

第 2 章 模型评价与损失函数

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，之后探讨分类问题的损失函数。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。本章主要介绍交叉熵损失函数，并从 K-L 散度和最大似然估计两种角度阐释交叉熵损失函数的原理。

阅读本章后，读者应当掌握了机器学习模型训练的一般概念和评价模型的方法，对交叉熵损失函数有了较深刻的理解。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 $\mathbf{w} = (w_1 \ w_2 \ \dots \ w_n)^T$ 和偏置值 b ，对于样本 $\mathbf{x} = (x_1 \ x_2 \ \dots \ x_n)^T$ ，逻辑回归模型预测其为 A 类的概率是：

$$f(\mathbf{x}) = \frac{1}{1+e^{-b-\mathbf{w}^T\mathbf{x}}} \quad (2.1)$$

式 (2.1) 中的 \mathbf{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 \mathbf{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。

训练样本的类别信息用一个实数 y 标识。例如，用 $y = 1$ 标识样本属于 A 类；用 $y = 0$ 标识样本属于 B 类。 y 值称为标签 (label)。标签的 1/0 编码只是方法的一种，还可以采用其他编码。后文会看到不同编码的用途。

训练集是如式 (2.2) 描述的集合。

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

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 $\mathbf{x}^i \in \mathbb{R}^n$ 是样本特征向量， $y^i \in \{0,1\}$ 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）。只有在测试集上对模型进行评价，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 \mathbf{x} ，逻辑回归给出的是它属于 A 类的概率 $p(\mathbf{x} \in A)$ 。人有主动权选定一个阈值 t ，当 $p(\mathbf{x} \in A) \geq t$ 时将 \mathbf{x} 判定为属于 A 类，否则判定 \mathbf{x} 属于 B 类：

$$\hat{y} = \begin{cases} 0, & p(\mathbf{x} \in A) < t \\ 1, & p(\mathbf{x} \in A) \geq t \end{cases} \quad (2.3)$$

\hat{y} 符号表示 \hat{y} 是模型预测的标签，与训练样本中的标签区分。对测试集 T 中的所有样本 $\mathbf{x}^i \in T$ 计算 $p(\mathbf{x} \in A)$ 。一旦选定了阈值 t ，根据式（2.3）就可以得出模型对每一个样本 \mathbf{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。

2.2.1 混淆矩阵

有了模型分类结果就可以绘制混淆矩阵（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

二分类问题的混淆矩阵是一个 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。

评价模型的指标都可以由混淆矩阵计算得出。下文介绍最常用的几个。

2.2.2 正确率

正确率（accuracy）的计算公式为：

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

正确率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。

2.2.3 查准率

查准率又称准确率（precision），其计算公式为：

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

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。同样也有 B 类查准率。

2.2.4 查全率

查全率又称召回率（recall），其计算公式为：

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

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）：

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

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。查全率，真阳率和假阳率也都可以对 B 类计算。

上述指标都基于分类结果，而分类结果依赖于可人为调节的概率阈值 t 。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， $recall_A/TPR_A$ 会升高。但同时也会把更多 B 类样本错判为 A 类，从而抬高 FPR_A ，降低 $precision_A$ 。反之，若 t 设得较高， $recall_A/TPR_A/FPR_A$ 会降低，但 $precision_A$ 将升高。所以选择阈值是对模型两种相反的倾向做权衡。依据是具体问题的需要。

2.2.5 ROC 曲线

FPR_A/TPR_A 这对指标随着 t 值变化同升同降。对于识别 A 类的问题来说，高 TPR_A 是我们愿意看到的，而高 FPR_A 是希望避免的。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 的变化行为可由 ROC (receiver operating characteristic) 曲线刻画。如图 2-1 所示。

