第 2 章　逻辑回归的训练

机器学习模型的训练是指根据“训练集”寻找最优模型参数的过程。训练集是指从现实样本分布——总体（population）中采样的包含类别信息的样本集合。

本章首先介绍模型训练的一般概念和模型评价的若干指标，之后探讨二分类问题的损失函数。特别是从衡量两个分布相似程度的 K-L 散度和最大似然两种角度阐述交叉熵损失函数的含义。

经由损失函数，模型训练问题归约成了以模型参数为自变量，在自变量空间中寻找损失函数最小值的函数优化问题。本章在回顾多元函数微积分的相关知识后介绍梯度下降法及其各种变体。

阅读本章后，读者应当理解了机器学习模型训练的原理和评价模型的方法。本章虽是在逻辑回归框架下进行讲解，但所有概念都可以直接用于神经网络和深度学习。

2.1　训练集与测试集

第 1 章已经介绍，给定权值向量$\boldsymbol{w} = (w_1 \quad w_2 \quad \cdots \quad w_n)^{\mathrm{T}}$和偏置值$b$，对于样本$\boldsymbol{x} = (x_1 \quad x_2 \quad \cdots \quad x_n)^{\mathrm{T}}$，逻辑回归模型预测其为 A 类的概率是：

$$f(x) = \frac{1}{1+e^{-b-\boldsymbol{w}^{\mathrm{T}}x}} \qquad (2.1)$$

公式（2.1）中的$\boldsymbol{w}$和b就是逻辑回归模型的参数（parameters）。所谓“训练”（training）就是寻找参数$\boldsymbol{w}$和b的值，使得模型可以很好地区分 A 类和 B 类样本。训练过程需要“训练集”（training set）。训练集由一批带类别信息的样本组成。这些样本是从现实中采样的属于 A 类或 B 类的样本。类别信息用一个实数y表示，例如用$y = 1$表示样本属于 A 类；$y = 0$表示样本属于 B 类。$\boldsymbol{y}$值称为标签（label）。标签的 1/0 编码只是方法的一种，还可以采用别的编码，例如 1/-1。后文会看到不同编码有不同的用途。于是训练集是如公式（2.2）描述的集合。

$$S = \{\boldsymbol{x}^i, y^i\}_{i=1}^{m} \qquad (2.2)$$

上标表示样本的编号。训练集 S 中一共包含 m 个样本。其中每一个$\boldsymbol{x}^i \in \mathbb{R}^n$是样本特征，$y^i \in \{0,1\}$是样本标签。

为了评价模型的表现，有必要取另一份带标签的样本集 T，称为测试集（test set）。在测试

集上对训练完成的模型进行评价才能得到客观无偏的评价指标。第 3 章“正则化”会介绍模型自由度、过拟合、偏置-方差平衡等概念。届时会阐述必须在独立的测试集上评价模型的原因。

2.2 分类模型的评价

对于已经训练完成的逻辑回归模型，可以在测试集 T 上评价它的表现。第 1 章曾提到：对于一个样本$\boldsymbol{x}$，逻辑回归给出的是它属于 A 类的概率$p_A(\boldsymbol{x})$。人有主动权选定一个阈值 t，当$p_A(\boldsymbol{x}) \geq t$时将$\boldsymbol{x}$判定为属于 A 类，否则判定$\boldsymbol{x}$属于 B 类：

$$\hat{y}^i = \begin{cases} 0, & p_A(\boldsymbol{x}^i) < t \\ 1, & p_A(\boldsymbol{x}^i) \geq t \end{cases} \qquad (2.3)$$

对训练集 T 中的所有样本$\boldsymbol{x}^i$计算逻辑回归模型的输出$p_A(\boldsymbol{x}^i)$。一旦选定了阈值 t，根据公式（2.3）就可以得出模型对每一个样本$\boldsymbol{x}^i$所判定的类别$\hat{y}^i$。和y^i一样，$\hat{y}^i$用 1/0 编码 A/B 类别。^符号表示$\hat{y}^i$是模型预测的标签，与训练集的标签y^i区分。

有了这份判定结果，可以绘制模型的混淆矩阵（confusion matrix）：

	预测 B 类	预测 A 类
真实 B 类	TN	FP
真实 A 类	FN	TP

表 2.1 二分类问题的混淆矩阵

对于二分类问题，混淆矩阵是一个2×2矩阵。从左上到右下每一个元素分别是：

- TN（True Negative）：真实为 B 类且模型判定为 B 类的样本个数；
- FP（False Positive）：真实为 B 类但模型判定为 A 类的样本个数（被错误地判定为 A 类）；
- FN（False Negative）：真实为 A 类但模型判定为 B 类的样本个数（被错误地判定为 B 类）；
- TP（True Positive）：真实为 A 类且模型判定为 A 类的样本个数。

