模式识别与机器学习

樊超

December 12, 2025

目 录

1 引言

探索数据中的模式（pattern）是一个根本性问题，其相关研究历史悠久。比如，16世纪第谷·布拉赫（Tycho Brahe）的大量天文观测，使约翰内斯·开普勒（Johannes Kepler）发现了行星运动的经验定律，这反过来又为经典力学的发展提供了跳板。同样，对原子光谱中规律性的发现，在20世纪早期量子物理学的发展和验证中发挥了关键作用。模式识别这一领域关注的是通过计算机算法自动发现数据中的规律，并利用这些规律来执行诸如将数据分类到不同类别等操作。

考虑识别手写数字的例子，如图1.1所示。每个数字对应一个28×28像素的图像，因此可以用一个由784个实数构成的向量 $\mathbf{x}$ 来表示。我们的目标是构建一台机器，它能够将这样的向量 $\mathbf{x}$ 作为输入，并输出数字 $0, \cdots, 9$ 的身份。这是一个非平凡的问题，因为手写体存在极大的多样性。可以通过手工制定规则或启发式方法，根据笔画的形状来区分数字，但在实践中，这种方法会导致规则及其例外情况的大量增加，结果不可避免地会很差。

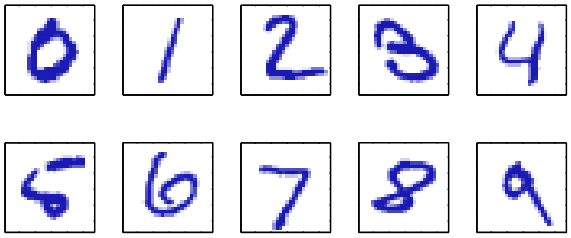

图 1.1 来自美国邮政编码（US zip codes）的手写数字示例

通过采用机器学习的方法，可以获得更好的结果。在这种方法中，使用一个包含 N 个数字的集合 $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$，称为训练集，来调整自适应模型的参数。训练集中的数字类别是事先已知的，通常通过逐个检查并手工标注得到。我们可以用目标向量 $\mathbf{t}$ 来表示一个数字的类别，它刻画了对应数字的身份。关于如何用向量来表示类别的合适技术，将在后文讨论。需要注意的是，对于每一个数字图像 $\mathbf{x}$，都有一个相应的目标向量 $\mathbf{t}$。

运行机器学习算法的结果可以表示为一个函数 $\mathbf{y}(\mathbf{x})$，它以一个新的数字图像 $\mathbf{x}$ 作为输入，并生成一个输出向量 y，其编码方式与目标向量相同。函数 $\mathbf{y}(\mathbf{x})$ 的精确形式是在训练阶段（也称为学习阶段）中根据训练数据确定的。一旦模型完成训练，它就可以识别新的数字图像，这些新的图像被称为测试集。能够正确分类与训练中使用的样本不同的新样本的能力，被称为泛化。在实际应用中，输入向量的变化范围通常非常大，以至于训练数据只能覆盖所有可能输入向量的一小部分，因此，泛化是模式识别中的核心目标。

在大多数实际应用中，原始输入变量通常会经过预处理，将其转换到某个新的变量空间中，在这个空间里，希望模式识别问题更容易解决。以数字识别问题为例，数字图像通常会被平移和缩放，使得每个数字都被包含在一个固定大小的方框中。这样大大减少了每个数字类别内部的变化，因为所有数字的位置和尺度都相同，从而使得后续的模

式识别算法更容易区分不同的类别。这个预处理阶段有时也被称为特征提取。需要注意的是，新的测试数据必须使用与训练数据相同的步骤进行预处理。

预处理也可能是为了加快计算速度。例如，如果目标是在高分辨率视频流中实现实时人脸检测，计算机必须每秒处理海量像素，而将这些像素直接输入复杂的模式识别算法在计算上可能不可行。相反，目标是找到既能快速计算、又能保留有用区分信息的特征，以便将人脸与非人脸区分开来。这些特征随后被用作模式识别算法的输入。例如，可以非常高效地计算图像在矩形子区域内的平均强度值（Viola 和 Jones，2004），而一组这样的特征在快速人脸检测中可能非常有效。由于这类特征的数量少于像素的数量，这种预处理也可以看作是一种降维。不过，在预处理过程中必须谨慎，因为往往会丢弃部分信息，如果这些信息对问题的解决是关键的，那么系统的整体准确率可能会受到影响。

