支持向量回归

前言

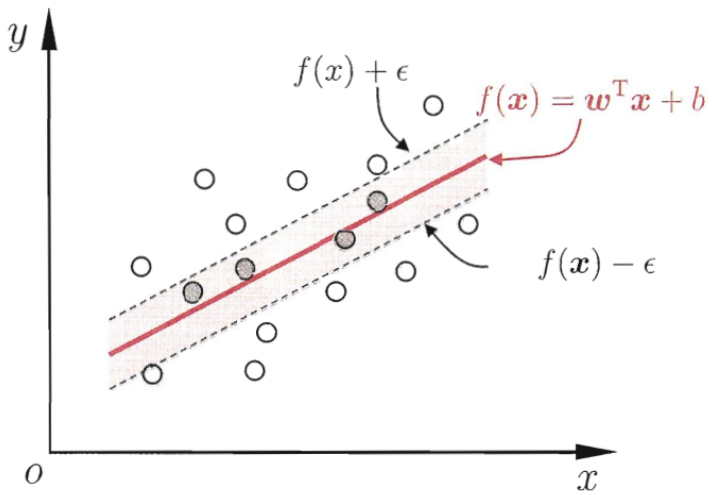
我们上一次已经介绍过了软间隔 + 核函数 + 对偶化下的 SVM，现在我们进而来思考回归问题。给定样本集，我们希望得到一个形如 $f(x)=\mathbf{w}^{\top}\mathbf{x}+b,\quad (6.7)$ 的回归模型，能使得 $f(\mathbf{x})$ 与 y 尽可能接近。也就是我们需要求解参数 $\mathbf{w}$ 和 b 。

注：*软间隔下的SVM主要用于分类任务，而这里所讲的是SVR，属于回归任务，强调误差 $>\epsilon$ 的部分*

支持向量回归 SVR

SVR 对偶化

给定训练样本 $D=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\},y_i\in\mathbb{R}$。对样本 $(\mathbf{x},y)$，传统回归模型一般直接基于模型输出 $f(\mathbf{x})$ 与真实输出 y 之间的差别来计算损失，当且仅当 $f(\mathbf{x})$ 与 y 完全相同时，损失才为零。与此不同，支持向量回归 SVR 假设我们能够容忍 $f(\mathbf{x})$ 与 y 之间最多有 ϵ 的偏差，即仅当 $f(\mathbf{x})$ 与 y 之间的差别绝对值大于 ϵ 时才计算损失。也就是我们在软间隔下所描述的那样“可容忍则不惩罚”。如👇图所示，相当于以 $f(\mathbf{x})$ 为中心，构建了一个宽度为 2ϵ 的隔离带，若训练样本落入此隔离带，则认为是被预测正确的（来自西瓜书）。

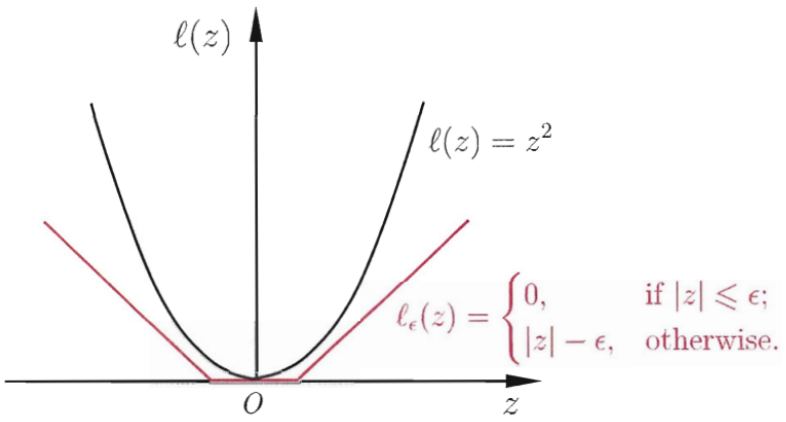


于是，SVR 问题可形式化为：

$$\min_{\mathbf{w},b}\ \frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{m}\ell_{\epsilon}\big(f(x_i)-y_i\big),\quad (6.43)$$

其中，根据我们在 软间隔与正则化 中提到的正则化，C 为正则化常数；ℓ_ϵ 是 $\epsilon-$ 不敏感损失函数(insensitive loss function)，函数和函数图像如下：

$$\ell_\epsilon(z)\begin{cases}0, & if\ |z|\leq\epsilon\\ |z|-\epsilon, & otherwise.\end{cases}, \quad (6.44)$$



