第 21 章 核逻辑回归

21.1 软间隔 SVM 的正则化视角

命题 21.1.1

1. 软间隔 **SVM** 的无约束优化形式

$$\min_{b,w} \underbrace{\frac{1}{2}\|w\|^2}_{\text{L2 正则项}} + C \underbrace{\sum_{n=1}^{N} \max\big(1 - y_n(w^\top z_n + b),\ 0\big)}_{\text{合页损失（hinge loss）}}$$

该目标函数由 L2 正则项与合页损失项组成，可简记为：

$$\min_{b,w} \frac{1}{2}\|w\|^2 + C\,\mathcal{L}_{\text{hinge}}(b, w),$$

即最小化带 L2 正则的 hinge loss，常被称为软间隔支持向量机的无约束形式。

2. 与经典正则化框架的对应关系

模型	正则方式	误差控制或损失
硬间隔 SVM	约束 $\|w\|^2 \le \text{const.}$	$E_{\text{in}} = 0$（完全可分）
软间隔 SVM	$\frac{1}{2}\|w\|^2$（L2 正则）	$\sum \max(0, 1 - y_n f(x_n))$（hinge loss）
L2 正则化模型	$\frac{\lambda}{N}\|w\|^2$	$\frac{1}{N}\sum \widehat{\text{err}}$（一般损失）

3. 统一视角下的理解

- 大间隔 $\Longleftrightarrow$ 较少可分超平面 $\Longleftrightarrow$ $\|w\|$ 小 $\Longleftrightarrow$ 强 L2 正则化；
- 软间隔 $\Longleftrightarrow$ 使用 hinge loss 对误差进行容忍，允许一定程度的误分；
- 正则强度由超参数 C 控制，等价地，也可使用 $\lambda = \frac{1}{2C}$：

  较大$C \Rightarrow$ 小正则，间隔小，拟合更强　较小$C \Rightarrow$ 大正则，间隔大，泛化更强

4. 拓展意义与建模连接

- 从正则化的角度理解 SVM，有助于与其他学习模型（如逻辑回归、岭回归）建立统一框架；
- 该视角也便于推广到核方法、稀疏学习、随机特征等更复杂场景；
- 尽管该无约束形式易于理解，但由于 $\max(\cdot, 0)$ 不可导，故优化上比标准 SVM 更难处理，不能直接使用核技巧。

21.2 SVM vs 逻辑回归

命题 21.2.1 (SVM 与逻辑回归的算法误差及关联)

1. 线性评分与误差度量

定义线性评分

$$s = w^\top z_n + b.$$

- 0/1 误差：$\mathrm{err}_{0/1}(s,y) = \mathbb{I}[ys \le 0]$。
- SVM（合页误差）：$\mathrm{err}_{\mathrm{SVM}}(s,y) = \max(1-ys,\,0)$，为 $\mathrm{err}_{0/1}$ 的凸上界，常称 **hinge loss**。
- 逻辑回归（缩放交叉熵）：$\mathrm{err}_{\mathrm{SCE}}(s,y) = \log_2(1+\exp(-ys))$，亦为 $\mathrm{err}_{0/1}$ 的凸上界。

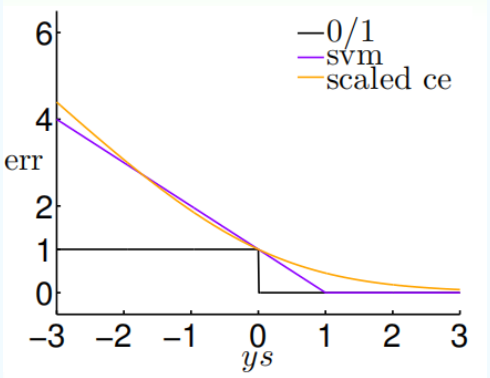

图 21.2.1: 二分类问题中不同误差度量对比图

2. SVM 与逻辑回归的近似关系

当 ys 远离 0 时，

$$\mathrm{err}_{\mathrm{SVM}}(s,y) \approx \ln 2 \cdot \mathrm{err}_{\mathrm{SCE}}(s,y),$$

因此 SVM 可视为 带 **L2** 正则的逻辑回归的近似。

表 21.2.1: 二分类线性模型对比

	PLA	软间隔 SVM	正则化逻辑回归
优化目标	最小化 0/1 误差	最小化正则化合页误差	最小化正则化交叉熵
	$\min_{w,b} \sum_n \mathbb{I}[y_n(w^\top z_n + b) \le 0]$	$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_n \max(0, 1-y_n f(x_n))$	$\min_{w,b} \frac{\lambda}{N}\|w\|^2 + \frac{1}{N}\sum_n \log(1+e^{-y_n f(x_n)})$
求解方法	逐点修正	二次规划 (QP)	梯度下降 / SGD 等
优点	线性可分时高效	优化简单，理论保证	优化简单，正则化自然
缺点	仅线性可分时收敛；否则需 pocket	合页误差对极负样本宽松	交叉熵对极负样本宽松

