第 22 章 支持向量回归

22.1 核岭回归

算法 22.1.1: 核岭回归（Kernel Ridge Regression, KRR）

问题表述 利用表示定理，将 L2 正则岭回归写成仅与系数 β 相关的核化形式：

$$\min_{\beta\in\mathbb{R}^N} \frac{\lambda}{N}\sum_{n=1}^{N}\sum_{m=1}^{N}\beta_n\beta_m K(x_n,x_m)+\frac{1}{N}\sum_{n=1}^{N}\Big(y_n-\sum_{m=1}^{N}\beta_m K(x_n,x_m)\Big)^2$$

矩阵形式 令 $\mathbf{K}\in\mathbb{R}^{N\times N}$ 为核矩阵，$\mathbf{y}=(y_1,\ldots,y_N)^\top$，则目标函数为

$$E_{\text{aug}}(\beta)=\frac{\lambda}{N}\beta^\top\mathbf{K}\beta+\frac{1}{N}\|\mathbf{y}-\mathbf{K}\beta\|^2=\frac{\lambda}{N}\beta^\top\mathbf{K}\beta+\frac{1}{N}(\beta^\top\mathbf{K}^\top\mathbf{K}\beta-2\beta^\top\mathbf{K}^\top\mathbf{y}+\mathbf{y}^\top\mathbf{y}).$$

解析解 对目标函数求梯度并令其为零：

$$\nabla E_{\text{aug}}(\beta)=\frac{2}{N}(\lambda\mathbf{K}^\top\mathbf{I}\beta+\mathbf{K}^\top\mathbf{K}\beta-\mathbf{K}^\top\mathbf{y})=0\Rightarrow(\lambda\mathbf{I}+\mathbf{K})\beta=\mathbf{y}.$$

由于 $\lambda>0$ 且 $\mathbf{K}$ 为半正定矩阵（满足 Mercer 条件），因此矩阵 $\lambda\mathbf{I}+\mathbf{K}$ 可逆，解为：

$$\beta=(\lambda\mathbf{I}+\mathbf{K})^{-1}\mathbf{y}.$$

预测函数 对任意新样本 x，其预测值为：

$$f(x)=\sum_{n=1}^{N}\beta_n K(x_n,x).$$

复杂度

- 训练复杂度：$\mathcal{O}(N^3)$（核矩阵求逆）；
- 预测复杂度：$\mathcal{O}(N)$（对每个测试样本计算 N 个核值）。

表 22.1.1: 线性岭回归 vs. 核岭回归

	线性岭回归	核岭回归
解析解	$(\lambda I+X^\top X)^{-1}X^\top y$	$(\lambda I+K)^{-1}y$
模型能力	受限（线性假设）	灵活（由核函数决定）
训练复杂度	$O(d^3+d^2N)$	$O(N^3)$
预测复杂度	$O(d)$	$O(N)$
适用场景	$N\gg d$ 时高效	大数据时困难

注 线性 vs. 核岭回归：在计算效率与模型灵活性之间权衡。

22.2 支持向量回归原问题

定义 22.2.1 (最小二乘支持向量机（Least-Squares SVM）)

最小二乘支持向量机（LS-SVM）是由 Suykens 和 Vandewalle 提出的一类监督学习方法，可视为传统支持向量机（SVM）的最小二乘版本。其核心特点包括：

- 优化目标：在正则化框架下最小化平方误差与权重范数的加权和

$$\min_{w,b,e} \frac{1}{2}w^\top w + \frac{\gamma}{2}\sum_{n=1}^{N} e_n^2,$$

  其中松弛变量 e_n 满足等式约束

$$y_n\big(w^\top \Phi(x_n) + b\big) = 1 - e_n, \quad n = 1, \dots, N.$$

- 线性系统求解：通过引入拉格朗日乘子 α_n，最终转化为线性 Karush-Kuhn-Tucker（KKT）方程组

$$\begin{bmatrix} 0 & \mathbf{1}^\top \\ \mathbf{1} & K + \gamma^{-1}I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix},$$

  其中 $K_{n,m} = y_n y_m k(x_n, x_m)$ 为核矩阵。
- 与传统 **SVM** 的区别：
  1) 使用等式约束代替不等式约束，避免二次规划；
  2) 解由线性方程组给出，计算速度快；
  3) 失去解的稀疏性（所有样本均贡献）；
  4) 对误差分布假设更敏感，需额外正则化或加权处理。
- 适用场景：回归、分类、函数估计及非线性系统建模，尤其适用于样本量中等但对计算效率要求较高的任务。

