\right)-y_i-\epsilon-\xi_i\right)=0 \\
\hat{\alpha}_i\left(y_i-f\left(\boldsymbol{x}_i\right)-\epsilon-\hat{\xi}_i\right)=0 \\
\alpha_i \hat{\alpha}_i=0, \xi_i \hat{\xi}_i=0 \\
\left(C-\alpha_i\right) \xi_i=0,\left(C-\hat{\alpha}_i\right) \hat{\xi}_i=0
\end{array}\right. \tag{13}
$$

（这里有一个问题：式（13）中的第三行是如何通过KKT条件得到的？）

可以看出，当且仅当$f\left(x_i\right)-y_i-\epsilon-\xi_i=0$时$\alpha_i$能取非零值，当且仅当$y_i-f\left(\boldsymbol{x}_i\right)-\epsilon-\hat{\xi}_i=0$时$\hat{\alpha}_i$能取非零值。换言之，仅当样本$(\boldsymbol{x}_i, y_i)$不落入$\epsilon-$间隔带中，相应的$\alpha_i$和$\hat{\alpha}_i$才能取非零值。此外，约束$f\left(\boldsymbol{x}_i\right)-y_i-\epsilon-\xi_i=0$和$y_i-f\left(\boldsymbol{x}_i\right)-\epsilon-\hat{\xi}_i=0$不能同时成立，因此$\alpha_i$和$\hat{\alpha}_i$中至少有一个为零。

将式（8）代入式（1），可得SVR的解：

$$
f(\boldsymbol{x})=\sum_{i=1}^{m}\left(\hat{\alpha}_i-\alpha_i\right) \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}+b \tag{14}
$$

能使式（14）中的$\left(\hat{\alpha}_i-\alpha_i\right) \neq 0$的样本即为SVR的支持向量，它们必落在$\epsilon-$间隔带之外。显然，SVR的支持向量仅是训练样本的一部分，即其解仍具有稀疏性。

由KKT条件可以看出，对每个样本$(\boldsymbol{x}_i, y_i)$都有$\left(C-\alpha_i\right) \xi_i=0$且$\alpha_i\left(f\left(x_i\right)-y_i-\epsilon-\xi_i\right)=0$。于是，在得到$\alpha_i$后，若$0<\alpha_i<C$，则必有$\xi_i=0$，进而有

$$
b=y_i+\epsilon-\sum_{i=1}^{m}\left(\hat{\alpha}_i-\alpha_i\right) \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x} \tag{15}
$$

因此，在求解式（12）得到α_i后，理论上来说，可任意选取满足$0<\alpha_i<C$的样本通过式（15）求得b。实践中常采用一种更鲁棒的方法：选取多个（或所有）满足条件$0<\alpha_i<C$的样本求解b后取平均值。

同样，我们可以在SVR中引入核函数，此时式（8）变为：

$$
\boldsymbol{w}=\sum_{i=1}^{m}\left(\hat{\alpha}_i-\alpha_i\right) \phi\left(\boldsymbol{x}_i\right) \tag{16}
$$

而SVR可表示为：

$$
f(\boldsymbol{x})=\sum_{i=1}^{m}\left(\hat{\alpha}_i-\alpha_i\right) \kappa\left(\boldsymbol{x}, \boldsymbol{x}_i\right)+b \tag{17}
$$

其中$\kappa\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)=\phi(\boldsymbol{x}_i)^{\mathrm{T}} \phi\left(\boldsymbol{x}_j\right)$为核函数。