评价模型表现的几个常用指标（metrics）如下：

$$accuracy = \frac{\mathrm{TN+TP}}{\mathrm{TN+FP+FN+TP}} \qquad (2.4)$$

正确率 *accuracy* 是混淆矩阵的对角线元素之和除以全体元素之和。它是模型正确分类的样本个数与全部样本个数之比。有时正确率并非一个良好的评价指标。假如测试集中 A 类样本和 B 类样本的数量比为 99:1，那么模型将所有样本判定为 A 类就能够得到 99%的正确率，但是该模型显然不是一个好模型。

$$precision_A = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \qquad (2.5)$$

A 类查准率$precision_A$是混淆矩阵右下角元素除以第二列元素之和。它是模型正确判定为 A 类的样本数量与全部判定为 A 类的样本数量之比。$precision_A$评价模型判定为 A 类的准确程度。$precision_A$越高则模型的断言越可靠。

$$recall_A = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \qquad (2.6)$$

A 类查全率$recall_A$是混淆矩阵右下角元素除以第二行元素之和。它是模型判断为 A 类的样本数量与全部 A 类样本数量之比。$recall_A$评价模型对 A 类的召回情况。$recall_A$越高则模型能把更多的 A 类样本识别出来。$recall_A$又称真阳率（TPR_A，True Positive Rate）。与之对应还有假阳率（FPR_A，False Positive Rate）：

$$FPR_A = \frac{\mathrm{FP}}{\mathrm{FP+TN}} \qquad (2.7)$$

FPR_A是所有 B 类样本中被模型错判成 A 类的比例。它越高则模型表现越差。$precision_A$，$recall_A$，FPR_A/TPR_A都可以针对 B 类计算。基于混淆矩阵还有其他评价指标，但最常用的是上述几个。

所有这些指标都基于分类结果，而分类结果依赖于概率阈值 *t*。*t* 是可人为调节的。假如 *t* 设得较低，可以想象低门槛将导致更多的样本被判定为 A 类，$recall_A/TPR_A$会较高。但同时也会把更多 B 类样本错判为 A 类，从而抬高FPR_A，拉低$precision_A$。反之，若 *t* 设得较高，$recall_A/TPR_A/FPR_A$将降低，$precision_A$将升高。所以，选择合适的阈值是在模型两种相反的倾向中进行权衡。权衡的准则依据具体问题的需要。

上述论述可知FPR_A/TPR_A这对指标随着 *t* 值变化同进退，一个升高另一个也升高。高TPR_A是

我们愿意看到的，而高FPR_A是我们希望避免的。我们希望在提高TPR_A的同时不要大幅度地提高FPR_A。FPR_A/TPR_A随着 t 的变化而变化的行为，可由模型的 ROC（receiver operating characteristic）曲线来表现。如图 2-1 所示。

图 2-1 ROC 曲线

ROC 曲线以FPR_A为横轴，以TPR_A为纵轴，将不同 t 值对应的FPR_A/TPR_A以散点的形式画在坐标系内。得到的图形是一条拱起的曲线。ROC 曲线上拱得越高，说明在较低的FPR_A水平能够得到较高的TPR_A。于是 ROC 曲线下的面积（Area Under Curve，AUC）可被用来衡量模型的质量。AUC 越大，ROC 曲线越上拱，模型的表现更优。AUC 不依赖于阈值 t 的选择，是一个全面衡量模型质量的指标。

我们希望训练得到的模型在测试集上有较优的评价，但是无法用测试集上的指标来指导模型参数的选择。因为评价指标不是模型参数的连续函数。参数在空间中的极微小位移会导致模型输出概率p_A的极微小变化。当这个微小变化不足以使p_A跨越阈值 t 时，模型对样本的分类不发生改变，上述各种评价指标也就不变。而一旦某个微小位移导致了p_A跨越阈值 t，各个指标将发生跳跃式变化。模型参数和评价指标之间缺乏一个显式的连续的映射，使我们无法利用评价指标来调整模型参数。

2.3 损失函数

存在一些非参数优化方法，例如遗传算法等。它们不依赖模型参数和评价指标之间的显式连续映射。但是这类的效率和资源占用是巨大的。在模型训练中，我们需要采用一种“代理”评价指标。它应该是一个关于模型参数的显式连续函数。这种“代理”评价指标称为损失函数（loss function）。损失函数以某种方式衡量模型的质量。模型的训练问题就变成了在参数空间中寻找损失函数最小值的问题。

