中国人民大學

RENMIN UNIVERSITY OF CHINA

XXXX

XXXXX

学院 _____

专业 _____

学号 _____

姓名 _____

2025 年 8 月 29 日

摘　要

XXXXX

关键词：XXX

目录

1　四级抽象与相互作用

1.1　总体框架

1. **Application / Data（应用与数据）**。首先判断数据是否带标签。若有标签，进一步区分分类（离散标签）与回归（连续标签）；若无标签，则多为聚类（相似性）或降维（定位）。
2. **Model（模型 / 假设空间）**。允许哪些假设：线性/多项式/逻辑回归/神经网络/最近邻/决策树等。模型容量影响过拟合与欠拟合，也影响可解释性与推断。
3. **Optimization Problem（优化问题）**。把任务转写成“变量 + 目标函数 + 约束”的形式。常见有：无约束最优化、凸规划、最小二乘、PCA 等。
4. **Optimization Algorithm（优化算法）**。选择实际的求解器，如梯度下降、单纯形法、SVD 等。

1.2　级联效应

1. 模型不同会改变目标形式与可解性，从而改变可用的算法。
2. 从线性分类器切换到神经网络，不仅假设空间改变，优化问题也由凸变为高度非凸，对应的算法与调参策略都要随之改变。

2　无约束优化与凸性

2.1　极小值与平滑性

1. 目标函数 $f(w)$ 连续可导且梯度连续时称为光滑（smooth）。
2. 全局极小 $w_\star$ 满足 $f(w_\star) \le f(v)$ 对任意 v 成立；局部极小在某个邻域内不可继续下降。

2.2　凸函数的关键性质

1. 定义：在凸域上，任意 x, y 与 $\beta \in [0,1]$，有

$$f\big(x + \beta(y - x)\big) \le (1 - \beta) f(x) + \beta f(y).$$

2. 凸函数之和仍凸；许多风险函数（如感知机风险、逻辑回归负对数似然）是多个样本损失的和，因此保持凸性。
3. 在封闭凸域上，连续凸函数要么无下界，要么只有一个局部极小，要么是由*全局等值极小点*构成的连通集合。后两种情形中，沿“下坡方向”迭代最终可达全局最优（或其等值集合）。

2.3　梯度下降与学习率

1. 基本迭代：$w_{t+1} = w_t - \varepsilon \nabla f(w_t)$。
2. ε 过大导致发散或震荡；过小导致收敛缓慢。一个简单的自适应策略是：若本步 f 上升则减小步长。

3. 在实践中，梯度下降通常是“趋近”极小值而不精确到达，这表现为*收敛*。

3 多维曲率、Hessian 与病态条件

3.1 方向曲率与 Hessian

定义 1 (Hessian)**.** 设 $f:\mathbb{R}^d\to\mathbb{R}$ 二阶可微，Hessian 为二阶导矩阵 $H(w)=\nabla^2 f(w)$。

命题 1 (方向曲率的二次型表示)**.** 对任意单位向量 v，令 $\phi(t)=f(w_0+tv)$，则

$$\phi''(0)=v^\top H(w_0)\,v.$$

证明要点：链式法则 $\phi'(t)=\nabla f(w_0+tv)^\top v$，再对 t 求导得 $\phi''(0)=v^\top\nabla^2 f(w_0)v$；或用二阶泰勒展开取 t^2 项系数即可。

命题 2 (特征值等于特征方向的曲率)**.** 当 H 为对称实矩阵（由二阶混合偏导可交换得知），存在正交分解 $H=Q\Lambda Q^\top$。若取方向 $v=q_i$（第 i 个特征向量），则

$$v^\top Hv=q_i^\top Hq_i=\lambda_i,$$

即对应特征值就是该方向的二阶曲率。

命题 3 (“Rayleigh–Ritz”上下界的朴素推导)**.** 任意单位 v 在特征向量基下写成 $v=\sum_i\alpha_i q_i$ 且 $\sum_i\alpha_i^2=1$。于是

$$v^\top Hv=\sum_i\lambda_i\alpha_i^2.$$

因此 $v^\top Hv$ 是特征值的加权平均，必位于 $[\lambda_{\min},\lambda_{\max}]$ 区间内，端点由 $v=q_{\min}$ 或 $v=q_{\max}$ 取得。

3.2 病态条件与学习率各向异性

1. 若 $\kappa(H)=\lambda_{\max}/\lambda_{\min}$ 很大，则不同方向曲率差异悬殊，等高线呈细长椭圆，称为病态（ill-conditioning）。
2. 单一学习率难以同时适配所有方向：对陡峭方向过大易发散，对平缓方向过小又收敛慢。
3. 典型缓解：特征缩放/标准化、预条件化、以及自适应学习率（Adam、RMSProp 等按坐标自调步长）。

3.3 鞍点

1. 鞍点是梯度为零但存在正负曲率方向并存的点，例如 $f(x,y)=x^2-y^2$ 在原点。
2. 在非凸问题（神经网络）中常见；SGD 的噪声有助于逃离部分鞍点。

4 线性规划（LP）：线性目标 + 线性不等式

4.1 形式与几何

1. 形式：max / min $c^\top w$ s.t. $Aw\le b$。可行域是凸多面体。

2. *活跃约束*：在最优解处以等号成立的约束。可能存在多重最优（如目标方向与某边界平行时一整条边都最优）；所有最优解的集合仍是凸的。

4.2 与“线性可分”的对应

1. 存在线性分类器等价于可行域非空：寻找 (w, α) 使 $y_i\,(w^\top x_i + \alpha) \geq 1$ 对所有 i。
2. LP 的挑战是确定最终会成为活跃的约束组合，组合数量呈指数级；常用**单纯形法**在多面体顶点间沿边移动直至最优。

