对数几率回归

张腾

2026 年 4 月 29 日

对数几率回归 (logisitic regression，LR) 虽然名字中包含“回归”二字，但其本质上是一种广泛使用的线性分类模型。它通过非线性变换将线性模型的输出映射到 $[0,1]$ 使结果具有概率意义。

1 二分类

对于二分类问题，设标记集合 $\mathcal{Y}=\{1,-1\}$。由贝叶斯定理，样本 $\boldsymbol{x}$ 属于正类的条件概率为：

$$\mathbb{P}(y=1|\boldsymbol{x})=\frac{\mathbb{P}(\boldsymbol{x},y=1)}{\mathbb{P}(\boldsymbol{x})}=\frac{\mathbb{P}(\boldsymbol{x},y=1)}{\mathbb{P}(\boldsymbol{x},y=1)+\mathbb{P}(\boldsymbol{x},y=-1)}=1\Big/\left(1+\frac{\mathbb{P}(\boldsymbol{x},y=-1)}{\mathbb{P}(\boldsymbol{x},y=1)}\right) \tag{1}$$

令

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

为正负类联合概率比值的对数，即对数几率 (log-odds)，则式(1)可写为：

$$\mathbb{P}(y=1|\boldsymbol{x})=\frac{1}{1+\exp(-a)}\triangleq\sigma(a)$$

其中 $\sigma:\mathbb{R}\mapsto(0,1)$ 为对数几率函数 (logisitic function)，其将对数几率映射为正类条件概率。LR 的核心假设是：对数几率 a 与输入特征 $\boldsymbol{x}$ 呈线性关系，即 $a=\boldsymbol{w}^\top\boldsymbol{x}+b$。因此，模型预测的概率为：

$$\mathbb{P}(y=1|\boldsymbol{x})=\sigma(\boldsymbol{w}^\top\boldsymbol{x}+b)=\frac{1}{1+\exp(-(\boldsymbol{w}^\top\boldsymbol{x}+b))}$$

$$\mathbb{P}(y=-1|\boldsymbol{x})=\frac{\exp(-(\boldsymbol{w}^\top\boldsymbol{x}+b))}{1+\exp(-(\boldsymbol{w}^\top\boldsymbol{x}+b))}=\frac{1}{1+\exp(\boldsymbol{w}^\top\boldsymbol{x}+b)}=\sigma(-(\boldsymbol{w}^\top\boldsymbol{x}+b))$$

给定数据集 $\mathcal{D}=\{(\boldsymbol{x}_i,y_i)\}_{i\in[m]}\subseteq\mathbb{R}^n\times\{1,-1\}$，我们采用极大似然估计法求解参数 $\boldsymbol{w}$ 和 b：

$$\begin{aligned}
\operatorname*{argmax}_{\boldsymbol{w},b}\ln\mathbb{P}(\mathcal{D}|\boldsymbol{w},b)&=\operatorname*{argmax}_{\boldsymbol{w},b}\ln\prod_{i\in[m]}\mathbb{P}(\boldsymbol{x}_i,y_i|\boldsymbol{w},b) && \text{(IID 假设)}\\
&=\operatorname*{argmax}_{\boldsymbol{w},b}\sum_{i\in[m]}\ln\mathbb{P}(\boldsymbol{x}_i,y_i|\boldsymbol{w},b) && (\ln(a\cdot b)=\ln a+\ln b)\\
&=\operatorname*{argmax}_{\boldsymbol{w},b}\sum_{i\in[m]}(\ln\mathbb{P}(y_i|\boldsymbol{x}_i,\boldsymbol{w},b)+\ln\mathbb{P}(\boldsymbol{x}_i|\boldsymbol{w},b)) && (\mathbb{P}(\boldsymbol{x}_i,y_i)=\mathbb{P}(y_i|\boldsymbol{x}_i)\mathbb{P}(\boldsymbol{x}_i))\\
&=\operatorname*{argmax}_{\boldsymbol{w},b}\sum_{i\in[m]}\ln\mathbb{P}(y_i|\boldsymbol{x}_i,\boldsymbol{w},b) && (\boldsymbol{x}_i\text{ 不依赖于 }\boldsymbol{w}、b\text{，故 }\mathbb{P}(\boldsymbol{x}_i|\boldsymbol{w},b)=\mathbb{P}(\boldsymbol{x}_i))\\
&=\operatorname*{argmax}_{\boldsymbol{w},b}\sum_{i\in[m]}\ln\frac{1}{1+\exp(-y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b))}\\
&=\operatorname*{argmin}_{\boldsymbol{w},b}\sum_{i\in[m]}\ln(1+\exp(-y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)))
\end{aligned}$$