在训练数据由输入向量及其对应的目标向量组成的应用中，这类问题被称为监督学习问题。像数字识别这种需要将每个输入向量分配到有限个离散类别之一的情况，被称为分类问题。如果期望的输出由一个或多个连续变量构成，那么任务就称为回归。一个回归问题的例子是预测化工制造过程中的产量，其输入由反应物的浓度、温度和压力组成。

在其他模式识别问题中，训练数据仅由一组输入向量 **x** 组成，而没有对应的目标值。在这种无监督学习问题中，目标可能是发现数据中相似样本的分组，这称为聚类；或者确定输入空间中数据的分布，这称为密度估计；或者将数据从高维空间投影到二维或三维空间，以便进行可视化。

最后，强化学习（Sutton 和 Barto，1998）所研究的问题是在给定情境下找到合适的动作，以使奖励最大化。与监督学习不同，这里的学习算法并不会得到最优输出的示例，而是必须通过试错的过程去发现它们。通常情况下，这涉及到一个状态和动作的序列，在其中学习算法不断与环境交互。在许多情形下，当前动作不仅影响即时奖励，还会对随后的所有时间步的奖励产生影响。

例如，通过使用合适的强化学习技术，神经网络可以学会高水平地玩西洋双陆棋（backgammon）（Tesauro，1994）。在这里，网络必须学会将棋盘位置与掷骰子的结果作为输入，并生成一个强有力的走棋作为输出。这通常通过让网络与它自己的副本对弈上百万局来实现。一个主要的挑战在于，一局西洋双陆棋可能包含数十步操作，而胜利这种奖励只会在游戏结束时才出现。此时必须将奖励合理地归因到导致它的所有动作上，即便其中有些动作很好，有些动作则不那么好。这就是所谓的信用分配问题（credit assignment problem）。

强化学习的一个普遍特征是探索与利用之间的权衡：探索是系统尝试新的动作方式以观察其效果，而利用则是系统使用那些已知能带来高奖励的动作。如果过度偏重探索或利用，都会导致较差的结果。强化学习仍然是机器学习研究中的一个活跃领域。然而，详细的讨论超出了本书的范围。

虽然这些任务各自需要不同的工具和技术，但支撑它们的许多关键思想在所有这

类问题中都是共通的。本章的主要目标之一，就是以一种相对非正式的方式介绍其中若干最重要的概念，并通过简单的例子加以说明。在本书后面的内容中，我们会看到这些思想再次出现，不过那时它们将出现在更复杂的模型背景下，而这些模型适用于现实世界中的模式识别应用。本章还提供了对三种重要工具的自成体系的介绍，即概率论、决策论和信息论。尽管这些主题听起来可能让人望而生畏，但实际上它们相对直观，并且要想在实际应用中最有效地使用机器学习技术，对它们的清晰理解是必不可少的

1.1 例子：多项式曲线拟合

我们首先介绍一个简单的回归（regression）问题，并将在本章中反复使用它作为示例，以引出若干关键概念。假设我们观察到一个实值输入变量 x，并希望利用这一观测来预测一个实值目标变量 t 的值。为了当前的目的，考虑一个使用人工合成数据的例子是有启发性的，因为这样我们就确切知道生成数据的过程，可以将其与任何学习到的模型进行比较。本例中的数据由函数 $\sin(2\pi x)$ 生成，并在目标值中加入了随机噪声。关于这一数据生成过程的详细描述见附录**??**。

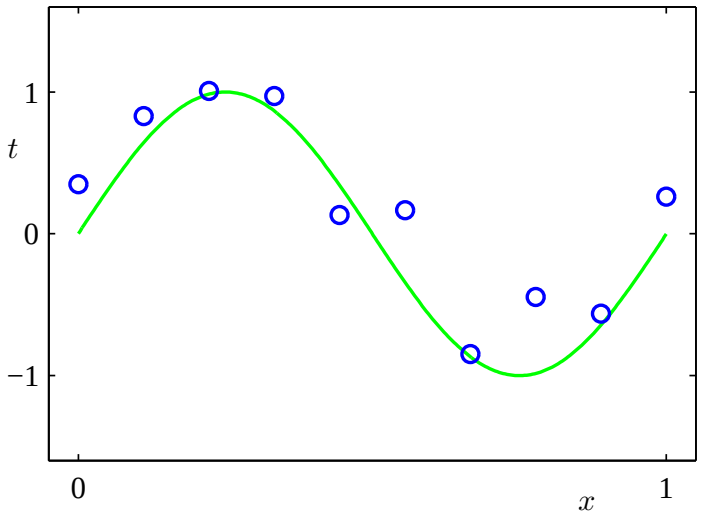

图 1.2 一个包含 $N = 10$ 个点的训练数据集的示意图，以蓝色圆点表示；每个点由输入变量 x 的一次观测及其对应的目标变量 t 构成。绿色曲线表示用于生成数据的函数 $\sin(2\pi x)$。我们的目标是在不知道这条绿色曲线的情况下，对某个新的 x 值预测对应的 t 值。

