第 10 章 逻辑回归

10.1 逻辑回归问题

定义 10.1.1 (软二分类（Soft Binary Classification）)

目标函数为条件概率

$$f(x) = P(+1 \mid x) \in [0,1].$$

- 理想（无噪）数据：观测标签 $y^\star = f(x)$，即直接给出真实概率。
- 实际（含噪）数据：观测标签 $y \sim P(y \mid x)$，其中

$$P(y=+1 \mid x) = f(x), \quad P(y=-1 \mid x) = 1 - f(x).$$

要点 输入数据 $\{(x_n, y_n)\}_{n=1}^N$ 与硬二分类完全相同，但目标函数不再是 $\{-1,+1\}$ 而是连续概率 $[0,1]$。

♣

定义 10.1.2 (逻辑假设（Logistic Hypothesis）)

给定病人特征向量

$$\mathbf{x} = (x_0, x_1, x_2, \ldots, x_d)^\top = (\text{年龄}, \text{性别}, \text{血压}, \text{胆固醇}, \ldots)^\top,$$

首先计算加权“风险分数”

$$s = \mathbf{w}^\top \mathbf{x} = \sum_{i=0}^{d} w_i x_i.$$

再用逻辑函数（logistic function）将分数映射为概率估计

$$\sigma(s) = \frac{1}{1+e^{-s}} \in (0,1).$$

于是逻辑假设定义为

$$h(\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^\top \mathbf{x})},$$

用于逼近目标条件概率

$$f(\mathbf{x}) = P(y=+1 \mid \mathbf{x}).$$

♣

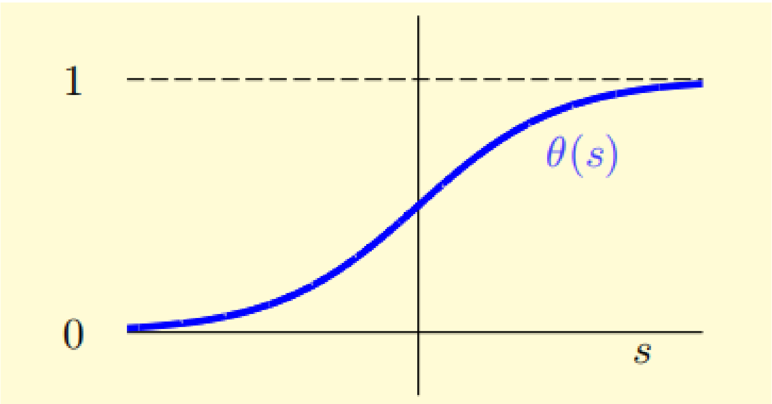

图 10.1.1: sigmoid 函数图像

例题 10.1Logistic 回归的二分类决策

设 Logistic 假设

$$h(\mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}^\top \mathbf{x}}} \in (0,1),$$

以阈值 0.5 将其转化为二分类预测：

$$\text{预测类别} = \operatorname{sign}\big(h(\mathbf{x}) - 0.5\big).$$

选项　上述决策等价于下列哪个公式？

1) $\operatorname{sign}(\mathbf{w}^\top \mathbf{x} - 0.5)$
2) $\operatorname{sign}(\mathbf{w}^\top \mathbf{x})$
3) $\operatorname{sign}(\mathbf{w}^\top \mathbf{x} + 0.5)$
4) 以上皆非

解答　正确选项为 $\boxed{2}$。由于当 $\mathbf{w}^\top \mathbf{x} = 0$ 时，$h(\mathbf{x}) = 0.5$，故在 0.5 处对 $h(\mathbf{x})$ 阈值化等价于在 0 处对 $\mathbf{w}^\top \mathbf{x}$ 阈值化。 ■

10.2 逻辑回归的误差

定义 10.2.1 (似然（Likelihood）)

给定目标函数

$$f(x) = P(y=+1 \mid x), \qquad P(y=-1 \mid x) = 1 - f(x),$$

及数据集

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}, \quad y_n \in \{+1, -1\}.$$

目标函数生成数据的概率

$$\prod_{n=1}^{N} P(x_n)\, f(x_n)^{\frac{1+y_n}{2}} \left[1 - f(x_n)\right]^{\frac{1-y_n}{2}}.$$

模型假设 h 生成数据的似然

$$\text{likelihood}(h) = \prod_{n=1}^{N} P(x_n)\, h(x_n)^{\frac{1+y_n}{2}} \left[1 - h(x_n)\right]^{\frac{1-y_n}{2}}.$$