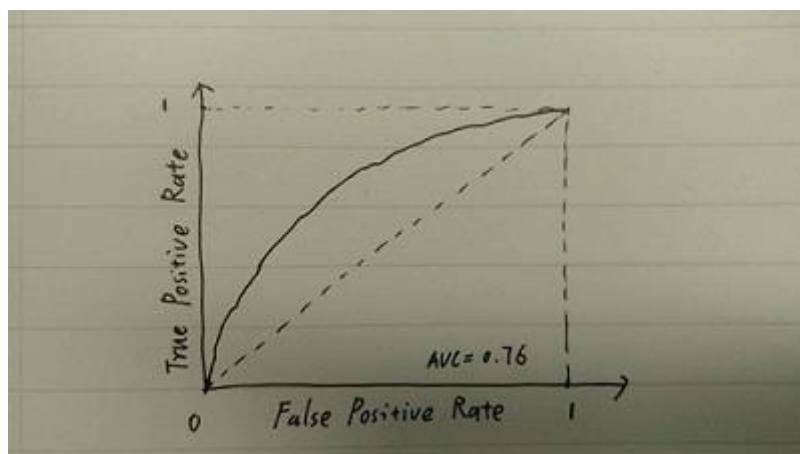


图 2-1 ROC 曲线

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 值对以散点的形式绘出。得到的图形是一条拱起的曲线。ROC 曲线上拱得高说明在较低的 FPR_A 水平能够得到较高的 TPR_A 。ROC 曲线下的面积 (Area Under Curve, AUC) 可以衡量模型的质量。高 AUC 意味着 ROC 曲线上拱，模型的表现更优。AUC 不依赖阈值 t ，是一个全面衡量模型质量的指标。

我们希望模型在测试集上有较优的表现，但我们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。参数在空间中的极小位移会导致模型输出概率的极小变化。当这个变化不足以使概率跨越阈值时，模型对样本的分类不改变，各种评价指标也就不变。而一旦某个微小位移导致了概率跨越了阈值，各种评价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优模型参数。

2.3 损失函数

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。于是模型训练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 $p(X = x_i)$ ，简写作 $p(x_i)$ 。将 X 看作一个信号源。观察到 $X = x_i$ 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：

$$I(X = x_i) = \log \frac{1}{p(x_i)} = -\log p(x_i), \quad i = 1 \dots k \quad (2.8)$$

式（2.8）中的对数可以以 2 为底，也可以取其他底，比如自然对数的底 e 。不同的底得到的信息量相差一个常系数。如果以 2 为底，信息量的单位是比特（bit）。 $I(X = x_i)$ 称为 $X = x_i$ 这条信息的自信息量（self-information）。 $I(X = x_i)$ 随着 $p(x_i)$ 变化的图像如图 2-2 所示。 $p(x_i)$ 趋向于 1 时， $I(X = x_i)$ 趋向于 0； $p(x_i)$ 趋向于 0 时， $I(X = x_i)$ 趋向于正无穷。

自信息量定义背后的洞见是：信息所告知的事件的概率越小，则信息的信息量越大。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。

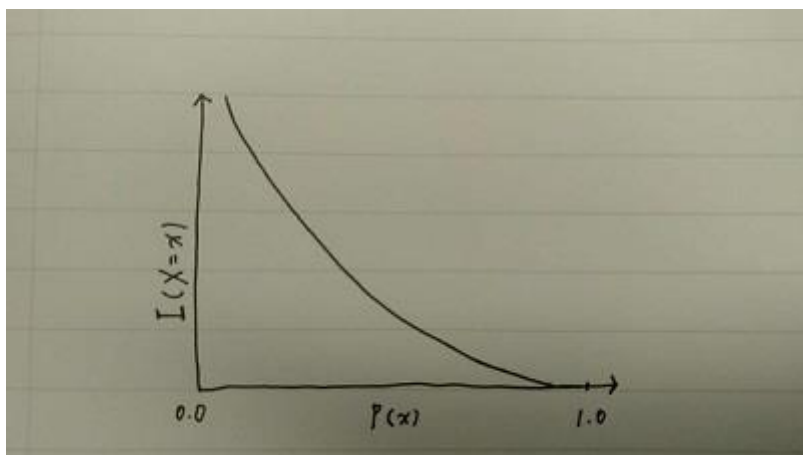


图 2-2 自信息量的图像

令信息源 X 取不同值 x_1, x_2, \dots, x_k 的概率为 $p(x_1), p(x_2), \dots, p(x_k)$ 。定义信息源的熵（entropy）为：

$$H(p) = \sum_{i=1}^k p(x_i) \log \frac{1}{p(x_i)} = -\sum_{i=1}^k p(x_i) \log p(x_i) \quad (2.9)$$

信息源由分布 p 描述，故将熵视为 p 的函数。熵的概念来自热力学。 $H(p)$ 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取平均。换句话说 $H(p)$ 是 $\log 1/p(x)$ 在分布 p 上的期望（expectation）。式（2.9）是针对离散型随机变量。对连续型随机变量应以积分取代求和。

