对数几率回归

张腾

2025 年 3 月 24 日

对数几率回归 (logisitic regression, LR) 虽然名字里有回归二字，但其实是一个用于分类的线性超平面模型。

1 二分类

对二分类问题，不妨设 $\mathcal{Y} = \{1, -1\}$，由 Bayes's 定理知

$$\begin{aligned}\mathbb{P}(y=1 \mid \boldsymbol{x}) &= \frac{\mathbb{P}(\boldsymbol{x}, y=1)}{\mathbb{P}(\boldsymbol{x})} = \frac{\mathbb{P}(\boldsymbol{x}, y=1)}{\mathbb{P}(\boldsymbol{x}, y=1) + \mathbb{P}(\boldsymbol{x}, y=-1)} \\ &= \frac{1}{1 + {}^{\mathbb{P}(\boldsymbol{x}, y=-1)}/_{\mathbb{P}(\boldsymbol{x}, y=1)}} = \frac{1}{1+\exp(-a)} = \sigma(a)\end{aligned}$$

其中 $\sigma : \mathbb{R} \mapsto (0,1)$ 为对数几率函数 (logisitic function)，其将正负类概率比 (几率，odds) 的对数

$$a = \ln \frac{\mathbb{P}(\boldsymbol{x}, y=1)}{\mathbb{P}(\boldsymbol{x}, y=-1)} = \ln \frac{\mathbb{P}(\boldsymbol{x}, y=1)}{1-\mathbb{P}(\boldsymbol{x}, y=1)} = \operatorname{logit}(\mathbb{P}(\boldsymbol{x}, y=1)) = \sigma^{-1}(\mathbb{P}(\boldsymbol{x}, y=1))$$

映射为正类概率。LR 用线性模型 $\boldsymbol{w}^\top \boldsymbol{x} + b$ 拟合 a，故名对数几率回归。另一方面，LR 也是在用线性模型 $\boldsymbol{w}^\top \boldsymbol{x} + b$ 输出的 σ 变换拟合正类概率：

$$\begin{aligned}\mathbb{P}(y=1 \mid \boldsymbol{x}) &= \sigma(\boldsymbol{w}^\top \boldsymbol{x} + b) = \frac{1}{1+\exp(-(\boldsymbol{w}^\top \boldsymbol{x} + b))} \\ \mathbb{P}(y=-1 \mid \boldsymbol{x}) &= 1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x} + b) = \frac{\exp(-(\boldsymbol{w}^\top \boldsymbol{x} + b))}{1+\exp(-(\boldsymbol{w}^\top \boldsymbol{x} + b))} = \frac{1}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x} + b)} = \sigma(-(\boldsymbol{w}^\top \boldsymbol{x} + b))\end{aligned}$$

给定 $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i\in[m]} \subseteq \mathbb{R}^n \times \{1,-1\}$，求解 $\boldsymbol{w}$ 和 b 可采用极大似然法，即

$$\begin{aligned}\operatorname*{argmax}_{\boldsymbol{w},b} \ln \mathbb{P}(\mathcal{D} \mid \boldsymbol{w}, b) &= \operatorname*{argmax}_{\boldsymbol{w},b} \ln \prod_{i\in[m]} \mathbb{P}(\boldsymbol{x}_i, y_i \mid \boldsymbol{w}, b) = \operatorname*{argmax}_{\boldsymbol{w},b} \sum_{i\in[m]} \ln \mathbb{P}(\boldsymbol{x}_i, y_i \mid \boldsymbol{w}, b) \\ &= \operatorname*{argmax}_{\boldsymbol{w},b} \sum_{i\in[m]} \ln \mathbb{P}(y_i \mid \boldsymbol{x}_i, \boldsymbol{w}, b) = \operatorname*{argmax}_{\boldsymbol{w},b} \sum_{i\in[m]} \ln \frac{1}{1+\exp(-y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b))} \\ &= \operatorname*{argmin}_{\boldsymbol{w},b} \sum_{i\in[m]} \ln(1+\exp(-y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)))\end{aligned}$$

上述形式化也可以通过交叉熵损失推导出来，类别标记的独热编码为 $[{}^{1+y}/_{2}, {}^{1-y}/_{2}]$，模型预测的概率分布为 $[{}^{1}/_{1+\exp(-(\boldsymbol{w}^\top \boldsymbol{x}+b))}, {}^{1}/_{1+\exp(\boldsymbol{w}^\top \boldsymbol{x}+b)}]$，故经验损失为

$$
\begin{aligned}
\ell_{\mathrm{CE}} &= \sum_{i \in [m]} \left(\frac{1+y_i}{2} \ln(1+\exp(-(\boldsymbol{w}^\top \boldsymbol{x}_i + b))) + \frac{1-y_i}{2} \ln(1+\exp(\boldsymbol{w}^\top \boldsymbol{x}_i + b)) \right) \\
&= \sum_{i \in [m]} \left(\mathbb{I}(y_i = 1) \ln(1+\exp(-(\boldsymbol{w}^\top \boldsymbol{x}_i + b))) + \mathbb{I}(y_i = -1) \ln(1+\exp(\boldsymbol{w}^\top \boldsymbol{x}_i + b)) \right) \\
&= \sum_{i \in [m]} \ln(1+\exp(-y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)))
\end{aligned}
$$