现在假设我们得到了一个包含 N 个观测值的训练集，其中输入记作 $\mathbf{x} \equiv (x_1, \ldots, x_N)^{\mathrm{T}}$，并且有与之对应的目标值观测，记作 $\mathbf{t} \equiv (t_1, \ldots, t_N)^{\mathrm{T}}$。图 1.2 显示了一个包含 $N = 10$ 个数据点的训练集。图 1.2 中的输入数据集 $\mathbf{x}$ 是通过在区间 $[0, 1]$ 上均匀选取 x_n 的值（$n = 1, \ldots, N$）生成的；目标数据集 $mathbft$ 则是先计算相应的函数值 $\sin(2\pi x)$，然后在每个点上加入一个服从高斯分布（Gaussian distribution，见 1.2.4 节）的随机噪声，从而得到对应的 t_n。通过这种方式生成数据，我们刻画了许多真实数据集的一个特性，即它们具有某种潜在的规律性，而我们希望学习到这一规律性，但同时单个观测值会受到随机噪声的干扰。这种噪声可能来自内在的随机过程（如放射性衰变），但更常见的情况是，由一些我们未能观测到的可变因素所导致。

我们的目标是利用这个训练集，在某个新的输入变量 $\hat{x}$ 的取值下预测目标变量 $\hat{t}$ 的值。正如我们稍后将看到的，这实际上涉及到隐式地试图发现潜在的函数 $\sin(2\pi x)$。由于我们必须从有限的数据集中进行泛化，这本质上是一个困难的问题。更进一步，观测数据中还包含噪声，因此对于给定的 $\hat{x}$，对应的 $\hat{t}$ 值存在不确定性。概率论（probability theory，见 1.2 节）提供了一个框架，可以以精确且定量的方式表达这种不确定性；而决策论（decision theory，见 1.5 节）则使我们能够利用这种概率表示，依据合适的准则做出最优的预测。

不过，就目前而言，我们将采用一种比较非正式的方法，考虑一种基于曲线拟合（curve fitting）的简单思路。具体来说，我们将使用如下形式的多项式函数来拟合数据：

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j \tag{1.1}$$

其中，M 表示多项式的阶数，x^j 表示 x 的 j 次幂。多项式的系数 $w_0, \ldots, w_M$ 共同记作向量 $\mathbf{w}$。需要注意的是，虽然多项式函数 $y(x, \mathbf{w})$ 是关于 x 的非线性函数，但它关于系数向量 $\mathbf{w}$ 却是线性的。像多项式这样的函数，在未知参数上是线性的，具有重要的性质，被称为线性模型（linear models），并将在第 3 章和第 4 章中进行深入讨论。

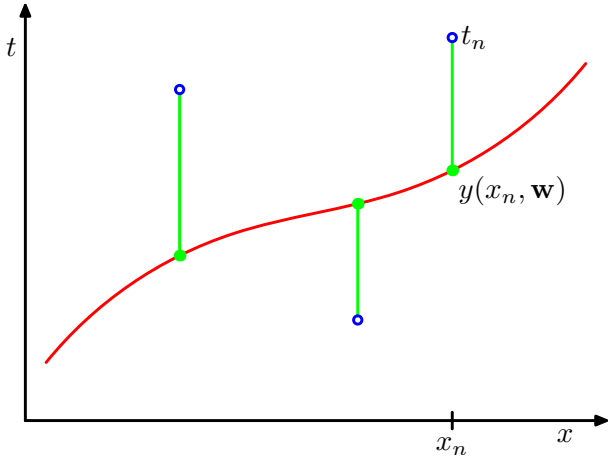

图 1.3 误差函数 (1.2) 对应于每个数据点与函数 $y(x, \mathbf{w})$ 之间偏差平方和的一半，其中这些偏差由垂直的绿色线段表示。

系数的取值将通过将多项式拟合到训练数据来确定。这可以通过最小化一个误差函数来实现，该误差函数用于度量函数 $y(x, \mathbf{w})$（对于任意给定的 $\mathbf{w}$）与训练数据点之间的不匹配程度。一种简单且被广泛使用的误差函数选择是：计算每个数据点 x_n 的预测值 $y(x_n, \mathbf{w})$ 与对应目标值 t_n 之间误差的平方和。于是我们需要最小化

