中國人民大學

RENMIN UNIVERSITY OF CHINA

XXXX

XXXXX

学院 ________________________________

专业 ________________________________

学号 ________________________________

姓名 ________________________________

2025 年 8 月 28 日

摘　要

XXXXX

关键词：XXX

目录

1 基本设定与增广记号

1. 训练集 $\{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathbb{R}^d$, $y_i \in \{\pm 1\}$；线性分类器 $f(x) = w \cdot x + b$，超平面 $H = \{x : w \cdot x + b = 0\}$。

2. 增广向量：$\tilde{x} = (x, 1) \in \mathbb{R}^{d+1}$，$\tilde{w} = (w, b) \in \mathbb{R}^{d+1}$，则 $f(x) = \tilde{w} \cdot \tilde{x}$；在 d+1 维里用过原点超平面统一表示一般超平面。

3. 预测 $\hat{y}(x) = \text{sign}(f(x))$；分类约束 $y_i(w \cdot x_i + b) \geq 0$。

2 x-space 与 w-space 的几何对偶

2.1 点–超平面对偶与分类约束的两种视角

命题 1 (对偶关系)**.** *在 x-space：超平面 $\{z : w \cdot z + b = 0\}$ 对应法向量 w；在 w-space：样本 x 诱导超平面 $\{z : x \cdot z + b = 0\}$。存在等价式*

$$x \in \{z : w \cdot z + b = 0\} \iff w \in \{z : x \cdot z + b = 0\}.$$

结论：把“在 x-space 找超平面”转化为“在 w-space 找点”。分类约束 $y_i(w \cdot x_i + b) \geq 0$ 在 x-space 表示样本在边界正确侧；在 w-space 表示 w 在样本诱导超平面的正确侧。所有半空间的交集为 w 的可行域。

2.2 “w 既是向量又是点”的说明

参数 $w = (w_1, \ldots, w_d)$ 在 w-space 中是点的坐标；在 x-space 中表现为法向量。优化问题“找 w”即“在 $\mathbb{R}^d$ 中找点”。

3 风险函数、形状与梯度

3.1 仅误分类计损的风险与感知机损失

1. 仅误分类计损的风险函数

$$R(w, b) = \sum_{i \in V} -y_i\,(w \cdot x_i + b), \quad V = \{i :\ y_i(w \cdot x_i + b) < 0\}.$$

   它在 w-space 呈分段线性凸“折面”，折痕与样本诱导直线对齐；$R = 0$ 的底部扇区即全体正确分类解集。

2. 感知机（凸）损失 $L_i(w, b) = \max(0, -y_i(w \cdot x_i + b))$，经验风险 $F(w, b) = \frac{1}{n} \sum_i L_i$；两者在误分类处的（次）梯度方向一致。

3.2 梯度、最速下降与更新

1. 多元一阶泰勒 $f(w+\Delta)\approx f(w)+\nabla f(w)^\top\Delta$；令 $\Delta=-\eta\nabla f(w)$ 得

$$w^{(t+1)}=w^{(t)}-\eta\nabla f\big(w^{(t)}\big).$$

2. 对 R 的梯度：$\nabla_w R(w)=-\sum_{i\in V}y_ix_i$, $\partial_b R(w)=-\sum_{i\in V}y_i$。对 L_i：若 $y_i(w\cdot x_i+b)<0$，则 $\partial_w L_i=-y_ix_i$, $\partial_b L_i=-y_i$，否则为 0。

4 GD 与 SGD：定义、复杂度与在线性

4.1 两种下降

1. 全量 GD：$w\leftarrow w-\eta\nabla F(w)$，每步 $O(nd)$，方向平稳。
2. SGD: 随机抽样 i，$w\leftarrow w-\eta\nabla L_i(w)$，每步 $O(d)$；由于 $F=\frac{1}{n}\sum_i L_i$，有 $\mathbb{E}[\nabla L_i(w)]=\nabla F(w)$，单步有噪声但无偏。

4.2 在线算法与步长

1. 感知机是在线算法：新样本到来可继续迭代。
2. 步长 η 不出现在错误次数的大 O 上界，但实际运行时间对 η 敏感：η 太小则慢；太大则可能跨越零风险区、在边界两侧振荡。可用线搜索、自适应步长或直接改用二次规划。

5 偏置与“增加一维”的统一

1. 一般超平面 $w\cdot x+\alpha=0$ 可通过增广 $\tilde{x}=(x,1)$、$\tilde{w}=(w,\alpha)$ 统一为 $\tilde{w}\cdot\tilde{x}=0$，从而用一条更新式覆盖 w 与偏置。
2. 在 $d+1$ 维空间中，训练点共面于 $x_{d+1}=1$ 的切片，仍可直接运行感知机或其他线性分类算法。

6 凸优化与 SGD 收敛

1. 最小化凸函数：局部极小即全局极小；合适步长（如衰减序列）下，SGD 在凸光滑情形以期望收敛。
2. 非凸目标（如深度网络）通常无法保证全局最优，但 SGD 仍实用；在本课的线性可分与感知机损失场景下，存在更强的有限步收敛结论。

