Libsvm 支持的各种 SVM 模型及原理

倪杰

2021 年 4 月 9 日

摘要

libsvm 主要支持 5 种主要的模型，分别是 C-SVC, ν-SVC, one-class SVM, ϵ-SVR, ν-SVR, 本文主要讲解一下这些模型的原理，区别，以及适用的场景。SVC 即 support vector classification, 用于处理分类问题，相应的，SVR 即 support vector regreesion，用于回归。

目录

1 C-SVC

C-SVC 对应的是用于分类的，具有软间隔机制的 SVM。用的损失函数是 hinge-loss.
优化目标是:

$$\min_{\mathbf{w},b}\ \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m} max(0, 1 - y_i(\mathbf{w}^T \cdot \phi(\mathbf{x_i}) + b))$$

引入松弛变量 $\xi \geq 0$, 可得下面的优化目标:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m}\xi_i \qquad (1) \\ subject\ to: \quad & y_i(\mathbf{w}^T \cdot \phi(\mathbf{x_i}) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0.\ i \in [1, m] \end{aligned}$$

通过拉格朗日乘子法可化为以下的二次规划问题:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \qquad (2) \\ subject\ to: \quad & Y^T\alpha = 0 \qquad (3) \\ & 0 \leq \alpha_i \leq C.\ i \in [1, m] \end{aligned}$$

其中 $Q_ij = y_iy_jK(x_i, x_j) = y_iy_j\phi(x_i)\phi(x_j)$.
得到 α 后，$\omega = \sum_{i=1}^{m} y_i\alpha_i\phi(x_i)$
在进行分类时，分类函数为 $sgn(\omega^T x + b) = sgn(\sum_{i=1}^{m} y_i\alpha_iK(x, x_i) + b)$

2 ϵ-SVR

ϵ-SVR 是正则化项采用 ϵ-不敏感损失函数的支持向量回归模型。其中 ϵ-不敏感损失函数可写为:

$$l_\epsilon(z) = \begin{cases} 0 & if|z| \leq \epsilon \\ |z| - \epsilon & otherwise \end{cases}$$

其中 ϵ 作为不计入误差的范围，作为参数需要事先给定

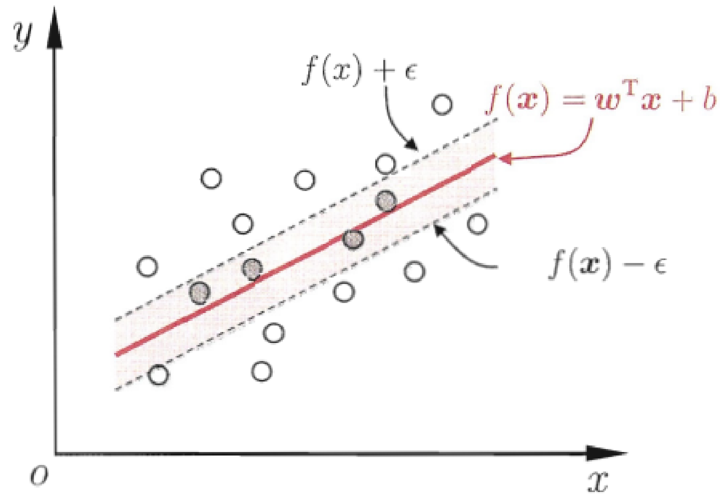


图 [illegible] 支持向量回归示意图. 红色显示出 ϵ-间隔带, 落入其中的样本不计算损失.

优化目标为:

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

引入松弛变量 x_i 和 $\hat{\xi}$, 可得:

$$\begin{aligned}
\min_{\mathbf{w},b,\xi,\hat{\xi}} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m}(\xi_i+\hat{\xi}_i) && (4)\\
subject\ to: \quad & \omega^T\phi(x_i)+b-y_i \leq \epsilon+\xi_i && (5)\\
& y_i-(\omega^T\phi(x_i)+b) \leq \epsilon+\hat{\xi}_i && (6)\\
& \xi_i \geq 0,\ \hat{\xi}_i \geq 0 \quad i\in[1,m] && (7)\\
& && (8)
\end{aligned}$$

通过拉格朗日乘子法可得到 ϵ-SVR 的对偶问题：

$$\begin{aligned}
\min_{\alpha} \quad & \frac{1}{2}(\alpha-\alpha^*)^T Q(\alpha-\alpha^*) + \epsilon\sum_{i=1}^{m}(\alpha_i+\alpha_i^*) + \sum_{i=1}^{m} y_i(\alpha_i-\alpha_i^*) && (9)\\
subject\ to: \quad & e^T(\alpha-\alpha^*)=0 && (10)\\
& 0\leq \alpha_i,\alpha_i^* \leq C.\ i\in[1,m]
\end{aligned}$$