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

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

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 $p(x_i) = q(x_i)$ ，即 $\log p(x_i)/q(x_i) = 0$ ，则 $KLD(p||q) = 0$ 。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。

注意式（2.10）第二个等号后。将第一项定义为分布 p 和 q 的交叉熵（cross entropy）：

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

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

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

分布 p 和 q 的交叉熵等于 p 和 q 的 K-L 散度加上 p 的熵。如果分布 p 不变，则 $H(p, q)$ 与 $KLD(p||q)$ 之间相差一个常数 $H(p)$ 。于是 $H(p, q)$ 也可用来衡量分布 p 和 q 的差异程度： $H(p, q)$ 越小则 p 和 q 越相似。

对于一个训练样本 $\{\mathbf{x}^i, y^i\}$ ，可以认为标签 y^i 给出了一个 \mathbf{x}^i 类别的伯努利分布：

$$p(\mathbf{x}^i \in A) = y^i, \quad p(\mathbf{x}^i \in B) = 1 - y^i, \quad i = 1 \dots m \quad (2.13)$$

当 \mathbf{x}^i 属于 A 类时 $y^i = 1$ ，该分布就是 $p(\mathbf{x}^i \in A) = 1, p(\mathbf{x}^i \in B) = 0$ ；当 \mathbf{x}^i 属于 B 类时 $y^i = 0$ ，该分布就是 $p(\mathbf{x}^i \in A) = 0, p(\mathbf{x}^i \in B) = 1$ 。这是一个“确定”的分布。

逻辑回归模型的输出也是一个伯努利分布：

$$q(\mathbf{x}^i \in A) = \frac{1}{1+e^{-b-\mathbf{w}^T \mathbf{x}^i}}, \quad q(\mathbf{x}^i \in B) = \frac{1}{1+e^{b+\mathbf{w}^T \mathbf{x}^i}}, \quad i = 1 \dots m \quad (2.14)$$

我们希望模型给出的分布与训练标签给出的分布越相似越好。于是可以将训练标签给出的分布 p 和模型给出的分布 q 的交叉熵作为在训练样本 $\{\mathbf{x}^i, y^i\}$ 上的损失：

$$loss(\mathbf{w}, b | \mathbf{x}^i, y^i) = -y^i \log \frac{1}{1+e^{-b-\mathbf{w}^T \mathbf{x}^i}} - (1 - y^i) \log \frac{1}{1+e^{b+\mathbf{w}^T \mathbf{x}^i}}, \quad i = 1 \dots m \quad (2.15)$$

$loss(\mathbf{w}, b | \mathbf{x}^i)$ 较大表示模型分布与训练标签分布之间的差异较大。反之亦然。

式 (2.15) 是在一个训练样本上的损失。在整个训练集上的损失就是在所有样本上的损失的平均：

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

式 (2.16) 就是交叉熵损失函数。它与训练集有关，以模型参数 \mathbf{w} 和 b 为自变量。逻辑回归模型的训练就是寻找使 $\text{loss}(\mathbf{w}, b)$ 尽可能小的 \mathbf{w} 和 b 。这就将模型训练问题转化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。 X 和 Y 是两个离散型随机变量。贝叶斯公式 (Bayes Rule) 是：

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

式 (2.17) 左边的 $p(X = x|Y = y)$ 称为后验概率 (posterior probability)。它是观察到事件 $Y = y$ 的前提下，事件 $X = x$ 发生的概率。右边分子上的 $p(X = x)$ 称为先验概率 (prior probability)。它是事件 $X = x$ 发生的概率。分子上的 $p(Y = y|X = x)$ 称为似然概率 (likelihood)。它是事件 $X = x$ 发生的前提下，事件 $Y = y$ 发生的概率。分母是事件 $Y = y$ 发生的边缘概率 (Marginal Probability)：

$$p(Y = y) = \sum_x p(Y = y, X = x) \quad (2.18)$$

由于是对 X 的所有可能取值求和，所以 $p(Y = y)$ 与 X 的取值无关。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 $p(X = x, Y = y)$ ——事件 $X = x$ 和 $Y = y$ 同时发生的概率。

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

$$p(\mathbf{w}, b|S) = \frac{p(S|\mathbf{w}, b)p(\mathbf{w}, b)}{p(S)} \quad (2.19)$$

观察到训练集 S 前提下参数值是 \mathbf{w} 和 b 的后验概率，等于参数值是 \mathbf{w} 和 b 的先验概率乘以参数值是 \mathbf{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。