损失函数有很多种，本书只介绍分类问题中最常用的交叉熵（cross entropy）损失函数。我们将从信息论和贝叶斯两种视角阐释交叉熵损失函数的含义。

2.3.1 K-L 散度与交叉熵

随机变量X有k种不同的取值：$x_1, x_2, \dots, x_k$。令X取x_i的概率为$p(\mathrm{X} = x_i)$，简写作$p(x_i)$。将X看作一个信号源，观察到$\mathrm{X} = x_i$就相当于收到了一条信息。克劳德·香农为一条信息的信息量做了定量定义：

$$I(\mathrm{X} = x_i) = log\frac{1}{p(x_i)} = -log^{p(x_i)} \qquad (2.8)$$

公式（2.8）中的对数log可以取以 2 为底，也可以取其他底，比如自然对数的底e。取不同的底计算出的信息量之间差一个常系数，不构成影响。如果以 2 为底，信息量的单位是比特（bit）。$I(\mathrm{X} = x_i)$ 称为$\mathrm{X} = x_i$这条信息的自信息量（self-information）。

$I(\mathrm{X} = x_i)$随着$p(x_i)$变化的图像如图 2-2 所示。$p(x_i)$趋向于 1 时，$I(\mathrm{X} = x_i)$趋向于 0；$p(x_i)$趋向于 0 时，$I(\mathrm{X} = x_i)$趋向于正无穷。这是出于这样的考虑：信息所告知的事件的概率越小，则这条信息的信息量越大。假如有人告诉你：“即将开奖的彩票中奖号码是 31415926”。这条信息非常有用，你愿意花大价钱购买它。假如有人告诉你：“明天太阳照常升起”。这条信息几乎是无用的。你不用别人告诉也知道明天太阳几乎肯定照常升起。前一条信息所告知的事件的概率极小，所以信息量很大；后一条信息所告知的事件的概率极大，所以信息量很小。

图 2-2　自信息量的图像

令信息源X取不同的取值$x_1, x_2, \dots, x_k$的概率分别为$p(x_1)$，$p(x_2)$，…，$p(x_k)$。定义信息源的信息容量为：

$$H(p) = \sum_{i=1}^{k} p(x_k)\, log\, \frac{1}{p(x_k)} = -\sum_{i=1}^{k} p(x_k) log^{p(x_k)} \qquad (2.9)$$

$H(p)$又被称为信息源的熵（entropy）。由于信息源由分布p描述，故将$H(p)$视为p的函数。熵是来自热力学的概念。$H(p)$又被称作平均自信息。因为$H(p)$是对X的所有取值以概率为权重计算加权平均。换句话说$H(p)$是$log\ 1/p(X)$在分布p上的期望（exception）。公式（2.9）是针对离散型随机变量的情况。对连续型随机变量的情况应以积分取代求和。

设有两个分布p和q，定义p与q的 K-L 散度（Kullback-Leibler Divergence）是：

$$KLD(p||q) = \sum_{i=1}^{k} p(x_k)\, log\, \frac{p(x_k)}{q(x_k)} = -\sum_{i=1}^{k} p(x_k) log^{q(x_k)} - H(p) \qquad (2.10)$$

根据公式（2.10），K-L 散度是$log\ p/q$在分布p上的期望。注意$KLD(p||q) \neq KLD(q||p)$。如果对于所有i有$p(x_i) = q(x_i)$，则$log\ p(x_i)/q(x_i) = 0$。这时$KLD(p||q) = 0$。也就是说：两个相同分布p和q的 K-L 散度为 0。K-L 散度用来衡量两个分布之间的差异程度。

注意公式（2.10）的第二步。将右边第一项定义为分布p和q的交叉熵（cross entropy）：

$$H(p, q) = -\sum_{i=1}^{k} p(x_k) log^{q(x_k)} \qquad (2.11)$$

$H(p,q)$是$log^{q(X)}$在分布p上的期望。根据公式（2.10）和（2.11），有：

$$H(p,q) = KLD(p||q) + H(p) \qquad (2.12)$$

分布p和q的交叉熵等于p和q的 K-L 散度加上p的熵。如果令分布p固定，则$H(p,q)$与$KLD(p||q)$之间相差一个常数$H(p)$。于是$H(p,q)$也可用来衡量p和q的差异程度：$H(p,q)$越小则p和q越相似。这就是交叉熵损失函数的思想。