其中 $Q_ij = y_iy_jK(x_i,x_j) = y_iy_j\phi(x_i)\phi(x_j)$. 我们可将 α 与 α^* 进行拼接，可得:

$$
\begin{aligned}
\min_{\alpha,\alpha^*} \quad & \frac{1}{2}[(\alpha^*)^T,\alpha^T]\begin{bmatrix} Q & -Q \\ -Q & Q \end{bmatrix}\begin{bmatrix} \alpha^* \\ \alpha \end{bmatrix} + [\epsilon e^T - Y^T, \epsilon e^T + Y^T]\begin{bmatrix} \alpha^* \\ \alpha \end{bmatrix} && (11) \\
subject\ to: \quad & E^T\begin{bmatrix} \alpha^* \\ \alpha \end{bmatrix} = 0 && (12) \\
& 0 \le \alpha_i, \alpha_i^* \le C.\ i \in [1,m]
\end{aligned}
$$

其中 $E = [e, -e]_{1\times 2m}$

把 $\begin{bmatrix} \alpha^* \\ \alpha \end{bmatrix}$ 看成一个变量，整个优化问题就变成一个大的二次规划问题

得到 α 和 α^* 后，SVR 的解可写为:

$$\sum_{i=1}^{m}(\alpha_i^* - \alpha_i)K(x,x_i) + b$$

3 ν-SVR

ν-SVR 与 ϵ-SVR 不同点在于 ϵ-SVR 的精度 ϵ 是事先给定的，而 ν-SVR 将 $\nu\epsilon$ 直接加到正则化项中，从而实现自动寻找比较小的精度 ϵ. 这里的 ν 取值在 (0,1] 之间，有一些实际的意义，之后会提及

优化目标为:

$$
\begin{aligned}
\min_{\mathbf{w},b,\epsilon} \quad & \frac{1}{2}||\mathbf{w}||^2 + C(\nu\epsilon + \frac{1}{m}\sum_{i=1}^{m}(\xi_i + \hat{\xi}_i)) && (13) \\
subject\ to: \quad & \omega^T\phi(x_i) + b - y_i \le \epsilon + \xi_i && (14) \\
& y_i - (\omega^T\phi(x_i) + b) \le \epsilon + \hat{\xi}_i && (15) \\
& \xi_i \ge 0,\ \hat{\xi}_i \ge 0 \quad i \in [1,m] && (16) \\
& && (17)
\end{aligned}
$$

与 ϵ-SVR 唯一的不同在于最小化项中多了 $\nu\epsilon$, 即 ϵ 不再作为事先给定的参数，而是在求解过程中被自动优化. 另一个用来求均值的 $\frac{1}{m}$ 仅仅影响参数 C 和 ν 的尺度。

对偶问题为:

$$\begin{aligned}
\min_{\alpha} \quad & \frac{1}{2}(\alpha-\alpha^*)^T Q(\alpha-\alpha^*) + Y^T(\alpha-\alpha^*) && (18)\\
subject\ to: \quad & e^T(\alpha-\alpha^*) = 0 && (19)\\
& e^T(\alpha+\alpha^*) \leq C\nu && (20)\\
& 0 \leq \alpha_i, \alpha_i^* \leq \frac{C}{m}.\ i \in [1,m]
\end{aligned}$$

有一个注意的点是优化问题中的不等式约束 $e^T(\alpha+\alpha^*) \leq C\nu$ 可以将不等号改为等号，因为优化目标和另一个等式约束只需要 $\alpha-\alpha^*$, 将 α,α^* 一起放大即可。问题实际上是:

$$\begin{aligned}
\min_{\alpha} \quad & \frac{1}{2}(\alpha-\alpha^*)^T Q(\alpha-\alpha^*) + Y^T(\alpha-\alpha^*) && (21)\\
subject\ to: \quad & e^T(\alpha-\alpha^*) = 0 && (22)\\
& e^T(\alpha+\alpha^*) = C\nu && (23)\\
& 0 \leq \alpha_i, \alpha_i^* \leq \frac{C}{m}.\ i \in [1,m] && (24)
\end{aligned}$$

由 (23),(24) 式可证明，在 $epsilon$>0 即是真的软间隔情形下,
间隔错误样本占比 $\leq v \leq$ 支持向量样本占比
得到 α 和 α^* 后，SVR 的解可写为:

$$\sum_{i=1}^{m}(\alpha_i^*-\alpha_i)K(x,x_i)+b$$

4 ν-SVC

ν-SVC 和 ν-SVR 一样，用参数 ν 代替了参数 ϵ，参数 ν 的意义和 SVR 中基本相同，即支持向量比例的下界，分类错误向量比例的上界。

$$\begin{aligned}
\min_{\mathbf{w},b,\xi,\rho} \quad & \frac{1}{2}||\mathbf{w}||^2 - \nu\rho + \frac{1}{m}\sum_{i=1}^{m}\xi_i && (25)\\
subject\ to: \quad & y_i(\omega^T\phi(x_i)+b) \geq \rho - \xi_i && (26)\\
& \xi_i \geq 0, \quad i \in [1,m] && (27)\\
& && (28)
\end{aligned}$$