训练的目标是寻找观察到训练集 S 前提下最有可能的参数值，也就是使后验概率 $p(\mathbf{w}, b|S)$ 最大的参数值：

$$\mathbf{w}^*, b^* = \operatorname{argmax}_{\mathbf{w}, b} p(\mathbf{w}, b|S) \quad (2.20)$$

假设先验分布 $p(\mathbf{w}, b)$ 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：

$$\mathbf{w}^*, b^* = \operatorname{argmax}_{\mathbf{w}, b} p(S|\mathbf{w}, b) \quad (2.21)$$

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

对于一个训练样本 $\{\mathbf{x}^i, y^i\}$ ， 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 \mathbf{x}^i 属于 y^i 所标识的类别的概率是：

$$p(y^i|\mathbf{w}, b, \mathbf{x}^i) = \left(\frac{1}{1+e^{-b-\mathbf{w}^T \mathbf{x}^i}} \right)^{y^i} \cdot \left(\frac{1}{1+e^{b+\mathbf{w}^T \mathbf{x}^i}} \right)^{1-y^i}, \quad i = 1 \dots m \quad (2.22)$$

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 $p(\mathbf{x}^i \in A)$ 或 $p(\mathbf{x}^i \in B)$ 。假设训练样本是独立的，可以得到：

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

因为 \log 是单调递增的。式（2.21）等价于寻找：

$$\mathbf{w}^*, b^* = \operatorname{argmax}_{\mathbf{w}, b} \log p(S|\mathbf{w}, b) \quad (2.24)$$

根据公式（2.22）、（2.23）和（2.24），最大似然估计是寻找 \mathbf{w}^* 和 b^* 使式（2.25）最大化：

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

最大化 $\log p(S|\mathbf{w}, b)$ 等价于最小化它的相反数。于是最大似然估计就是寻找 \mathbf{w}^* 和 b^* 使式(2.26)最小化：

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

除了一个常系数 $1/m$ ，式(2.26)和式(2.16)相同。所以最小化交叉熵(2.16)的 \mathbf{w}^* 和 b^* 正是最大似然估计。

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。

2.3.3 从几何角度理解逻辑回归的交叉熵损失

第1章介绍过，逻辑回归只能形成超平面分界面（以下把2维直线、3维平面以及更高维的超平面统称超平面）。如果以 $p(\mathbf{x} \in A) = 0.5$ 为阈值，则分界超平面的方程是：

$$p(\mathbf{x} \in A) = \frac{1}{1+e^{-b-\mathbf{w}^T \mathbf{x}}} = 0.5 \quad (2.27)$$

经过简单的计算，可知式(2.27)等价于：

$$b + \mathbf{w}^T \mathbf{x} = 0 \quad (2.28)$$

满足式(2.28)的所有点构成一个以 \mathbf{w} 为法向量的超平面 \mathbb{P} 。 \mathbb{P} 上的点 \mathbf{x} 与 \mathbf{w} 的内积是常数，即 \mathbb{P} 上的向量 \mathbf{x} 向 \mathbf{w} 方向的投影长度都相同：

$$\frac{\mathbf{w}^T \mathbf{x}}{\|\mathbf{w}\|} = -\frac{b}{\|\mathbf{w}\|}, \quad \mathbf{x} \in \mathbb{P} \quad (2.29)$$

若空间某一点向 \mathbf{w} 方向的投影长度大于 $-b/\|\mathbf{w}\|$ ，则它位于 \mathbb{P} 的一侧，称之为正侧；反之若某一点向 \mathbf{w} 方向的投影长度小于 $-b/\|\mathbf{w}\|$ ，则它位于 \mathbb{P} 的另一侧，称之为负侧。 \mathbf{x} 的投影长度与 $-b/\|\mathbf{w}\|$ 之差的绝对值 $|b + \mathbf{w}^T \mathbf{x}|/\|\mathbf{w}\|$ 是 \mathbf{x} 与超平面 \mathbb{P} 之间的距离。如图2-3所示。