♣

表 22.2.1: 软间隔 SVM vs. 最小二乘 SVM (LS-SVM)

	软间隔 SVM	**最小二乘 SVM (LS-SVM)**
本质	最大间隔分类器	核岭回归用于分类
决策边界	稀疏支撑向量决定	与软间隔 SVM 近似
支撑向量	稀疏 α（少量 SV）	稠密 β（所有样本）
预测复杂度	$O(\lvert\mathcal{I}_{\text{SV}}\rvert)$	$O(N)$
实际影响	预测快	预测慢，内存占用大
稀疏性	期望稀疏 α	期望稀疏 β（难点）

注 LS-SVM 与核逻辑回归一样产生稠密系数；标准 SVM 的稀疏性仍是高效预测的关键。

命题 22.2.1 (Tube 回归（ε-不敏感回归）)

误差定义　给定容忍宽度 $\varepsilon > 0$，定义 ε-不敏感误差

$$\mathrm{err}(y,s) = \max\big(0,\ |s-y|-\varepsilon\big) = \begin{cases} 0, & |s-y| \le \varepsilon, \\ |s-y|-\varepsilon, & |s-y| > \varepsilon. \end{cases}$$

几何意义

- 在 $|s-y| \le \varepsilon$ 的“管”内：误差为 0；
- 在管外：误差为到管边界的垂直距离。

与平方误差对比

$$\mathrm{err}_{\mathrm{sq}}(y,s) = (s-y)^2.$$

- 当 $|s-y| \ll \varepsilon$ 时，$\mathrm{err}(y,s) \approx 0$ 而平方误差仍较小；
- 当 $|s-y| \gg \varepsilon$ 时，管误差线性增长，对离群点更不敏感；
- 当 $|s-y|$ 较小时，$\mathrm{err}_{tube} \approx \mathrm{err}_{sq}$，且受异常值的影响较小。

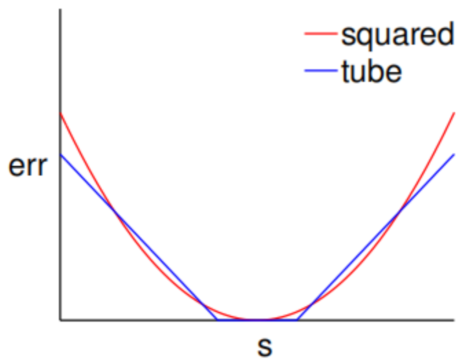

图 22.2.1: 平方误差与管误差对比图

命题 22.2.2 (L2 正则化 Tube 回归与 SVR 的统一视角)

原始无约束问题

$$\min_{w}\ \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\max\big(0,\ |w^\top z_n + b - y_n| - \varepsilon\big)$$

因含不可微的 $\max(\cdot)$ 而难以直接优化。借鉴标准 SVM 的做法，引入松弛变量并转化为二次规划：

支持向量回归（**SVR**）的原始形式

$$\min_{b,w,\xi^\wedge,\xi^\vee}\quad \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n$$
$$\text{s.t.}\quad |y_n - w^\top z_n - b| \le \varepsilon + \xi_n,\quad \xi_n \ge 0.$$

化简支持向量回归（SVR）的原始形式

$$
\begin{aligned}
\min_{b,w,\xi^\wedge,\xi^\vee} \quad & \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}(\xi_n^\wedge + \xi_n^\vee) \\
\text{s.t.} \quad & y_n - w^\top z_n - b \le \varepsilon + \xi_n^\wedge, \\
& w^\top z_n + b - y_n \le \varepsilon + \xi_n^\vee, \\
& \xi_n^\wedge, \xi_n^\vee \ge 0, \quad n = 1, \dots, N.
\end{aligned}
$$

后续步骤

- 利用表示定理将 w 表示为核函数的线性组合，实现核化；
- 构造拉格朗日对偶，导出仅含对偶变量 α_n 的 QP；
- KKT 条件将自动给出稀疏解：仅 $|w^\top z_n + b - y_n| \ge \varepsilon$ 的样本成为支撑向量。