对于一个带标签的训练样本$\{\boldsymbol{x}^i, y^i\}$，可以认为它给出了一个$\boldsymbol{x}^i$属于 A/B 类的分布：

$$p(\boldsymbol{x}^i \in \mathrm{A}) = y^i, p(\boldsymbol{x}^i \in \mathrm{B}) = 1 - y^i \qquad (2.13)$$

当$\boldsymbol{x}^i$属于 A 类时$y^i = 1$，该分布就是$p(\boldsymbol{x}^i \in \mathrm{A}) = 1, p(\boldsymbol{x}^i \in \mathrm{B}) = 0$；当$\boldsymbol{x}^i$属于 B 类时$y^i = 0$，该分布就是$p(\boldsymbol{x}^i \in \mathrm{A}) = 0$，$p(\boldsymbol{x}^i \in \mathrm{B}) = 1$。这是一个“确定”的分布。前文说过，逻辑回归模型的输出也是一个分布：

$$q(\boldsymbol{x}^i \in \mathrm{A}) = \frac{1}{1+\mathrm{e}^{-b-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^i}},\ q(\boldsymbol{x}^i \in \mathrm{B}) = \frac{1}{1+\mathrm{e}^{b+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^i}} \qquad (2.14)$$

将训练样本给出的真实分布p和模型给出的分布q的交叉熵作为模型在样本$\boldsymbol{x}^i$上的损失：

$$loss(\boldsymbol{w}, b|\boldsymbol{x}^i) = -y^i log\frac{1}{1+e^{-b-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^i}} - (1-y^i) log\frac{1}{1+e^{b+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^i}} \qquad (2.15)$$

$loss(\boldsymbol{w}, b|\boldsymbol{x}^i)$较大表示模型给出的分布与真实分布之间的差异度较大。反之亦然。公式（2.15）是在模型一个训练样本上的损失。模型在整个训练集上的损失就是在所有样本上的损失的平均：

$$loss(\boldsymbol{w}, b) = \frac{1}{m}\sum_{i=1}^{m}\left(-y^i log\frac{1}{1+e^{-b-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^i}} - (1-y^i) log\frac{1}{1+e^{b+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^i}}\right) \qquad (2.16)$$

公式（2.16）就是交叉熵损失函数。它与训练集有关，以模型参数$\boldsymbol{w}$和b为自变量。给定一

套参数就得到一个损失函数值$loss(\boldsymbol{w},b)$。逻辑回归模型的训练就是寻找使$loss(\boldsymbol{w},b)$尽可能小的$\boldsymbol{w}$和b。这就将模型训练问题转化成为一个函数优化问题。

交叉熵作为“代理”评价指标，与 2.2 节介绍的各种评价指标没有直接的、显式的关系。但通过最小化交叉熵，我们拉近了模型的预测类别分布与训练样本的真实类别分布之间的“距离”。经过训练，可以期待模型抓住数据背后的分布规律，从而在测试集上获得较好的效果。

2.3.2 最大似然估计

本节从最大似然估计的视角阐释交叉熵损失的含义。用X和Y表示随机变量，贝叶斯公式（Bayes Rule）是：

$$p(\mathrm{X}=x|\mathrm{Y}=y)=\frac{p(\mathrm{Y}=y|\mathrm{X}=x)p(\mathrm{X}=x)}{p(\mathrm{Y}=y)} \qquad (2.17)$$

公式(2.17)左边的$p(\mathrm{X}=x|\mathrm{Y}=y)$称为后验概率(posterior probability)。它是观察到事件$\mathrm{Y}=y$的前提下，事件$\mathrm{X}=x$发生的概率。等式右边分子上的第二项$p(\mathrm{X}=x)$称为先验概率（prior probability）。它是事件$\mathrm{X}=x$发生的概率。分子上第一项$p(\mathrm{Y}=y|\mathrm{X}=x)$称为似然概率（likelihood）。它是事件$\mathrm{X}=x$发生的前提下，事件$\mathrm{Y}=y$发生的概率。分母是事件$\mathrm{Y}=y$发生的边缘概率（Marginal Probability）：

$$p(\mathrm{Y}=y)=\sum_x p(\mathrm{Y}=y,\mathrm{X}=x) \qquad (2.18)$$

由于是对X的所有可能取值求和，所以$p(\mathrm{Y}=y)$与X无关。公式（2.17）的证明很简单：将右边的分母乘到左边，根据条件概率的定义，等号两边都是$p(\mathrm{X}=x，\mathrm{Y}=y)$——事件$\mathrm{X}=x$和$\mathrm{Y}=y$同时发生的概率。