5 二次规划（QP）与最大间隔 SVM

5.1 QP 基本形态与可解性

1. 形式：min $f(w) = w^\top Q w + c^\top w$ s.t. $Aw \leq b$，其中 Q 对称半正定则凸可解，Q 正定则解唯一。
2. 若 Q 不定，目标非凸，普遍为 NP-hard。

5.2 从“可分”到“最大间隔”的推导

1. 目标是最大化几何间隔
$$\gamma = \min_i \frac{y_i\,(w^\top x_i + b)}{\|w\|}.$$
2. 通过缩放标准化约束 $y_i(w^\top x_i + b) \geq 1$，此时 $\gamma = 1/\|w\|$。
3. 因而“最大间隔”等价于
$$\min_{w,b}\ \tfrac{1}{2}\|w\|^2 \quad \text{s.t.} \quad y_i(w^\top x_i + b) \geq 1,$$
即二次目标 + 线性不等式约束的 QP。
4. 最优点处贴边的样本构成*支持向量*，是几何上的活跃约束。

5.3 软间隔与合页损失的等价

1. 带松弛变量的带约束形式：
$$\min_{w,b,\xi}\ \tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad \text{s.t.} \quad y_i(w^\top x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0.$$
2. 由约束得 $\xi_i \geq \max(0,\, 1 - y_i(w^\top x_i + b))$；由于目标对 ξ_i 单调，在最优处取下界，故等价为*无约束*：
$$\min_{w,b}\ \tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\max\bigl(0,\, 1 - y_i(w^\top x_i + b)\bigr),$$
第二项即合页损失（hinge loss）。

5.4 核技巧一言以蔽之

1. 将内积 $x^\top z$ 替换为核函数 $K(x, z) = \langle \Phi(x), \Phi(z) \rangle$，在隐式高维中作线性分类，从而得到非线性判别边界。

5.5 QP 求解算法与对比

1. **单纯形类（Simplex-like）**：适用广、能精确解一般 QP；但大规模慢、内存压力大；更偏“组合/离散”风格。
2. **SMO（Sequential Minimal Optimization）**：把对偶 QP 分解成反复求解 2 变量子问题；中等规模核 SVM 高效，LIBSVM 与 `sklearn.SVC` 采用；但超大规模或核矩阵存储时仍受限。
3. **坐标下降（Coordinate Descent）**：逐坐标优化，极擅长超大规模稀疏线性 SVM（文本/点击率等），LIBLINEAR 与 `LinearSVC` 采用；但不直接适配核 SVM，收敛速度依赖坐标选择策略。

6 实践要点与易错清单

6.1 学习率与收敛

1. 一维：步长过大发散，过小缓慢；多维病态时单一步长无法兼顾各方向。
2. 简便诊断：若某步 f 上升，缩小步长；使用学习率衰减或自适应策略（Adam、RMSProp）。

6.2 数据与特征处理

1. 线性模型前建议标准化特征；对病态问题可考虑预条件化。
2. 核 SVM 的核宽度、正则系数 C 需网格或贝叶斯调参；大数据优先线性 SVM 或近似核方法。

6.3 从 LP 到 QP 的识别技巧

1. 只需“可分”的分类 $\Rightarrow$ LP 可行性问题：判断不等式系统是否有解。
2. 追求“最大间隔” $\Rightarrow$ QP：最小化 $\frac{1}{2}\|w\|^2$ 加线性约束。
3. 不可分且希望鲁棒 $\Rightarrow$ 软间隔 + 合页损失（等价无约束）。

7 概念速查表

术语	要点
局部/全局极小	全局对所有点最小；局部在邻域内最小
凸函数	连线在图像之上；凸和仍凸；许多经验风险是凸的
Hessian	二阶导矩阵；方向曲率为 $v^\top H v$
病态（条件数）	$\kappa = \lambda_{\max}/\lambda_{\min} \gg 1$，不同方向曲率差异大
鞍点	梯度为零但存在正负曲率方向并存
LP	线性目标 + 线性不等式；可行域多面体；单纯形法
QP	二次凸目标 + 线性不等式；$Q \succeq 0$ 则凸可解
SVM（硬间隔）	$\min \frac{1}{2}\|w\|^2$ s.t. $y_i(w^\top x_i + b) \geq 1$
SVM（软间隔）	加松弛 ξ_i；等价于 $\frac{1}{2}\|w\|^2 + C\sum \max(0, 1 - y_i f_i)$
核技巧	用 $K(x, z) = \langle \Phi(x), \Phi(z) \rangle$ 替代内积
SMO	核 SVM 常用；逐个两变量子问题精解
坐标下降	线性 SVM 大规模稀疏数据利器

8 关键推导与证明摘要

A. 方向曲率 = 二次型

设 $\phi(t) = f(w_0 + tv)$，链式法则给出 $\phi''(0) = v^\top \nabla^2 f(w_0)\, v$。因此 Hessian 的特征值是特征方向上的曲率；任意方向曲率为特征值的非负加权平均，从而被夹在 $[\lambda_{\min}, \lambda_{\max}]$ 内。

B. 软间隔 SVM $\Rightarrow$ 合页损失

约束 $y_i(w^\top x_i + b) \geq 1 - \xi_i$ 与 $\xi_i \geq 0$ 蕴含 $\xi_i \geq \max(0, 1 - y_i f_i)$。目标对 ξ_i 单调，最优取下界，代回得 $\min\ \frac{1}{2}\|w\|^2 + C\sum \max(0, 1 - y_i f_i)$。