算法 22.2.2: 支持向量回归（Support Vector Regression, SVR）

原始问题（Primal） 引入松弛变量 $\xi_n^+, \xi_n^- \ge 0$ 分别表示上、下管外违规量，SVR 的原始优化问题为

$$
\begin{aligned}
\min_{b,w,\xi^\wedge,\xi^\vee} \quad & \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}(\xi_n^\wedge + \xi_n^-) \\
\text{s.t.} \quad & y_n - w^\top z_n - b \le \varepsilon + \xi_n^\wedge, \\
& w^\top z_n + b - y_n \le \varepsilon + \xi_n^\vee, \\
& \xi_n^\wedge \ge 0, \quad \xi_n^\vee \ge 0, \quad n = 1, \dots, N.
\end{aligned}
$$

参数解释

- $C > 0$：正则化与管外违规的权衡；
- $\varepsilon > 0$：垂直管宽度（容忍误差范围）。

二次规划规模

- 变量：$\tilde{d} + 1 + 2N$（$b, w, \xi^\wedge, \xi^\vee$）；
- 约束：$4N$ 条线性不等式。

下一步 通过拉格朗日对偶消除对维度 $\tilde{d}$ 的依赖，实现核技巧。

例题 22.1 选择题：SVR 松弛变量的计算

考虑使用 $\epsilon = 0.05$ 求解支持向量回归（SVR）。在最优解处，已知 $\mathbf{w}^T\mathbf{z}_1 + b = 1.234$ 和 $y_1 = 1.126$。则松弛变量 $\xi_1^\vee$ 和 $\xi_1^\wedge$ 的值为：

1) $\xi_1^\vee = 0.108, \xi_1^\wedge = 0.000$
2) $\xi_1^\vee = 0.000, \xi_1^\wedge = 0.108$
3) $\xi_1^\vee = 0.058, \xi_1^\wedge = 0.000$
4) $\xi_1^\vee = 0.000, \xi_1^\wedge = 0.058$

解答 正确选项为 $\boxed{3}$。计算偏差 $y_1 - (\mathbf{w}^T\mathbf{z}_1 + b) - \epsilon = -0.158$，由于小于 0，故 $\xi_1^\vee = 0.000$；计算偏差 $(\mathbf{w}^T\mathbf{z}_1 + b) - y_1 - \epsilon = 0.058$，故 $\xi_1^\wedge = 0.058$。 ■

22.3 支持向量回归对偶

命题 22.3.1 (SVR 的对偶推导与解的稀疏性)

1. 拉格朗日函数与对偶变量　引入两组拉格朗日乘子

$$\alpha_n^\wedge \geq 0 \quad \text{对应} \quad y_n - (w^\top z_n + b) \leq \varepsilon + \xi_n^\wedge, \qquad \alpha_n^\vee \geq 0 \quad \text{对应} \quad (w^\top z_n + b) - y_n \leq \varepsilon + \xi_n^\vee.$$

原始拉格朗日函数（已消去松弛变量 $\xi_n^\wedge$ 和 $\xi_n^\vee$[利用求导求极值的方式]）：

$$\mathcal{L} = \frac{1}{2}\|w\|^2 + \sum_{n=1}^{N}\Big[\alpha_n^\wedge\big(-\varepsilon + y_n - w^\top z_n - b\big) + \alpha_n^\vee\big(-\varepsilon - y_n + w^\top z_n + b\big)\Big].$$

2. 对偶问题　通过标准推导得到

$$\min_{\alpha^\wedge,\alpha^\vee} \quad \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(\alpha_n^\wedge - \alpha_n^\vee)(\alpha_m^\wedge - \alpha_m^\vee)\,K(x_n, x_m) + \varepsilon\sum_{n=1}^{N}(\alpha_n^\wedge + \alpha_n^\vee) - \sum_{n=1}^{N} y_n(\alpha_n^\wedge - \alpha_n^\vee)$$

$$\text{s.t.} \quad \sum_{n=1}^{N}(\alpha_n^\wedge - \alpha_n^\vee) = 0, \quad 0 \leq \alpha_n^\wedge, \alpha_n^\vee \leq C, \quad n = 1, \ldots, N.$$

3. 稀疏性　令 $\beta_n = \alpha_n^\wedge - \alpha_n^\vee$，则

$$w = \sum_{n=1}^{N}(\alpha_n^\wedge - \alpha_n^\vee)z_n = \sum_{n=1}^{N}\beta_n z_n \quad \text{且} \quad f(x) = w^\top z + b = \sum_{n=1}^{N}\beta_n K(x_n, x) + b.$$