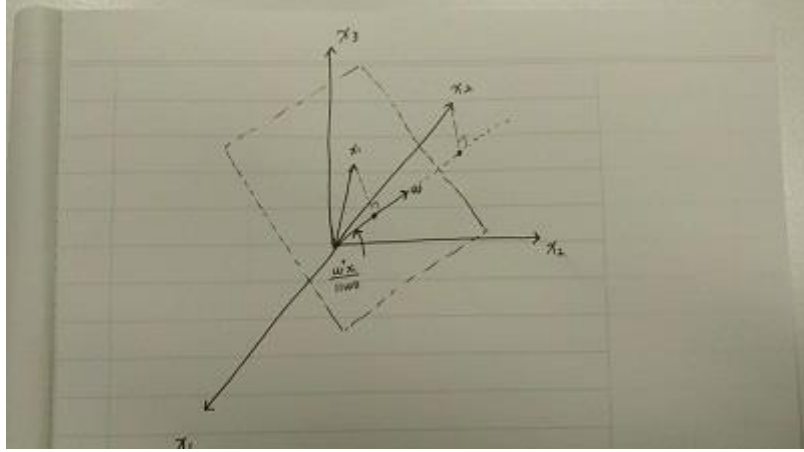


图 2-3 超平面将空间分为两侧以及点到超平面的距离

现在用另一种编码标识训练样本的类别：

$$\tilde{y}^i = \begin{cases} 1, & \mathbf{x}^i \in A \\ -1, & \mathbf{x}^i \in B \end{cases}, \quad i = 1 \dots m \quad (2.30)$$

假设一个逻辑回归模型能够完美地将 A/B 两类训练样本分开。所有 $\tilde{y}^i = 1$ 的样本有 $p(\mathbf{x}^i \in A) > 0.5$ 。于是 $b + \mathbf{w}^T \mathbf{x}^i > 0$ ，即 \mathbf{x}^i 位于 \mathbb{P} 的正侧。此时 $\tilde{y}^i(b + \mathbf{w}^T \mathbf{x}^i) > 0$ ；所有 $\tilde{y}^i = -1$ 的样本有 $p(\mathbf{x}^i \in A) < 0.5$ 。于是 $b + \mathbf{w}^T \mathbf{x}^i < 0$ ，即 \mathbf{x}^i 位于 \mathbb{P} 的负侧。此时仍有 $\tilde{y}^i(b + \mathbf{w}^T \mathbf{x}^i) > 0$ 。

当 $\tilde{y}^i(b + \mathbf{w}^T \mathbf{x}^i) < 0$ ，即 \tilde{y}^i 和 $b + \mathbf{w}^T \mathbf{x}^i$ 符号相反时，模型分类错误——A 类样本位于 \mathbb{P} 的负侧， $p(\mathbf{x}^i \in A) < 0.5$ ；B 类样本位于 \mathbb{P} 的正侧， $p(\mathbf{x}^i \in A) > 0.5$ 。损失函数应该对分错的情况施以惩罚。

交叉熵损失 (2.16) 中的 y^i 用 1/0 标识 A/B 类别。对于所有 i ， y^i 与 \tilde{y}^i 的关系是：

$$y^i = \frac{1 + \tilde{y}^i}{2}, \quad i = 1 \dots m \quad (2.31)$$

将式 (2.31) 代入交叉熵损失 (2.16)，得到：

$$\text{loss}(\mathbf{w}, b) = \frac{1}{m} \sum_{i=1}^m \left(-\frac{1 + \tilde{y}^i}{2} \log \frac{1}{1 + e^{-b - \mathbf{w}^T \mathbf{x}^i}} - \frac{1 - \tilde{y}^i}{2} \log \frac{1}{1 + e^{b + \mathbf{w}^T \mathbf{x}^i}} \right) \quad (2.32)$$

每一个训练样本 $\{\mathbf{x}^i, \tilde{y}^i\}$ 对交叉熵损失的贡献是：

$$\begin{aligned} loss(\mathbf{w}, b | \mathbf{x}^i, \tilde{y}^i) = & -\frac{1+\tilde{y}^i}{2} \log \frac{1}{1+e^{-b-\mathbf{w}^T \mathbf{x}^i}} - \frac{1-\tilde{y}^i}{2} \log \frac{1}{1+e^{b+\mathbf{w}^T \mathbf{x}^i}} = \frac{1}{2} \left(\tilde{y}^i \log \frac{1+e^{b+\mathbf{w}^T \mathbf{x}^i}}{1+e^{-b-\mathbf{w}^T \mathbf{x}^i}} - \right. \\ & \left. \log \frac{1}{(1+e^{b+\mathbf{w}^T \mathbf{x}^i})(1+e^{-b-\mathbf{w}^T \mathbf{x}^i})} \right), \quad i = 1 \dots m \end{aligned} \quad (2.33)$$

