Logistic Regression 推导

Soul Walker

2020 年 7 月 22 日

1 Logistic 回归模型

对于我们想要得到一个类别的概率的问题，我们需要能够输出 [0, 1] 区间的值的函数，若这里我们考虑二分类模型，我们如果想要用贝叶斯定理对 $p(C|\mathbf{x})$ 建立模型, C_1，C_2 分别对应两个类别

$$p(C_1|\mathbf{x}) = \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_1)p(C_1) + p(\mathbf{x}|C_2)p(C_2)} \tag{1}$$

这里我们取 $\theta = ln\frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)}$，那么式 (1) 就可以作出如下转换

$$p(C_1|\mathbf{x}) = \frac{1}{1 + exp(-\theta)} \tag{2}$$

也就得出了 Logistic Sigmoid 函数 ($\sigma(\theta) = \frac{1}{1+e^{-\theta}}$)，参数 θ 表示了两个类联合概率比值的对数，我们不关心这个参数的具体值，模型假设对 θ 进行，那么 Logistic 回归的模型假设就是

$$\theta = w^T\mathbf{x} \tag{3}$$

还可以有一种理解，如果我们想要用线性回归问题来映射出线性分类问题，那么我们需要在线性回归问题 ($w^T\mathbf{x}$) 之后加上一个激活函数 (Activation Function)，比如这里的 Sigmoid 函数。Principal Components Analysis 我们的问题就转换为在这个模型下寻找 w 的最佳值以得到最佳模型，这里因为是概率判别模型我们最常用最大似然估计来对参数进行估计

输入：训练数据集

$$T = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N))\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$$

其中，$\mathbf{x}_i \in \chi \in \mathbb{R}^N$ 为实例的特征向量，$\mathbf{x}_i = (\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, ..., \mathbf{x}_i^{(n)})^T$ ，$y_i \in \{0, 1\}$ 为实例的类别，$i = 1, 2, \cdots, N$ ，实例特征向量 $\mathbf{x}$，$\mathbf{X}$ 为训练样本集，形状为 (N, p)，$\mathbf{Y}$ 是训练标签集，形状为 $(N, 1)$。

输出：实例 $\mathbf{x}$ 所属的类别 y。

那么我们可以得到分类 y 的概率如下

$$\begin{aligned} p_1 &= p(y = 1|\mathbf{x}) = \sigma(w^T\mathbf{x}) = \frac{1}{1 + e^{-w^T\mathbf{x}}} \\ p_0 &= 1 - p(y = 1|\mathbf{x}) = 1 - \frac{1}{1 + e^{-w^T\mathbf{x}}} = \frac{e^{-w^T\mathbf{x}}}{1 + e^{-w^T\mathbf{x}}} \end{aligned} \tag{4}$$

对上式综合可得

$$p(y|\mathbf{x}) = p_1^y p_0^{1-y} \tag{5}$$

2 Logistic 回归模型参数估计

此处对于 N 次独立全同的观测最大似然估计得

$$\begin{aligned}
\hat{w} &= \arg\max_w J(w) \\
&= \arg\max_w logp(\mathbf{Y}|\mathbf{X}) \\
&= \arg\max_w log\prod_{i=1}^{N} p(y_i|\mathbf{x}_i) \\
&= \arg\max_w \sum_{i=1}^{N} logp(y_i|\mathbf{x}_i) \\
&= \arg\max_w \sum_{i=1}^{N} (y_i logp_1 + (1-y_i)logp_0)
\end{aligned} \tag{6}$$

我们可以看到 $y_i logp_1 + (1-y_i)logp_0$ 是交叉熵 (cross entropy) 表达式的相反数

我们可以对 $J(w)$ 做求导

$$J'(w) = \sum_{i=1}^{N} y_i(1-p_1)\mathbf{x}_i - p_1\mathbf{x}_i + y_i p_1 \mathbf{x}_i = \sum_{i=1}^{N}(y_i - p_1)\mathbf{x}_i \tag{7}$$

由于概率值的非线性，放在求和符号中时，这个式子无法直接求解。我们可以使用不同大小的批量随机梯度上升（对于最小化就是梯度下降）来获得这个函数的极大值。这里我们直接选择解决多分类问题如下

这里我们选择使用 $\theta = w^T\mathbf{x} + b$ 的回归模型

$$\begin{aligned}
L(w,b) &= -\frac{1}{N}\sum_{i=1}^{N}[y_i logp_i + (1-y_i)log(1-p_i)] \\
&= -\frac{1}{N}\sum_{i=1}^{N}[y_i(w^T\mathbf{x}_i + b) - log(1+e^{w^T\mathbf{x}_i+b})]
\end{aligned} \tag{8}$$

假设 η 为学习率

$$\begin{aligned}
\frac{\partial L(w,b)}{\partial w} &= -\frac{1}{N}\sum_{i=1}^{N}(y_i\mathbf{x}_i - \frac{x_i e^{w^T\mathbf{x}_i+b}}{1+e^{w^T\mathbf{x}_i+b}}) \\
&= -\frac{1}{N}\sum_{i=1}^{N}(y_i - \frac{e^{\mathbf{x}^T\mathbf{x}_i+b}}{1+e^{w^T\mathbf{x}_i+b}})\mathbf{x}_i \\
&= -\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)\mathbf{x}_i
\end{aligned} \tag{9}$$

$$\begin{aligned}
\frac{\partial L(w,b)}{\partial b} &= -\frac{1}{N}\sum_{i=1}^{N}(y_i - \frac{e^{w^T\mathbf{x}_i+b}}{1+e^{w^T\mathbf{x}_i+b}}) \\
&= -\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)
\end{aligned} \tag{10}$$

$$\begin{aligned}
w &= w + \eta(y_i - \hat{y}_i)\mathbf{x}_i \\
b &= b + \eta(y_i - \hat{y}_i)
\end{aligned} \tag{11}$$