上述目标函数也可从交叉熵的角度来推导：类别标记的独热编码可表示为 $[(1+y)/2,(1-y)/2]$，模型预测分布为 $[\sigma(\boldsymbol{w}^\top\boldsymbol{x}+b),\sigma(-(\boldsymbol{w}^\top\boldsymbol{x}+b))]$，其经验交叉熵损失 ℓ_{CE} 化简后与上述极大似然形式一致：

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

2 多分类

对于多分类问题，设标记集合 $\mathcal{Y}=[c]$。LR 引入 c 个线性判别函数 $\boldsymbol{w}_j^\top\boldsymbol{x}+b_j$，其中 $j\in[c]$，并通过 softmax 函数将输出转化为概率分布：

$$\mathbb{P}(y=k|\boldsymbol{x})=\sigma_k(\boldsymbol{x})=\frac{\exp(\boldsymbol{w}_k^\top\boldsymbol{x}+b_k)}{\sum_{j\in[c]}\exp(\boldsymbol{w}_j^\top\boldsymbol{x}+b_j)}=\frac{\exp((\boldsymbol{w}_k-\boldsymbol{w}_c)^\top\boldsymbol{x}+b_k-b_c)}{\sum_{j\in[c-1]}\exp((\boldsymbol{w}_j-\boldsymbol{w}_c)^\top\boldsymbol{x}+b_j-b_c)+1}\tag{2}$$

其中最后一个等号是因为 softmax 函数的参数存在冗余性，分子、分母同除以某个 $\exp(\boldsymbol{w}_j^\top\boldsymbol{x}+b_j)$ 概率值不变，故通常令 $\boldsymbol{w}_c=\boldsymbol{0}$、$b_c=0$，即 $\exp(\boldsymbol{w}_c^\top\boldsymbol{x}+b_c)=1$ 以消除冗余。当 $c=2$ 时，式(2)即退化为二分类的对数几率函数：

$$\mathbb{P}(y=1|\boldsymbol{x})=\frac{\exp(\boldsymbol{w}_1^\top\boldsymbol{x}+b_1)}{\exp(\boldsymbol{w}_1^\top\boldsymbol{x}+b_1)+1}=\frac{1}{1+\exp(-(\boldsymbol{w}_1^\top\boldsymbol{x}+b_1))}$$

给定数据集 $\mathcal{D}=\{(\boldsymbol{x}_i,y_i)\}_{i\in[m]}\subseteq\mathbb{R}^n\times[c]$，极大似然估计的目标是最大化对数似然：

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

为了求解该优化问题，我们计算其梯度：

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

于是对数似然关于 $\boldsymbol{w}_k$ 的梯度为

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

同理可得对数似然关于 b_k 的梯度为

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

令梯度为零，我们得到最优解应满足的平衡方程组：

$$\sum_{i\in[m]} \mathbb{I}(k=y_i)\boldsymbol{x}_i = \sum_{i\in[m]} \sigma_k(\boldsymbol{x}_i)\boldsymbol{x}_i, \quad \sum_{i\in[m]} \mathbb{I}(k=y_i) = \sum_{i\in[m]} \sigma_k(\boldsymbol{x}_i), \quad \forall k \in [c] \tag{3}$$