注 正则化逻辑回归 ≈ SVM；反之，SVM 也可视为带合页误差的逻辑回归近似。

21.3 软二分类的 SVM

命题 21.3.1 (两层学习：SVM 变换后再用逻辑回归微调)

定义最终模型

$$g(x) = \sigma\big(A\,(w_{\text{SVM}}^{\top}\Phi(x) + b_{\text{SVM}}) + B\big), \quad \sigma(z) = \frac{1}{1+e^{-z}},$$

其中

- **SVM** 变体：固定超平面方向与位置 $(w_{\text{SVM}}, b_{\text{SVM}})$，由核方法给出；
- 逻辑回归变体：仅用两个标量参数 $A > 0, B \approx 0$ 对超平面进行 缩放与平移，使其更接近最大似然。

第二层的逻辑回归优化

将 SVM 输出

$$\psi_{\text{SVM}}(x_n) = w_{\text{SVM}}^{\top}\Phi(x_n) + b_{\text{SVM}}$$

作为新特征，求解

$$\min_{A,B} \frac{1}{N}\sum_{n=1}^{N} \log\Big(1 + \exp\big[-y_n\big(A\,\psi_{\text{SVM}}(x_n) + B\big)\big]\Big).$$

总结　此“两层学习”策略把 SVM 的非线性变换结果当作逻辑回归的新输入，实现 **SVM** 变换后的逻辑回归微调，兼具 SVM 的结构性与逻辑回归的概率校准优势。

算法 21.3.1: Platt 概率 SVM（Platt’s Probabilistic SVM）

输入: 训练数据 $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$；核函数 $K(\cdot,\cdot)$；惩罚参数 C

输出: 概率输出软二分类器 $g(x) = \sigma\big(A \cdot f_{\text{SVM}}(x) + B\big)$，其中 $\sigma(z) = \dfrac{1}{1+e^{-z}}$

步骤 1：训练核软间隔 SVM

$$f_{\text{SVM}}(x) = w_{\text{SVM}}^{\top}\Phi(x) + b_{\text{SVM}} = \sum_{n\in\mathcal{I}_{\text{SV}}} \alpha_n y_n K(x_n, x) + b_{\text{SVM}}.$$

步骤 2：构造新特征

$$z_n = f_{\text{SVM}}(x_n), \quad n = 1, \ldots, N.$$

步骤 3：逻辑回归微调

在新特征空间 $\{(z_n, y_n)\}_{n=1}^{N}$ 上求解带特殊正则化的逻辑回归

$$\min_{A,B} \sum_{n=1}^{N} \log\Big(1 + \exp\big[-y_n(Az_n + B)\big]\Big) + \text{正则项}.$$

（实际实现中采用梯度下降、SGD 或其他高效方法，因仅含两个变量）

输出

$$g(x) = \sigma\big(A\,f_{\text{SVM}}(x) + B\big), \quad \text{提供概率而非硬标签。}$$

备注

- 概率边界可能与纯 SVM 边界不同，主要受偏置 B 影响；
- 若直接在 z 空间做精确逻辑回归，则等价于核化逻辑回归（下一节给出）。

21.4 核逻辑回归

命题 21.4.1 (核技巧的核心：统一表示定理)

设最优权重向量可写成训练样本的线性组合

$$w^* = \sum_{n=1}^{N} \beta_n z_n, \qquad z_n = \Phi(x_n).$$

则对任意输入 x，

$$w^{*\top} z = \sum_{n=1}^{N} \beta_n z_n^\top z = \sum_{n=1}^{N} \beta_n K(x_n, x),$$

其中 $K(x_n, x) = \Phi(x_n)^\top \Phi(x)$ 为核函数。因此无需显式计算高维特征 $\Phi(x)$，只需核值即可预测。

三种经典算法的统一形式

$$w = \sum_{n=1}^{N} \alpha_n y_n z_n$$

系数来源：

- SVM：α_n 来自对偶最优解；
- PLA：α_n 为样本被误分并用于修正的次数；
- 逻辑回归（SGD）：α_n 为历次梯度更新中该样本的贡献累积。

结论　核技巧的本质是权重由数据线性表示，使所有线性模型在高维特征空间中统一且高效。

定理 21.4.1 (表示定理 Representer Theorem)

对任意带 L2 正则的线性模型

$$\min_w \ \frac{\lambda}{N}\|w\|^2 + \frac{1}{N}\sum_{n=1}^{N} \mathrm{err}(y_n, w^\top z_n),$$