2 多分类

对多分类问题，不妨设 $\mathcal{Y} = [c]$，LR 引入 c 个线性模型 $\boldsymbol{w}_i^\top \boldsymbol{x} + b_i$，其中 $i \in [c]$，并用 c 个模型的输出的 softmax 变换拟合各类概率

$$
\begin{aligned}
\mathbb{P}(y = 1 \mid \boldsymbol{x}) = \sigma_1(\boldsymbol{x}) &= \frac{\exp(\boldsymbol{w}_1^\top \boldsymbol{x} + b_1)}{\sum_{i \in [c]} \exp(\boldsymbol{w}_i^\top \boldsymbol{x} + b_i)} = \frac{\exp((\boldsymbol{w}_1 - \boldsymbol{w}_c)^\top \boldsymbol{x} + b_1 - b_c)}{1 + \sum_{i \in [c-1]} \exp((\boldsymbol{w}_i - \boldsymbol{w}_c)^\top \boldsymbol{x} + b_i - b_c)} \\
&\vdots \\
\mathbb{P}(y = c-1 \mid \boldsymbol{x}) = \sigma_{c-1}(\boldsymbol{x}) &= \frac{\exp(\boldsymbol{w}_{c-1}^\top \boldsymbol{x} + b_{c-1})}{\sum_{i \in [c]} \exp(\boldsymbol{w}_i^\top \boldsymbol{x} + b_i)} = \frac{\exp((\boldsymbol{w}_{c-1} - \boldsymbol{w}_c)^\top \boldsymbol{x} + b_{c-1} - b_c)}{1 + \sum_{i \in [c-1]} \exp((\boldsymbol{w}_i - \boldsymbol{w}_c)^\top \boldsymbol{x} + b_i - b_c)} \\
\mathbb{P}(y = c \mid \boldsymbol{x}) = \sigma_c(\boldsymbol{x}) &= \frac{\exp(\boldsymbol{w}_c^\top \boldsymbol{x} + b_c)}{\sum_{i \in [c]} \exp(\boldsymbol{w}_i^\top \boldsymbol{x} + b_i)} = \frac{1}{1 + \sum_{i \in [c-1]} \exp((\boldsymbol{w}_i - \boldsymbol{w}_c)^\top \boldsymbol{x} + b_i - b_c)}
\end{aligned}
$$

也可以直接学习 $\boldsymbol{w}_i' = \boldsymbol{w}_i - \boldsymbol{w}_c$ 和 $b_i' = b_i - b_c$，其中 $i \in [c-1]$，这样变量少一些。当 $c = 2$ 时退化为二分类，$\boldsymbol{w}_1 - \boldsymbol{w}_2$ 和 $b_1 - b_2$ 就是前面学习的 $\boldsymbol{w}$ 和 b，因此 softmax 变换就是 σ 变换的推广。

给定 $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i \in [m]} \subseteq \mathbb{R}^n \times [c]$，用极大似然求解

$$
\begin{aligned}
\operatorname*{argmax}_{\boldsymbol{w}_j, b_j, j \in [c]} \ln \mathbb{P}(\mathcal{D} \mid \boldsymbol{w}_j, b_j, j \in [c]) &= \operatorname*{argmax}_{\boldsymbol{w}_j, b_j, j \in [c]} \ln \prod_{i \in [m]} \mathbb{P}(\boldsymbol{x}_i, y_i \mid \boldsymbol{w}_j, b_j, j \in [c]) \\
&= \operatorname*{argmax}_{\boldsymbol{w}_j, b_j, j \in [c]} \sum_{i \in [m]} \ln \mathbb{P}(\boldsymbol{x}_i, y_i \mid \boldsymbol{w}_j, b_j, j \in [c]) \\
&= \operatorname*{argmax}_{\boldsymbol{w}_j, b_j, j \in [c]} \sum_{i \in [m]} \ln \mathbb{P}(y_i \mid \boldsymbol{x}_i, \boldsymbol{w}_j, b_j, j \in [c]) \\
&= \operatorname*{argmax}_{\boldsymbol{w}_j, b_j, j \in [c]} \sum_{i \in [m]} \ln \frac{\exp(\boldsymbol{w}_{y_i}^\top \boldsymbol{x}_i + b_{y_i})}{\sum_{j \in [c]} \exp(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j)} = \operatorname*{argmax}_{\boldsymbol{w}_j, b_j, j \in [c]} \sum_{i \in [m]} \ln \sigma_{y_i}(\boldsymbol{x}_i)
\end{aligned}
$$