由互补松弛条件

$$\alpha_n^\wedge(\varepsilon + \xi_n^\wedge - y_n + w^\top z_n + b) = 0, \qquad \alpha_n^\vee(\varepsilon + \xi_n^\vee + y_n - w^\top z_n - b) = 0,$$

可得

- 严格位于管内 $|f(x_n) - y_n| < \varepsilon$ 时 $\xi_n^\wedge = \xi_n^\vee = 0 \Rightarrow \alpha_n^\wedge = \alpha_n^\vee = 0 \Rightarrow \beta_n = 0$;
- 仅位于管边界或管外的样本才成为支撑向量，$\beta_n \neq 0$。

故 SVR 同样具有稀疏解。

例题 22.2 选择题：SVR 对偶问题的变量数量

支持向量回归（SVR）对偶问题的二次规划（QP）变量数量为：

1) $\tilde{d} + 1$
2) $\tilde{d} + 1 + 2N$
3) N
4) $2N$

解答　正确选项为 $\boxed{4}$。SVR 对偶问题的变量为拉格朗日乘数 $\alpha_i^\vee$ 和 $\alpha_i^\wedge$（各 N 个），总数量为 $2N$。■

22.4 核模型小结

表 22.4.1: 线性模型一览（Map of Linear Models）

	优化目标	求解方式	说明
PLA / Pocket	直接最小化 0/1 误差	迭代修正	仅线性可分时有效
线性软间隔 SVM	正则化合页损失	QP（二次规划）	大间隔，稀疏解
线性岭回归	正则化平方损失	解析解	稠密系数，无稀疏性
正则化逻辑回归	正则化交叉熵	GD/SGD 等	概率输出，稠密系数
线性 SVR（Tube 回归）	正则化 ε-不敏感损失	QP	稀疏支撑向量

注 线性软间隔 SVM、线性岭回归、正则化逻辑回归、线性 SVR 均为 LIBLINEAR 库中的常用算法。

表 22.4.2: 线性 / 核模型综合一览（Comprehensive Map of Linear / Kernel Models）

类别	模型	求解方式	稀疏性	使用情况	说明
线性	PLA / Pocket	迭代修正	稀疏（误分修正）	性能较差，较少使用	仅线性可分时有效
	线性软间隔 SVM	QP（二次规划）	稀疏 α	中等规模常用	大间隔，稀疏 α
	线性岭回归	解析解	稠密	小数据常用	稠密系数
	正则化逻辑回归	GD / SGD	稠密	概率输出常用	概率输出，稠密系数
	线性 SVR	QP	稀疏 α	回归任务常用	ε-不敏感损失，稀疏 α
核化	核岭回归	解析解	稠密 β（所有样本）	因稠密 β，大数据少用	非线性映射
	核逻辑回归	无约束优化	稠密 β	因稠密 β，大数据少用	概率输出
	核 SVM	对偶 QP	稀疏 α	LIBSVM 主力模型	非线性边界
	核 SVR	对偶 QP	稀疏 α	LIBSVM 主力模型	非线性回归
	概率 SVM	SVM-变换后逻辑回归	稀疏 α	LIBSVM 扩展模型	概率校准

注 1. 核 SVM、核 SVR、核岭回归、核逻辑回归、概率 SVM 均为 LIBSVM 库中的常用核化算法。
2. 线性模型因性能受限而较少采用；核化稠密模型（核岭回归、核逻辑回归）因预测时需全部核值，计算代价高，大数据场景下使用受限。

22.5 总结

笔记 [支持向量回归]

- 核岭回归：在岭回归上应用表示定理，实现核化。
- 支持向量回归原问题：最小化带正则的 ε-tube 误差。
- 支持向量回归对偶：与 SVM 对偶类似，仍为二次规划（QP）。

- 核模型小结：力量越大，责任越大（使用高表达能力核函数需谨慎）。

笔记 [支持向量回归总结] 支持向量回归通过核技巧与 ε-不敏感损失，在保留线性模型优势的同时实现非线性回归，核岭回归提供核化的解析解但系数稠密，SVR 则借助对偶优化获得稀疏解（仅依赖支持向量），显著提升预测效率。这一框架需平衡模型灵活性（核函数选择）与复杂度（参数 C、ε 调节），避免过拟合。整体而言，支持向量回归是连接线性方法与复杂回归任务的重要工具，其稀疏性与核化能力使其在中等规模数据场景中表现突出。