在上述(6.43)中，$f(x_i)$ 是我们所熟知的预测值，而 y_i 是真实值，两者之差就是我们所说的损失，再代入 $\ell_\epsilon(\cdot)$ 得到损失函数，求和得到损失期望，也就是经验风险。

所以，我们可以代换，引入松弛变量 ξ_i 和 $\hat{\xi}_i$，可将式(6.43)重写为（注意，损失函数里的符号由负变正）：

$$\begin{aligned}\min_{\mathbf{w},b,\xi_i,\hat{\xi}_i}\ & \frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{m}\ell_\epsilon(\xi_i+\hat{\xi}_i)\\ \text{s.t.}\ & f(x_i)-y_i\leq\epsilon+\xi_i,\\ & y_i-f(x_i)\leq\epsilon+\hat{\xi}_i,\\ & \xi_i\geq 0,\ \hat{\xi}_i\geq,\ i=1,2,...,m.\end{aligned}, \quad (6.45)$$

同样的，直接引入拉格朗日乘子 $\mu_i\geq 0,\hat{\mu}_i\geq 0,\alpha_i\geq 0,\hat{\alpha}_i\geq 0$，由拉格朗日乘子法得到对应的拉格朗日函数：

$$\begin{aligned}&L(w,b,\alpha,\hat{\alpha},\xi,\hat{\xi},\mu,\hat{\mu})\\ &=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}(\xi_i+\hat{\xi}_i)-\sum_{i=1}^{m}\mu_i\xi_i-\sum_{i=1}^{m}\hat{\mu}_i\hat{\xi}_i\\ &+\sum_{i=1}^{m}\alpha_i\big(f(x_i)-y_i-\epsilon-\xi_i\big)+\sum_{i=1}^{m}\hat{\alpha}_i\big(y_i-f(x_i)-\epsilon-\hat{\xi}_i\big).\end{aligned}, \quad (6.46)$$

将 $f(x)=\mathbf{w}^\top\mathbf{x}+b$，(6.7) 代入，对参数求偏导（多少遍了？不用教怎么求偏导吧？）、分别等于零，得到对应的式子：

$$\mathbf{w}=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)x_i, \quad (6.47)$$

$$0 = \sum_{i=1}^{m} (\hat{\alpha}_i - \alpha_i), \quad (6.48)$$

$$C = \alpha_i + \mu_i, \quad (6.49)$$

$$C = \hat{\alpha}_i + \hat{\mu}_i, \quad (6.50)$$

全部代入(6.46)，得到对偶问题：

$$\begin{aligned} \max_{\alpha,\hat{\alpha}} \quad & \sum_{i=1}^{m} y_i(\hat{\alpha}_i - \alpha_i) - \epsilon(\hat{\alpha}_i + \alpha_i) \\ & - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)x_i^T x_j \\ \text{s.t.} \quad & \sum_{i=1}^{m}(\hat{\alpha}_i - \alpha_i) = 0, \\ & 0 \le \alpha_i, \hat{\alpha}_i \le C. \end{aligned}, \quad (6.51)$$

得到 KKT 条件：

$$\begin{cases} \alpha_i(f(\mathbf{x}_i) - y_i - \epsilon - \xi_i) = 0, ① \\ \hat{\alpha}_i(y_i - f(\mathbf{x}_i) - \epsilon - \hat{\xi}_i) = 0, ② \\ \alpha_i\hat{\alpha}_i = 0, ③ \, \xi_i\hat{\xi}_i = 0, ④ \\ (C - \alpha_i)\xi_i = 0, ⑤ \, (C - \hat{\alpha}_i)\hat{\xi}_i = 0. ⑥ \end{cases}, \quad (6.52)$$