注意

$$
\begin{aligned}
\frac{\partial \sigma_{y_i}(\boldsymbol{x}_i)}{\partial \boldsymbol{w}_{y_i}} &= \frac{\exp(\boldsymbol{w}_{y_i}^\top \boldsymbol{x}_i) \boldsymbol{x}_i \sum_{j \in [c]} \exp(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \exp(\boldsymbol{w}_{y_i}^\top \boldsymbol{x}_i + b_{y_i}) \exp(\boldsymbol{w}_{y_i}^\top \boldsymbol{x}_i + b_{y_i}) \boldsymbol{x}_i}{(\sum_{j \in [c]} \exp(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j))^2} \\
&= \sigma_{y_i}(\boldsymbol{x}_i) \boldsymbol{x}_i - \sigma_{y_i}^2(\boldsymbol{x}_i) \boldsymbol{x}_i = \sigma_{y_i}(\boldsymbol{x}_i)(1 - \sigma_{y_i}(\boldsymbol{x}_i)) \boldsymbol{x}_i \\
\frac{\partial \sigma_{y_i}(\boldsymbol{x}_i)}{\partial \boldsymbol{w}_k} &= \frac{-\exp(\boldsymbol{w}_{y_i}^\top \boldsymbol{x}_i + b_{y_i}) \exp(\boldsymbol{w}_k^\top \boldsymbol{x}_i + b_k) \boldsymbol{x}_i}{(\sum_{j \in [c]} \exp(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j))^2} = -\sigma_{y_i}(\boldsymbol{x}_i) \sigma_k(\boldsymbol{x}_i) \boldsymbol{x}_i, \ k \neq y_i
\end{aligned}
$$

合并起来可写成 ${\partial \sigma_{y_i}(\boldsymbol{x}_i)}/{\partial \boldsymbol{w}_k} = \sigma_{y_i}(\boldsymbol{x}_i)(\mathbb{I}(k=y_i) - \sigma_k(\boldsymbol{x}_i))\boldsymbol{x}_i$，于是

$$\frac{\partial \sum_{i\in[m]} \ln \sigma_{y_i}(\boldsymbol{x}_i)}{\partial \boldsymbol{w}_k} = \sum_{i\in[m]} \frac{1}{\sigma_{y_i}(\boldsymbol{x}_i)} \frac{\partial \sigma_{y_i}(\boldsymbol{x}_i)}{\partial \boldsymbol{w}_k} = \sum_{i\in[m]} (\mathbb{I}(k=y_i) - \sigma_k(\boldsymbol{x}_i))\boldsymbol{x}_i$$

同理可得

$$\frac{\partial \sum_{i\in[m]} \ln \sigma_{y_i}(\boldsymbol{x}_i)}{\partial b_k} = \sum_{i\in[m]} \frac{1}{\sigma_{y_i}(\boldsymbol{x}_i)} \frac{\partial \sigma_{y_i}(\boldsymbol{x}_i)}{\partial b_k} = \sum_{i\in[m]} (\mathbb{I}(k=y_i) - \sigma_k(\boldsymbol{x}_i))$$

故最优的 $\boldsymbol{w}_k^\star, b_k^\star, k \in [c]$ 应满足如下的非线性方程组

$$\sum_{i\in[m]} \mathbb{I}(k=y_i)\boldsymbol{x}_i = \sum_{i\in[m]} \sigma_k(\boldsymbol{x}_i)\boldsymbol{x}_i, \quad \sum_{i\in[m]} \mathbb{I}(k=y_i) = \sum_{i\in[m]} \sigma_k(\boldsymbol{x}_i), \quad \forall k \in [c] \tag{1}$$

其中左边是第 k 类样本和 (数)，右边是所有样本加权和 (数)，权重为 softmax 变换给出的概率。式(1)与马尔可夫过程里的细致平稳 (detailed balance) 方程很像，这里我们也称其为平衡方程组。

3 最大熵

最大熵 (maximum entropy) 是对随机事件进行概率分配的一个准则，在满足所有约束的条件下，不再做任何额外假设，也就是对未知的都当作等概率处理，因为此时熵最大。举个简单的例子，给你一个骰子，问掷出 6 的概率是多少，由于你对这个骰子一无所知，只能回答 $1/6$；假如有个先知跟你说这个骰子掷出 1 的概率是 $1/4$，再问你掷出 6 的概率是多少，那你肯定会回答 $3/20$。