A 类样本的 $\tilde{y}^i = 1$ ，公式（2.33）成为：

$$loss(\mathbf{w}, b | \mathbf{x}^i, \tilde{y}^i) = \frac{1}{2} \log \left(1 + e^{b+\mathbf{w}^T \mathbf{x}^i} \right)^2 = \log \left(1 + e^{b+\mathbf{w}^T \mathbf{x}^i} \right), \quad i = 1 \dots m \quad (2.34)$$

B 类样本的 $\tilde{y}^i = -1$ ，公式（2.33）成为：

$$loss(\mathbf{w}, b | \mathbf{x}^i, \tilde{y}^i) = \frac{1}{2} \log \left(1 + e^{b+\mathbf{w}^T \mathbf{x}^i} \right)^2 = \log \left(1 + e^{-(b+\mathbf{w}^T \mathbf{x}^i)} \right), \quad i = 1 \dots m \quad (2.35)$$

结合公式（2.34）和（2.35），得到：

$$loss(\mathbf{w}, b | \mathbf{x}^i, \tilde{y}^i) = \log \left(1 + e^{\tilde{y}^i(b+\mathbf{w}^T \mathbf{x}^i)} \right), \quad i = 1 \dots m \quad (2.36)$$

将 $loss(\mathbf{w}, b | \mathbf{x}^i, \tilde{y}^i)$ 视作 $\tilde{y}^i(b + \mathbf{w}^T \mathbf{x}^i)$ 的函数，其图像如图 2-4 所示。

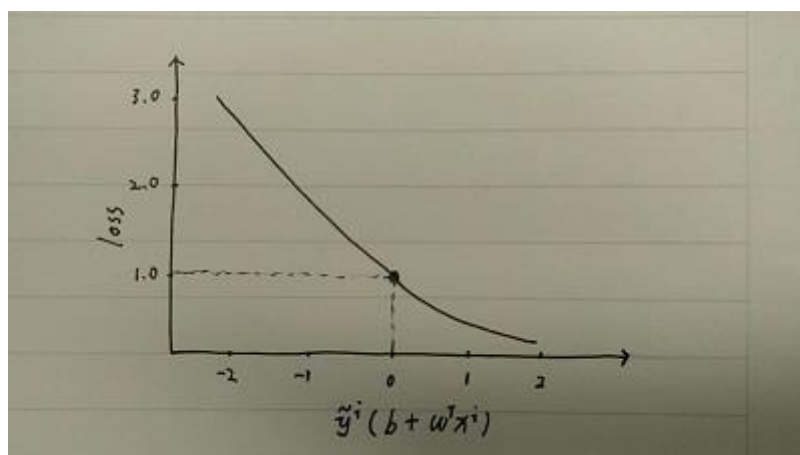


图 2-4 交叉熵损失作为 $\hat{y}^i(b + \mathbf{w}^T \mathbf{x}^i)$ 的函数的图像

以 $p(\mathbf{x} \in A) = 0.5$ 为阈值。如果对某一个训练样本 $\{\mathbf{x}^i, \tilde{y}^i\}$ 有 $\tilde{y}^i(b + \mathbf{w}^T \mathbf{x}^i) > 0$ ，则分类正确。 $\tilde{y}^i(b + \mathbf{w}^T \mathbf{x}^i)$ 越大 \mathbf{x}^i 距离分界面 \mathbb{P} 越远。如果 $\tilde{y}^i(b + \mathbf{w}^T \mathbf{x}^i) < 0$ 则分类错误。 $\tilde{y}^i(b + \mathbf{w}^T \mathbf{x}^i)$ 越小，即 $|\tilde{y}^i(b + \mathbf{w}^T \mathbf{x}^i)|$ 越大则 \mathbf{x}^i 在错误的一侧距离分界面 \mathbb{P} 越远。后一种情况下损失函数的值应该更大。如图 2-4 所示，交叉熵损失函数恰当地惩罚了分类错误。

2.4 小结

本章介绍了分类模型的评价指标和损失函数的概念。介绍了交叉熵损失函数，并从分布差异度——K-L 散度和最大似然估计两种视角阐释了交叉熵损失的原理和含义。引入损失函数后，寻找最优模型参数的训练过程就成为了损失函数的优化过程。至于优化损失函数的方法，是下一章讨论的内容。