- 条件一：α_i 是与约束 $f(\mathbf{x}_i) - y_i \le \epsilon + \xi_i$ 对应的拉格朗日乘子。该条件表明，对于每个样本 i ，要么 $\alpha_i = 0$ ，此时约束 $f(\mathbf{x}_i) - y_i \le \epsilon + \xi_i$ 是“松弛”的（不紧）；要么当 $\alpha_i > 0$ 时，必须有 $f(\mathbf{x}_i) - y_i = \epsilon + \xi_i$ ，即约束是“紧”的。这体现了拉格朗日乘子与约束的互补性。
- 条件二：$\hat{\alpha}_i$ 是与约束 $y_i - f(\mathbf{x}_i) \le \epsilon + \hat{\xi}_i$ 对应的拉格朗日乘子。类似地，对于每个样本 i ，要么 $\hat{\alpha}_i = 0$ ，此时约束 $y_i - f(\mathbf{x}_i) \le \epsilon + \hat{\xi}_i$ 是松弛的；要么当 $\hat{\alpha}_i > 0$ 时，必须有 $y_i - f(\mathbf{x}_i) = \epsilon + \hat{\xi}_i$ ，约束是紧的。
- 条件三：这个条件说明，对于同一个样本 i ， α_i 和 $\hat{\alpha}_i$ 不能同时大于0。从 SVR 的几何意义和优化角度看，这是因为两个约束 $f(\mathbf{x}_i) - y_i \le \epsilon + \xi_i$ 和 $y_i - f(\mathbf{x}_i) \le \epsilon + \hat{\xi}_i$ 不能同时以“紧”的方式被满足（否则会导致矛盾的等式关系），所以对应的拉格朗日乘子 α_i 和 $\hat{\alpha}_i$ 最多只有一个能取正值。
- 条件四：ξ_i 和 $\hat{\xi}_i$ 是松弛变量，分别对应两个约束的松弛程度。该条件表明，对于同一个样本 i ，ξ_i 和 $\hat{\xi}_i$ 不能同时大于 0。这是因为如果 $\xi_i > 0$ ，说明 $f(\mathbf{x}_i) - y_i > \epsilon$ ；如果 $\hat{\xi}_i > 0$ ，说明 $y_i - f(\mathbf{x}_i) > \epsilon$ ，这两个情况不能同时发生，所以 ξ_i 和 $\hat{\xi}_i$ 最多只有一个为正。
- 条件五：C 是 SVR 中的正则化参数，控制对训练误差的惩罚程度。对于每个样本 i ，要么 $\xi_i = 0$ ，此时 α_i 可以取小于 C 的值；要么当 $\xi_i > 0$ 时，必须有 $\alpha_i = C$ 。这体现了松弛变量 ξ_i 与拉格朗日乘子 α_i 的互补性，即当样本在“误差管”（宽度为 2ϵ）外时（$\xi_i > 0$），对应的拉格朗日乘子 α_i 达到上限 C 。

- 条件六：类似条件⑤，对于每个样本 i ，要么 $\hat{\xi}_i = 0$，此时 $\hat{\alpha}_i$ 可以取小于 C 的值；要么当 $\hat{\xi}_i > 0$ 时，必须有 $\hat{\alpha}_i = C$。这体现了松弛变量 $\hat{\xi}_i$ 与拉格朗日乘子 $\hat{\alpha}_i$ 的互补性，即当样本在“误差管”外（$\hat{\xi}_i > 0$）时，对应的拉格朗日乘子 $\hat{\alpha}_i$ 达到上限 C。

求解

将式(6.47)代入(6.7)，得到 SVR 的方程：

$$f(x) = \sum_{i=1}^{m} (\hat{\alpha}_i - \alpha_i) x_i^{\top} \mathbf{x} + b, \quad (6.53)$$

在该式中，能使 $\hat{\alpha}_i - \alpha_i \neq 0$ 的样本即为 SVR 的支持向量，它们必定落在隔离带之外。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。

再看 KKT 条件，每个样本 (x_i, y_i) 都有条件五和条件一。于是，在得到 α_i 后，若 $0 < \alpha_i < C$，则必有 $\xi_i = 0$，进而有：

$$b = y_i + \epsilon - \sum_{i=1}^{m} (\hat{\alpha}_i - \alpha_i) x_i^{\top} \mathbf{x}, \quad (6.54)$$

因此，在由(6.51)解得 α_i 后，理论上来说，可以任意选取满足 $0 < \alpha_i < C$ 的样本通过(6.54)求得 b 。但在实践中，我们一般使用取多个满足 $0 < \alpha_i < C$ 的样本，并求取平均值，这样具有更好的鲁棒性。

核函数

若考虑特征映射形式(6.19)，则相应的，式(6.47)将形如

$$\mathbf{w} = \sum_{i=1}^{m} (\hat{\alpha}_i - \alpha_i) \phi(\mathbf{x}_i), \quad (6.55)$$

将式(6.55)代入(6.19)，则 SVR 可表示为

$$f(\mathbf{x}) = \sum_{i=1}^{m} (\hat{\alpha}_i - \alpha_i) \kappa(\mathbf{x}, \mathbf{x}_i) + b, \quad (6.56)$$