其最优解 w^* 必落在训练样本张成的子空间中，即存在 $\{\beta_n\}_{n=1}^N \subset \mathbb{R}$，使得

$$w^* = \sum_{n=1}^{N} \alpha_n z_n, \qquad z_n = \Phi(x_n).$$

证明要点

设 $w^* = w_\parallel + w_\perp$，其中 $w_\parallel \in \mathrm{span}\{z_n\}$，$w_\perp \perp \mathrm{span}\{z_n\}$。

- 误差项只与 $w_\parallel$ 有关：$\mathrm{err}(y_n, w^{*\top} z_n) = \mathrm{err}(y_n, w_\parallel^\top z_n)$；
- 正则项满足 $\|w^*\|^2 = \|w_\parallel\|^2 + \|w_\perp\|^2 \geq \|w_\parallel\|^2$；
- 若 $w_\perp \neq 0$，则 $w_\parallel$ 更优，矛盾！故 $w_\perp = 0$。

推论　任意 L2 正则线性模型均可核化：

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

算法 21.4.2: 核逻辑回归（Kernel Logistic Regression）

目标函数 利用表示定理，将 L2 正则化逻辑回归写成仅与系数 β_n 相关的核化形式：

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

算法步骤

1. 构造核矩阵 $\mathbf{K}\in\mathbb{R}^{N\times N}$，其中 $K_{n,m}=K(x_n,x_m)$；
2. 任选优化算法（GD、SGD、LBFGS 等）对无约束目标

$$J(\beta)=\frac{\lambda}{N}\beta^\top\mathbf{K}\beta+\frac{1}{N}\sum_{n=1}^{N}\log\Big(1+\exp\big[-y_n(\mathbf{K}\beta)_n\big]\Big)$$

求解最优 β^*；

3. 预测函数

$$g(x)=\sigma\Big(\sum_{n=1}^{N}\beta_n^* K(x_n,x)\Big),\quad \sigma(z)=\frac{1}{1+e^{-z}}.$$

复杂度

- 训练：$\mathcal{O}(N^3)$（核矩阵）+ 迭代优化开销；
- 预测：$\mathcal{O}(N)$（全部样本）或稀疏近似。

命题 21.4.2 (核逻辑回归的另一种视角：核特征空间中的线性模型)

核逻辑回归（KLR）可等价地视为在两种不同空间中的线性模型：

1. 核特征空间中的线性模型

目标函数

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

等价于在变换后的特征空间

$$\big(K(x_1,x_n),\,K(x_2,x_n),\,\ldots,\,K(x_N,x_n)\big)\in\mathbb{R}^N$$

上学习线性模型

$$f(x)=\beta^\top\Phi_{\mathbf{K}}(x),$$

其中 $\Phi_{\mathbf{K}}(x)=\big(K(x_1,x),\ldots,K(x_N,x)\big)^\top$，并采用核正则化项 $\frac{\lambda}{N}\beta^\top\mathbf{K}\beta$。

2. 嵌入核的原始空间中的线性模型

同样可视为在无限维或高维嵌入特征空间 $\Phi(x)$ 中学习线性模型

$$f(x)=w^\top\Phi(x),$$

并施加 L2 正则化 $\frac{\lambda}{N}\|w\|^2$。

与 **SVM** 的对比

- SVM：α_n 稀疏，仅支撑向量非零；
- KLR：β_n 通常非零，整体稠密，预测需全部核值。

♠

例题 21.1 选择题：KLR 线性模型的空间维度

当将核逻辑回归（KLR）视为具有嵌入核变换和核正则化的 β 线性模型时，该线性模型操作的 $\mathcal{Z}$ 空间

的维度是：

1) d（原始空间 $\mathcal{X}$ 的维度）
2) N（训练样本数）
3) $\tilde{d}$（核函数隐式定义的特征变换维度）
4) λ（正则化参数）

解答 正确选项为 $\boxed{2}$。对于任意样本 $\mathbf{x}$，核变换后的数据为 $(K(\mathbf{x}_1,\mathbf{x}),K(\mathbf{x}_2,\mathbf{x}),\ldots,K(\mathbf{x}_N,\mathbf{x}))$，其维度为 N。 ■

21.5 总结

笔记 [核逻辑回归]

- 软间隔 SVM 的正则化视角：带 L2 正则项的合页误差（hinge loss）。
- SVM vs 逻辑回归：SVM 近似于带 L2 正则的逻辑回归。
- 软二分类的 SVM：常用“两层学习”流程。
- 核逻辑回归：利用表示定理，将 L2 正则化逻辑回归推广到核空间。