回到逻辑回归的语境下，事件X是模型参数为特定值$\boldsymbol{w}$和b，事件Y是观察到训练集S。代入贝叶斯公式：

$$p(\boldsymbol{w},b|S)=\frac{p(S|\boldsymbol{w},b)p(\boldsymbol{w},b)}{p(S)} \qquad (2.19)$$

观察到训练集S前提下参数值为$\boldsymbol{w}$和b的后验概率，等于参数值是$\boldsymbol{w}$和b的先验概率乘以参数值是$\boldsymbol{w}$和b前提下观察到训练集S的似然概率，再除以观察到训练集S的概率。

模型训练的目标是寻找观察到训练集S前提下可能性最大的参数值，也就是使后验概率最大

的参数值：

$$\boldsymbol{w}^*, b^* = argmax_{\boldsymbol{w},b}\ p(\boldsymbol{w}, b|S) \qquad (2.20)$$

假设先验分布$p(\boldsymbol{w}, b)$是均匀的，与参数取值无关，那么问题转化为寻找似然概率最大的参数值：

$$\boldsymbol{w}^*, b^* = argmax_{\boldsymbol{w},b}\ p(S|\boldsymbol{w}, b) \qquad (2.21)$$

$\boldsymbol{w}^*, b^*$称为最大似然估计（Maximum Likelihood Estimate，MLE）。

对于一个训练样本$\{\boldsymbol{x}^i, y^i\}$，$y^i$用 1 或 0 标识样本属于 A 类或 B 类。逻辑回归模型预测$\boldsymbol{x}^i$属于y^i所标识的类别的概率是：

$$p(y^i|\boldsymbol{w}, b, \boldsymbol{x}^i) = \left(\frac{1}{1+e^{-b-\boldsymbol{w}^{\mathrm{T}}x^i}}\right)^{y^i} \cdot \left(\frac{1}{1+e^{b+\boldsymbol{w}^{\mathrm{T}}x^i}}\right)^{1-y^i} \qquad (2.22)$$

公式（2.22）的技巧是利用任何数的 0 次方都等于 1 的事实，根据y^i是 1 还是 0 选择$\boldsymbol{x}^i \in$ A或$\boldsymbol{x}^i \in$ B的概率。假设训练集合样本是独立的，可以得到：

$$p(S|\boldsymbol{w}, b) = \prod_{i=1}^{m} p(y^i|\boldsymbol{w}, b, \boldsymbol{x}^i) \qquad (2.23)$$

因为log是单调递增的。寻找公式（2.21）的$\boldsymbol{w}^*$和b^*等价于寻找：

$$\boldsymbol{w}^*, b^* = argmax_{\boldsymbol{w},b}\ log^{p(S|\boldsymbol{w},b)} \qquad (2.24)$$

根据公式（2.22）、（2.23）和（2.24），目的是寻找$\boldsymbol{w}^*$和b^*使（2.25）最大化：

$$log^{p(S|\boldsymbol{w},b)} = log^{\prod_{i=1}^{m} p(y^i|\boldsymbol{w},b,\boldsymbol{x}^i)} = \sum_{i=1}^{m} y^i log\frac{1}{1+e^{-b-\boldsymbol{w}^{\mathrm{T}}x^i}} + \left(1-y^i\right) log\frac{1}{1+e^{b+\boldsymbol{w}^{\mathrm{T}}x^i}} \qquad (2.25)$$

最大化$log^{p(S|\boldsymbol{w},b)}$等价于最小化$-log^{p(S|\boldsymbol{w},b)}$，于是最大似然估计就是寻找参数值使（2.26）

最小化：

$$-log^{p(S|\boldsymbol{w},b)} = -\sum_{i=1}^{m}\left(y^i log\frac{1}{1+e^{-b-\boldsymbol{w}^{\mathrm{T}}x^i}} + \left(1-y^i\right)log\frac{1}{1+e^{b+\boldsymbol{w}^{\mathrm{T}}x^i}}\right) \quad (2.26)$$

除了一个常系数$1/m$，公式（2.26）和公式（2.16）是相同的。所以使交叉熵损失函数（2.16）最小的$\boldsymbol{w}^*$和b^*就是最大似然估计。

最大似然估计使似然概率最大化。但其实我们想要的是最大化后验概率。在假设模型参数的先验分布是均匀的前提下，此二者等价。在第 3 章“正则化”中我们将看到，为损失函数加上“正则化项”相当于取一个参数先验分布，然后最大化后验概率。正则化的强度与先验分布的方差有关。