这里 ρ 的一种简单理解是当 $\xi=0$, 两个平面间的距离为 $\frac{2\rho}{||\omega||}$, 我自己理解 ρ 越大可以相当于间隔越大，所以前面是负号。
对偶问题为:

$$
\begin{aligned}
\min_{\alpha} \quad & \frac{1}{2}\alpha^T Q\alpha && (29)\\
subject\ to: \quad & e^T\alpha \geq \nu && (30)\\
& Y^T\alpha = 0 && (31)\\
& 0 \leq \alpha_i \leq \frac{1}{m}.\ i \in [1,m]
\end{aligned}
$$

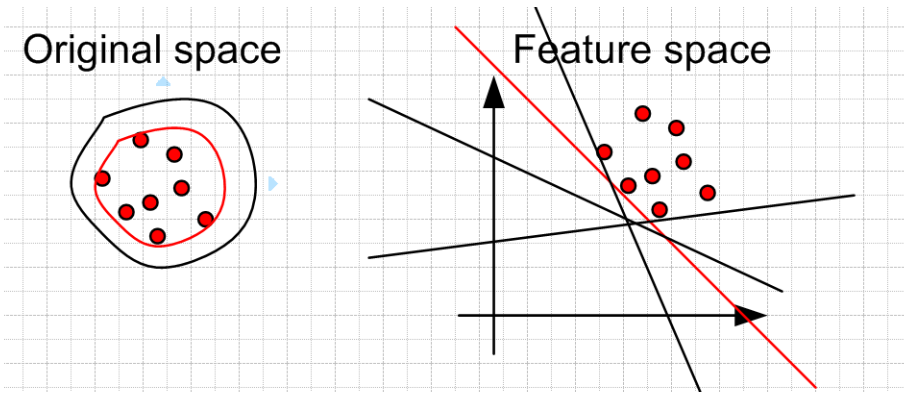
其中 $Q_ij = y_iy_jK(x_i,x_j) = y_iy_j\phi(x_i)\phi(x_j)$. 该问题有可行解当且仅当 $\nu \leq \frac{2min\{\#y_i=+1\#y_i=-1\}}{l}$ 注意到 $^T\alpha \geq \nu$ 实际上可转化为 $e^T\alpha = \nu$，将整个 α 成比例缩小至与 ν 相等即可
对偶问题实际上为:

$$
\begin{aligned}
\min_{\alpha} \quad & \frac{1}{2}\alpha^T Q\alpha && (32)\\
subject\ to: \quad & e^T\alpha = \nu && (33)\\
& Y^T\alpha = 0 && (34)\\
& 0 \leq \alpha_i \leq \frac{1}{m}.\ i \in [1,m]
\end{aligned}
$$

分类函数为 $sgn(\omega^T x + b) = sgn(\sum_{i=1}^{m} y_i\alpha_i K(x,x_i) + b)$

5 one-class SVM

one-class 就是只有一类数据训练，分类时可能出现训练数据以外的类，但是若待分类样本与根据训练出的模型有很大差别，则判断不是该类。一般适用于异常数据很少或没有的情况。
在 SVM 的意义上实际上在核函数对应的映射 ϕ 映射后，用平面尽可能紧地将其与原点分开。



将分出的半空间映射回去实际上就是估计的一个只有正类可能出现的范围 (支集)。软间隔意义下允许部分点被分错。大部分的点应该半平面落在映回去的空间里，可以允许有一部分落在外面。相当于对一个概率分布进行截尾，所以这种 one-class 的估计正类出现范围的方法被称为" 分布估计".
优化目标为:

$$
\begin{aligned}
\min_{\mathbf{w},\rho,\xi} \quad & \frac{1}{2}||\mathbf{w}||^2 - \rho + \frac{1}{\nu m}\sum_{i=1}^{m}\xi_i && (35)\\
subject\ to: \quad & \omega^T\phi(x_i) \geq \rho - \xi_i && (36)\\
& \xi_i \geq 0 \quad i \in [1, m] && (37)\\
& && (38)
\end{aligned}
$$

这边的 ρ 实际上就是其他 SVM 里边的常数项 b。ξ_i 就是部分点落在平面和原点之间的惩罚项.
$f(x) = \omega^T X - \rho$ 到原点的距离为 $\frac{|\rho|}{||\omega||}$。直观上理解，要想尽可能离原点远，$\omega$ 要小,ρ 要大, 所以 ρ 的符号是负的. 对于这个 ν，与之前的 ν-SVR 与 ν-SVM 类似，是错误分类的最大比例和支持向量的最小比例。以概率分布估计的意义来讲相当于置信水平 α
对偶问题为:

$$
\begin{aligned}
\min_{\alpha} \quad & \frac{1}{2}\alpha^T Q\alpha && (39)\\
subject\ to: \quad & e^T\alpha = 1 && (40)\\
& 0 \leq \alpha_i \leq \frac{1}{\nu m}.\ i \in [1, m]
\end{aligned}
$$

其中 $Q_{ij} = y_i y_j K(x_i, x_j) = y_i y_j \phi(x_i)\phi(x_j)$.
分类函数为 $sgn(\sum_{i=1}^{m} \alpha_i K(x, x_i) - \rho)$