7 感知机收敛定理与错误上界

7.1 陈述与参数

定义半径 $R=\max_i\|\tilde{x}_i\|$。若存在单位向量 $\|\tilde{w}^\star\|=1$ 使得几何间隔下界

$$\gamma=\min_i y_i(\tilde{w}^\star\cdot\tilde{x}_i)>0,$$

则在线更新（仅误分类样本）$\tilde{w} \leftarrow \tilde{w} + \eta\, y_i \tilde{x}_i$ 的犯错次数 T 满足

$$T \le \Big(\frac{R}{\gamma}\Big)^2 \frac{1}{\eta^2}.$$

常用归一化 $R \le 1,\ \eta = 1$ 时，$T \le 1/\gamma^2$。

7.2 证明（Novikoff）

1. 投影线性累增：每次犯错使 $\langle \tilde{w}^{(t+1)}, \tilde{w}^\star \rangle \ge \langle \tilde{w}^{(t)}, \tilde{w}^\star \rangle + \eta\gamma$，累加得 $\langle \tilde{w}^{(T)}, \tilde{w}^\star \rangle \ge T\eta\gamma$。
2. 范数二次受限：犯错时 $y_t(\tilde{w}^{(t)} \cdot \tilde{x}_t) < 0$，故 $\|\tilde{w}^{(t+1)}\|^2 \le \|\tilde{w}^{(t)}\|^2 + \eta^2 \|\tilde{x}_t\|^2 \le \|\tilde{w}^{(t)}\|^2 + \eta^2 R^2$，从而 $\|\tilde{w}^{(T)}\| \le \eta R \sqrt{T}$。
3. Cauchy–Schwarz：$T\eta\gamma \le \|\tilde{w}^{(T)}\| \le \eta R \sqrt{T}$，解得结论。

8 间隔（margin）的本质与最大化间隔

8.1 函数间隔与几何间隔

定义 1 (函数间隔)**.** $\hat{\gamma}_i = y_i(w \cdot x_i + \alpha)$，*受 w 的缩放影响。*

定义 2 (几何间隔)**.** $\gamma_i = \dfrac{y_i(w \cdot x_i + \alpha)}{\|w\|}$ *为点到超平面的带符号欧氏距离，整体间隔* $\gamma = \min_i \gamma_i$。

8.2 定标、条带与最大化

为消除缩放不定性并排除 $w = 0$，施加

$$y_i(w \cdot x_i + \alpha) \ge 1, \quad i = 1, \ldots, n.$$

此时最近样本（支持向量）满足等式，故

$$\gamma = \frac{1}{\|w\|}, \qquad \text{条带宽度} = 2/\|w\|, \quad \text{中线} w \cdot x + \alpha = 0.$$

最大化间隔等价于最小化 $\|w\|$（实际用光滑的 $\frac{1}{2}\|w\|^2$）。

9 硬间隔 SVM 的凸二次规划与权重空间几何

9.1 原始问题（线性可分）

$$\boxed{\min_{w,\alpha} \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad y_i(w \cdot x_i + \alpha) \ge 1,\ i = 1, \ldots, n.}$$

1. 目标严格凸、约束线性，线性可分时解唯一；最优几何间隔为 $1/\|w\|$。
2. 采用 $\|w\|^2$ 而非 $\|w\|$ 的理由是可微性与计算便利，解不变。

9.2 权重空间三维/截面直观

1. 在 (w_1, w_2, α) 空间，每个训练点诱导一个线性约束平面；可行域为这些半空间交集。
2. 目标只惩罚 w 的长度（到 α 轴的水平距离），最优解是可行域中离 α 轴最近的点；二维截面中呈若干直线交点（支持向量）给定的解。

10 从感知机到 SVM：动机与实践

1. 感知机只求可分，解不唯一、对步长敏感、整体仍可能慢；SVM 通过最大化间隔给出唯一且鲁棒的解，二次规划求解更稳定高效。
2. 数据不可分时，引入软间隔 $y_i(w \cdot x_i + \alpha) \geq 1 - \xi_i$, $\xi_i \geq 0$ 并最小化 $\frac{1}{2}\|w\|^2 + C\sum_i \xi_i$；进一步可用核方法得到非线性边界。

11 历史与备注

1. Rosenblatt（1957）在康奈尔航空实验室提出并以硬件实现 Mark I 感知机（用于 20×20 像素图像），感知机是深度学习历史的起点之一。
2. 现代优化与统计学习理论推动了从“可分即可”到“最大间隔与泛化”的转变。

12 常见疑问与澄清

1. “最优点是否一定在超平面交点”：最优解位于可行域边界；在 d 维可为若干约束的交点/交线/交面。
2. “是否降维”：属于解集自由度因约束而降低，并非数据维度降维（不同于 PCA）。
3. “为何用 $-\nabla f$ 更新”：来自多元一阶泰勒与最速下降方向。
4. “单样本梯度为何无偏”：$F = \frac{1}{n}\sum_i L_i \Rightarrow \nabla F = \frac{1}{n}\sum_i \nabla L_i$，均匀抽样 $\mathbb{E}[\nabla L_i] = \nabla F$。
5. “为何右端取 1 而非 0”：消除缩放不定性、排除 $w = 0$，并让几何间隔与 $\|w\|$ 建立一一关系，从而把“最大化间隔”转为“最小化 $\|w\|$”。