当 $h \approx f$ 时，$\text{likelihood}(h)$ 接近使用真实 f 的概率，而真实概率通常较大。 ♣

命题 10.2.1 (Logistic 假设的似然)

设 Logistic 假设

$$h(\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}^\top \mathbf{x}}}, \quad \text{满足} \quad 1 - h(\mathbf{x}) = h(-\mathbf{x}).$$

给定数据集 $\{(x_n, y_n)\}_{n=1}^{N}$，其中 $y_n \in \{+1, -1\}$，则似然函数

$$\begin{aligned} likelihood(h) &= P(\mathbf{x}_1)h(\mathbf{x}_1) \times P(\mathbf{x}_2)(1 - h(\mathbf{x}_2)) \times \cdots \times P(\mathbf{x}_N)(1 - h(\mathbf{x}_N)) \\ &= P(\mathbf{x}_1)h(+\mathbf{x}_1) \times P(\mathbf{x}_2)h(-\mathbf{x}_2) \times \cdots \times P(\mathbf{x}_N)h(-\mathbf{x}_N) \end{aligned}$$

可以看出似然函数正比于

$$\text{likelihood}(h) \propto \prod_{n=1}^{N} h(y_n x_n) = \prod_{n=1}^{N} \sigma(y_n \mathbf{w}^\top \mathbf{x}_n).$$

最优参数通过最大似然估计获得

$$\mathbf{w} = \arg\max_{\mathbf{w}} \prod_{n=1}^{N} \sigma(y_n \mathbf{w}^\top \mathbf{x}_n).$$

定义 10.2.2 (交叉熵误差（Cross-Entropy Error）)

给定数据集 $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$，其中 $y_n \in \{-1, +1\}$，逻辑回归假设

$$h(x; w) = \sigma(w^\top x) = \frac{1}{1 + \exp(-w^\top x)}.$$

似然函数

$$\text{likelihood}(w) \propto \prod_{n=1}^{N} h(x_n; w)^{\frac{1+y_n}{2}} \left[1 - h(x_n; w)\right]^{\frac{1-y_n}{2}} = \prod_{n=1}^{N} \sigma(y_n w^\top x_n).$$

对数似然与交叉熵误差　最大化对数似然等价于最小化交叉熵误差

$$E_{\text{in}}(w) = \frac{1}{N} \sum_{n=1}^{N} \underbrace{\ln\big(1 + \exp(-y_n w^\top x_n)\big)}_{\text{err}(w, x_n, y_n)}.$$

10.3 逻辑回归误差的梯度

命题 10.3.1 (最小化交叉熵误差 $E_{\text{in}}(w)$)

交叉熵误差

$$E_{\text{in}}(w) = \frac{1}{N} \sum_{n=1}^{N} \ln\big(1 + \exp(-y_n w^\top x_n)\big)$$

连续、可微、二阶可微且凸。

梯度计算

$$\nabla E_{\text{in}}(w) = \frac{1}{N} \sum_{n=1}^{N} \underbrace{\sigma(-y_n w^\top x_n)}_{\theta_n}(-y_n x_n) = \frac{1}{N} \sum_{n=1}^{N} \theta_n (-y_n x_n),$$

其中 $\theta_n \in (0, 1)$ 为样本权重。

最优条件

$$\nabla E_{\text{in}}(w) = \mathbf{0} \implies \sum_{n=1}^{N} \theta_n y_n x_n = \mathbf{0}.$$

- 若数据线性可分，则 $\forall n$, $y_n w^\top x_n > 0$，所有 $\theta_n \to 0$，梯度趋于零；
- 一般情形下上式为非线性方程，无闭式解，需数值迭代（如梯度下降、Newton 法）。

算法 10.3.1: 感知机学习算法（PLA）：迭代优化视角

输入: 数据集 $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$，其中 $y_n \in \{-1, +1\}$

输出: 权重向量 w（作为最终假设 g）

初始化 $w_0 \leftarrow \mathbf{0}$;

for $t = 0, 1, 2, \ldots$ **do**

任选一个误分类样本 $(x_{n(t)}, y_{n(t)})$ 满足

$$\mathrm{sign}(w_t^\top x_{n(t)}) \neq y_{n(t)};$$

更新权重

$$w_{t+1} \leftarrow w_t + y_{n(t)} x_{n(t)};$$

if 所有样本被正确分类 **then break;**

return $w \leftarrow w_t$;