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 \tag{1.2}$$

其中的因子 $1/2$ 是为了后续推导的方便而引入的。我们将在本章后面讨论选择这种误差函数的动机。这里先简单指出，该误差函数始终是非负的，并且当且仅当函数 $y(x, \mathbf{w})$ 恰好通过每一个训练数据点时，其值才为零。平方和误差函数的几何解释如图 1.3 所

示。我们可以通过选择使 $E(\mathbf{w})$ 尽可能小的 $\mathbf{w}$ 来解决曲线拟合问题。由于误差函数是系数 $\mathbf{w}$ 的二次函数，它对各个系数的导数关于 $\mathbf{w}$ 的各分量是线性的，因此该误差函数的最小化具有唯一解，记作 $\mathbf{w}^*$，并且可以通过闭式形式求解。最终得到的多项式由函数 $y(x, \mathbf{w}^*)$ 给出。

还剩下一个问题，即如何选择多项式的阶数 M。正如我们将看到的，这将成为一个重要概念的例子，该概念被称为模型比较（model comparison）或模型选择（model selection）。在图 1.4 中，我们展示了将阶数分别为 $M = 0, 1, 3, 9$ 的多项式拟合到图 1.2 所示数据集上的四个结果示例。

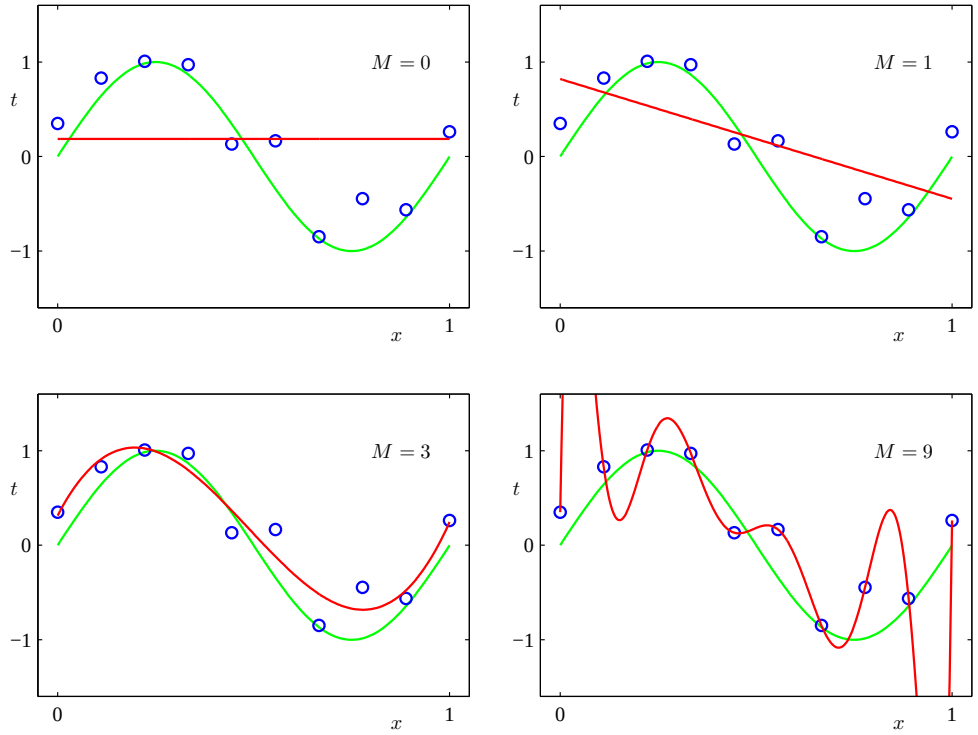

图 1.4 不同阶数 M 的多项式拟合结果示意图，以红色曲线表示，并拟合到图 1.2 所示的数据集上。

我们注意到，常数多项式（$M = 0$）和一次多项式（$M = 1$）对数据的拟合效果较差，因此对函数 $\sin(2\pi x)$ 的表示也很差。三次多项式（$M = 3$）似乎在图 1.4 所示的几个例子中给出了对函数 $\sin(2\pi x)$ 的最佳拟合。当我们采用更高阶的多项式（$M = 9$）时，得到的结果在训练数据上的拟合非常好。事实上，该多项式恰好通过了每一个数据点，并且满足 $E(\mathbf{w}^*) = 0$。然而，拟合曲线出现剧烈震荡，对函数 $\sin(2\pi x)$ 的表示反而非常糟糕。这种现象称为过拟合（over-fitting）。