平衡方程组是基于 softmax 变换推导出来的，下面证明在满足式(1)约束的条件下，最大熵准则给出的概率变换就是 softmax 变换，即对数几率回归是最大熵准则指导下的线性分类模型。

设 $\pi_k(\boldsymbol{x}_i)$ 为将样本 $\boldsymbol{x}_i$ 预测为第 k 类的概率，其熵为 $-\sum_{k\in[c]} \pi_k(\boldsymbol{x}_i) \ln \pi_k(\boldsymbol{x}_i)$，考虑如下优化问题

$$\begin{aligned}
\max_{\pi_1,\ldots,\pi_c} \ & \sum_{i\in[m]} \left(-\sum_{k\in[c]} \pi_k(\boldsymbol{x}_i) \ln \pi_k(\boldsymbol{x}_i) \right) \\
\text{s.t.} \ & \pi_k(\boldsymbol{x}_i) \geq 0, \ \forall i \in [m], \ \forall k \in [c] \\
& \sum_{k\in[c]} \pi_k(\boldsymbol{x}_i) = 1, \ \forall i \in [m] \\
& \sum_{i\in[m]} \mathbb{I}(k=y_i)\boldsymbol{x}_i = \sum_{i\in[m]} \pi_k(\boldsymbol{x}_i)\boldsymbol{x}_i, \ \forall k \in [c] \\
& \sum_{i\in[m]} \mathbb{I}(k=y_i) = \sum_{i\in[m]} \pi_k(\boldsymbol{x}_i), \ \forall k \in [c]
\end{aligned}$$

其中前两个约束要求 $[\pi_1, \ldots, \pi_c]$ 是一个分布，后两个约束就是平衡方程组。下面先不考虑第一个非负性约束 (事实上它是冗余的)，对后三个约束分别引入拉格朗日乘子 $\alpha_1, \ldots, \alpha_m$、$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_c$、$\gamma_1, \ldots, \gamma_c$ 得到拉格朗日函数

$$L(\pi_k, \alpha_i, \boldsymbol{\beta}_k, \gamma_k) = -\sum_{i\in[m]} \sum_{k\in[c]} \pi_k(\boldsymbol{x}_i) \ln \pi_k(\boldsymbol{x}_i) - \sum_{i\in[m]} \alpha_i \left(\sum_{k\in[c]} \pi_k(\boldsymbol{x}_i) - 1 \right)$$

$$
- \sum_{k \in [c]} \boldsymbol{\beta}_k^\top \left(\sum_{i \in [m]} (\mathbb{I}(k = y_i) - \pi_k(\boldsymbol{x}_i)) \boldsymbol{x}_i \right) - \sum_{k \in [c]} \gamma_k \left(\sum_{i \in [m]} (\mathbb{I}(k = y_i) - \pi_k(\boldsymbol{x}_i)) \right)
$$

对 $\pi_k(\boldsymbol{x}_i)$ 求偏导可得

$$
\frac{\partial L(\pi_k, \alpha_i, \boldsymbol{\beta}_k, \gamma_k)}{\partial \pi_k(\boldsymbol{x}_i)} = -\ln \pi_k(\boldsymbol{x}_i) - 1 - \alpha_i + \boldsymbol{\beta}_k^\top \boldsymbol{x}_i + \gamma_k \Longrightarrow \pi_k(\boldsymbol{x}_i) = \exp(-1 - \alpha_i + \boldsymbol{\beta}_k^\top \boldsymbol{x}_i + \gamma_k)
$$

于是

$$
\begin{aligned}
\sum_{k \in [c]} \exp(-1 - \alpha_i + \boldsymbol{\beta}_k^\top \boldsymbol{x}_i + \gamma_k) = 1 &\Longrightarrow \exp(\alpha_i) = \sum_{k \in [c]} \exp(-1 + \boldsymbol{\beta}_k^\top \boldsymbol{x}_i + \gamma_k) \\
&\Longrightarrow \pi_k(\boldsymbol{x}_i) = \frac{\exp(-1 + \boldsymbol{\beta}_k^\top \boldsymbol{x}_i + \gamma_k)}{\sum_{j \in [c]} \exp(-1 + \boldsymbol{\beta}_j^\top \boldsymbol{x}_i + \gamma_j)} = \frac{\exp(\boldsymbol{\beta}_k^\top \boldsymbol{x}_i + \gamma_k)}{\sum_{j \in [c]} \exp(\boldsymbol{\beta}_j^\top \boldsymbol{x}_i + \gamma_j)}
\end{aligned}
$$

显然这就是 softmax 变换。