式(3)表明：在最优解处，第 k 类样本的特征之和 (及数量) 等于所有样本以 softmax 变换给出的概率为权重的加权和。这与马尔可夫过程中的细致平稳 (detailed balance) 条件形式相似。

3 最大熵视角

最大熵 (maximum entropy) 原理是概率建模的一个重要准则：在满足已知约束的前提下，应选择熵最大的分布，即对未知信息不做任何额外假设 (视为等概率)。例如，对于一个未知骰子，在没有任何信息的情况下，只能假设掷出各数的概率均为 $1/6$；若已知掷出 1 的概率为 $1/4$，则掷出其它数的概率应均分剩余的 $3/4$。

定理 3.1. 对数几率回归是满足平衡方程组(3)约束下的最大熵模型。

证明. 设 $\pi_k(\boldsymbol{x}_i)$ 为模型将样本 $\boldsymbol{x}_i$ 预测为第 k 类的概率，最大化所有样本的预测熵之和：

$$\max_{\pi_1,\dots,\pi_c} \sum_{i\in[m]} \left(-\sum_{k\in[c]} \pi_k(\boldsymbol{x}_i) \ln \pi_k(\boldsymbol{x}_i) \right)$$

$$\text{s.t. } \pi_k(\boldsymbol{x}_i) \geq 0, \ \forall i \in [m], \ \forall k \in [c] \qquad \text{(概率非负)}$$

$$\sum_{k\in[c]} \pi_k(\boldsymbol{x}_i) = 1, \ \forall i \in [m] \qquad \text{(概率归一化)}$$

$$\sum_{i\in[m]} \mathbb{I}(k=y_i)\boldsymbol{x}_i = \sum_{i\in[m]} \pi_k(\boldsymbol{x}_i)\boldsymbol{x}_i, \ \forall k \in [c] \qquad \text{(特征矩匹配)}$$

$$\sum_{i\in[m]} \mathbb{I}(k=y_i) = \sum_{i\in[m]} \pi_k(\boldsymbol{x}_i), \ \forall k \in [c] \qquad \text{(样本数匹配)}$$

引入拉格朗日乘子：α_i 对应概率归一化、$\boldsymbol{\beta}_k$ 对应特征矩匹配、γ_k 对应样本数匹配，构建拉格朗日函数 (暂忽略概率非负约束，因其由指数形式自动满足)：

$$\begin{aligned} L(\pi_k, \alpha_i, \boldsymbol{\beta}_k, \gamma_k) = &- \sum_{i\in[m]} \sum_{k\in[c]} \pi_k(\boldsymbol{x}_i) \ln \pi_k(\boldsymbol{x}_i) - \sum_{i\in[m]} \alpha_i \left(\sum_{k\in[c]} \pi_k(\boldsymbol{x}_i) - 1 \right) \\ &- \sum_{k\in[c]} \boldsymbol{\beta}_k^\top \left(\sum_{i\in[m]} (\mathbb{I}(k=y_i) - \pi_k(\boldsymbol{x}_i))\boldsymbol{x}_i \right) - \sum_{k\in[c]} \gamma_k \left(\sum_{i\in[m]} (\mathbb{I}(k=y_i) - \pi_k(\boldsymbol{x}_i)) \right) \end{aligned}$$

对 π_k 求偏导并令其为零：

$$\frac{\partial L}{\partial \pi_k} = -\ln \pi_k(\boldsymbol{x}_i) - 1 - \alpha_i + \boldsymbol{\beta}_k^\top \boldsymbol{x}_i + \gamma_k = 0$$

解得：

$$\pi_k(\boldsymbol{x}_i) = \exp(-1 - \alpha_i + \boldsymbol{\beta}_k^\top \boldsymbol{x}_i + \gamma_k)$$

利用归一化约束 $\sum_{k\in[c]}\pi_k(\boldsymbol{x}_i)=1$ 消去 α_i：

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

这正是 softmax 函数的形式。 ♠