10.4 梯度下降

算法 10.4.2: 梯度下降法最小化 $E_{\text{in}}(w)$

输入: 初始权重 $w_0 \in \mathbb{R}^{d+1}$，学习率 $\eta > 0$，迭代次数 T

输出: 最终权重 $g = w_T$

for $t = 0$ **to** $T - 1$ **do**

计算梯度：$\mathbf{g}_t = \nabla E_{\text{in}}(w_t)$;

更新权重：$w_{t+1} \leftarrow w_t - \eta \dfrac{\mathbf{g}_t}{\|\mathbf{g}_t\|}$;

return w_T;

命题 10.4.1 (梯度下降的原理)

利用一阶泰勒近似

$$E_{\text{in}}(w_t + \eta v) \approx E_{\text{in}}(w_t) + \eta v^\top \nabla E_{\text{in}}(w_t), \quad \|v\| = 1,$$

当步长 η 足够小时，最优方向为负梯度

$$v^\star = -\frac{\nabla E_{\text{in}}(w_t)}{\|\nabla E_{\text{in}}(w_t)\|}.$$

因此迭代公式

$$w_{t+1} = w_t + \eta v^\star = w_t - \eta \frac{\nabla E_{\text{in}}(w_t)}{\|\nabla E_{\text{in}}(w_t)\|}$$

保证了 E_{in} 沿“下坡”方向单调下降，是简单且流行的优化工具。♠

命题 10.4.2 (固定学习率的梯度下降启发式)

设当前梯度为 $\mathbf{g}_t = \nabla E_{\text{in}}(w_t)$。若将学习率 η 设为与梯度范数单调相关的量（例如 $\eta \propto \|\mathbf{g}_t\|$），则步长

$$w_{t+1} = w_t - \eta \frac{\mathbf{g}_t}{\|\mathbf{g}_t\|}$$

退化为固定步长。实践中更常用的是固定学习率（fixed learning rate）$\eta > 0$ 的梯度下降：

$$\boxed{w_{t+1} \leftarrow w_t - \eta \nabla E_{\text{in}}(w_t).}$$

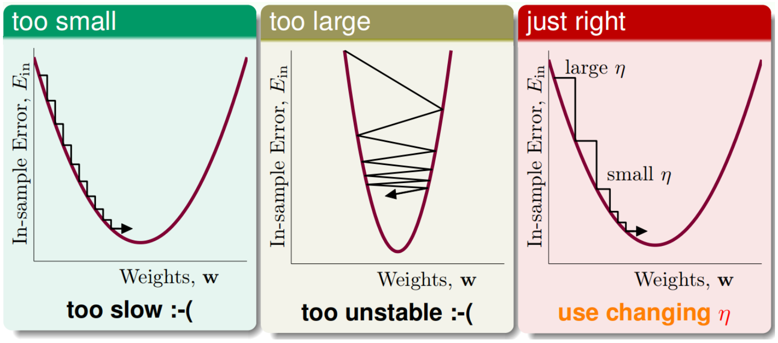

图 10.4.1: 学习率对优化过程影响示意图

算法 10.4.3: 逻辑回归 (梯度下降)

输入: 训练集 $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$，其中 $y_n \in \{-1, +1\}$；学习率 $\eta > 0$；最大迭代次数 T

输出: 最终权重向量 g

初始化 $w_0 \in \mathbb{R}^{d+1}$;

for $t = 0, 1, \ldots, T-1$ **do**

计算梯度

$$\nabla E_{\text{in}}(w_t) = \frac{1}{N} \sum_{n=1}^{N} \sigma(-y_n w_t^\top x_n)(-y_n x_n);$$

更新权重

$$w_{t+1} \leftarrow w_t - \eta \nabla E_{\text{in}}(w_t);$$

if $\|\nabla E_{in}(w_{t+1})\| < \varepsilon$ **then break;**

return $g \leftarrow w_{t+1}$;

注 每轮迭代时间复杂度与 Pocket 算法一轮相当。

10.5 总结

笔记 [逻辑回归]

- 逻辑回归问题：以 $P(+1|x)$ 为目标，用 $w^\top x$ 作为假设。
- 逻辑回归的误差：交叉熵（负对数似然）。
- 逻辑回归误差的梯度：以权重 A 加权的样本向量之和。
- 梯度下降：沿负梯度 $-\nabla E_{\text{in}}(w)$ 向“山下”滚动更新。