1 分类与逻辑回归

1.1 多分类问题

考虑这样一个分类问题，响应值y可以为指定的k个值中的任意一个，即$y \in \{1, 2, \ldots, k\}$。举个例子，我们可能想要将邮件划分为三种类型，比如垃圾邮件、个人邮件及工作邮件，而不是只划分为垃圾邮件与非垃圾邮件（这是一个二分类问题）。标签或响应值仍然是离散的，但可以取两个以上的值。因此我们将使用多项式分布对其进行建模。

在这种情况下，$p(y|x;\theta)$是基于k个可能的离散值的分布，因此是一个多项式分布。对于包含k个值的多项式分布，$\phi_1, \ldots, \phi_k$表示每一种可能的概率，必须满足约束$\sum_{i=1}^{k} \phi_i = 1$。我们将设计一个参数化模型，在给出输入$x$的前提下，输出满足这个约束的$\phi_1, \ldots, \phi_k$。

我们引入k组参数$\theta_1, \ldots, \theta_k$，每一组参数都是空间$\mathbb{R}^d$中的一个向量。根据直觉，我们应该可以使用$\theta_1^T x, \ldots, \theta_k^T x$来表示$\phi_1, \ldots, \phi_k$，即概率$P(y=1|x;\theta), \ldots, P(y=k|x;\theta)$。然而，采用这种直接的办法有两个问题，首先，$\theta_j^T x$不一定在$[0,1]$内，其次，$\sum_{j=1}^{k} \theta_j^T x$不一定为1。因此，我们将使用softmax函数将向量$(\theta_1, \ldots, \theta_k)$转化为每个元素都是非负的并且和为1的概率向量。

定义softmax函数$\mathrm{softmax}: \mathbb{R}^{\mathrm{k}} \to \mathbb{R}^{\mathrm{k}}$为

$$\mathrm{softmax}(\mathrm{t}_1, \ldots, \mathrm{t}_\mathrm{k}) = \begin{bmatrix} \frac{\exp(t_1)}{\sum_{j=1}^{k} \exp(t_j)} \\ \vdots \\ \frac{\exp(t_k)}{\sum_{j=1}^{k} \exp(t_j)} \end{bmatrix} \tag{1.1}$$

softmax函数的输入，向量t一般被称为logits，在定义中，softmax函数的输出必为每个元素都是非负的并且和为1的概率向量。

令$(t_1, \ldots, t_k) = (\theta_1^T x, \ldots, \theta_k^T x)$，我们将$(t_1, \ldots, t_k)$作为softmax函数的输入，将softmax函数的输出作为概率$P(y=1|x;\theta), \ldots, P(y=k|x;\theta)$，我们得到如下概率模型：

$$\begin{bmatrix} P(y=1|x;\theta) \\ \vdots \\ P(y=k|x;\theta) \end{bmatrix} = \mathrm{softmax}(\mathrm{t}_1, \ldots, \mathrm{t}_\mathrm{k}) = \begin{bmatrix} \frac{\exp(\theta_1^T x)}{\sum_{j=1}^{k} \exp(\theta_j^T x)} \\ \vdots \\ \frac{\exp(\theta_k^T x)}{\sum_{j=1}^{k} \exp(\theta_j^T x)} \end{bmatrix} \tag{1.2}$$

为了表示方便，我们令$\phi_i = \frac{\exp(\theta_i^T x)}{\sum_{j=1}^{k} \exp(\theta_j^T x)}$，上面等式可以简写为：

$$P(y=i|x;\theta) = \phi_i = \frac{\exp(t_i)}{\sum_{j=1}^{k} \exp(t_j)} = \frac{\exp(\theta_i^T x)}{\sum_{j=1}^{k} \exp(\theta_j^T x)} \tag{1.3}$$

接下来，我们计算一个样例(x, y)的负对数-似然（log-likehood）。

$$-\log P(y|x,\theta) = -\log\left(\frac{\exp(t_y)}{\sum_{j=1}^{k} \exp(t_j)}\right) = -\log\left(\frac{\exp(\theta_y^T x)}{\sum_{j=1}^{k} \exp(\theta_j^T x)}\right) \tag{1.4}$$

因此，训练数据的负对数-似然，即损失函数可以写为：

$$\ell(\theta) = \sum_{i=1}^{n} -\log\left(\frac{\exp(\theta_{y^{(i)}}^T x^{(i)}}{\sum_{j=1}^{k} \exp(\theta_j^T x^{(i)})}\right) \tag{1.5}$$