正如前面提到的，我们的目标是通过对新数据做出准确预测来实现良好的泛化。为了定量分析泛化性能对 M 的依赖关系，我们可以考虑一个单独的测试集，该测试集包含 100 个数据点，这些数据点的生成方式与训练集完全相同，但在目标值中加入的随机噪声取值不同。对于每一个 M 的选择，我们都可以计算训练数据对应的残差值 $E(\mathbf{w}^\star)$(由公式 1.2 给出)，同时也可以计算测试数据集上的 $E(\mathbf{w}^\star)$。有时使用均方根误差

（root-mean-square error，RMS error）会更加方便，其定义为

$$E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N} \tag{1.3}$$

在这里，除以 N 使我们能够在比较不同规模的数据集时保持一致，而开平方则确保 E_{RMS} 与目标变量 t 的量纲和单位相同。图 1.5 给出了在不同 M 值下，训练集和测试集的均方根误差的曲线。测试集误差衡量了我们在新输入 x 的观测值上预测目标变量 t 的效果。由图 1.5 可见，当 M 较小时，测试集误差相对较大，这是因为相应的多项式缺乏灵活性，无法刻画函数 $\sin(2\pi x)$ 的振荡特性。当 $3 \leq M \leq 8$ 时，测试集误差较小，同时对生成函数 $\sin(2\pi x)$ 的表示也较为合理，例如从图 1.4 中 $M = 3$ 的情况可以清楚地看到这一点。

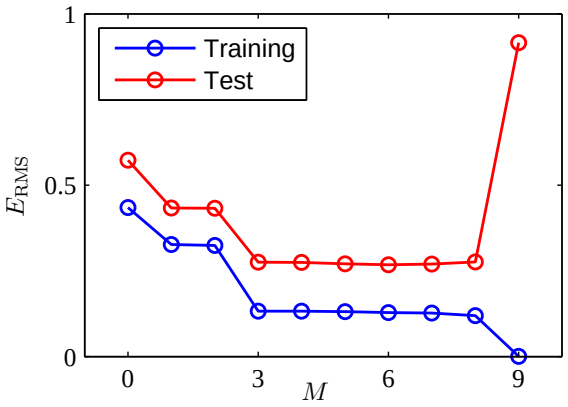

图 1.5 在不同 M 取值下，训练集与独立测试集上的均方根误差（由公式 1.3 定义）曲线图。

当 $M = 9$ 时，训练集误差为零，这是意料之中的结果，因为该多项式包含 10 个自由度，对应于 10 个系数 $w_0, \ldots, w_9$，因此可以完全拟合训练集中的 10 个数据点。然而，此时测试集误差却变得非常大，并且正如图 1.4 所示，相应的函数 $y(x, \mathbf{w}^*)$ 出现了剧烈的震荡。

这看起来似乎是一个悖论，因为一个给定阶数的多项式包含所有低阶多项式作为特殊情形。因此，$M = 9$ 的多项式理应能够给出至少与 $M = 3$ 多项式一样好的结果。进一步来说，我们或许会认为，对新数据的最佳预测器应当是生成数据的函数 $\sin(2\pi x)$（我们将在后面看到，这确实如此）。我们知道，函数 $\sin(2\pi x)$ 的幂级数展开包含所有阶的项，因此我们可能会预期，随着 M 的增大，结果应当单调改进。

我们可以通过考察不同阶数多项式所得到的系数 $\mathbf{w}^\star$ 的数值（如表 1.1 所示），对这一问题获得一些直观认识。可以看到，随着 M 的增加，系数的数值大小通常会变得更大。特别是在 $M = 9$ 的多项式中，这些系数被精细地调节为较大的正值和负值，从而使得相应的多项式函数能够恰好通过所有数据点，但在数据点之间（尤其是在区间两端附近），函数表现出图 1.4 中所观察到的剧烈震荡。直观地讲，发生的情况是：当 M 较大时，多项式的灵活性增强，它们逐渐开始拟合目标值中的随机噪声。

考察在数据集规模变化时给定模型的行为也很有趣，如图 1.6 所示。可以看到，对

表 1.1 不同阶数多项式对应的系数 $\mathbf{w}^\star$ 表格。可以看到，随着多项式阶数的增加，系数的典型大小急剧增大。

	$M=0$	$M=1$	$M=6$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.1
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				-557682.99
$w_8^\star$				125201.43

于一个固定的模型复杂度，随着数据集规模的增大，过拟合问题会减轻。换句话说，数据集越大，我们就能承受将越复杂（即更灵活）的模型拟合到数据中。有一种经验性启发认为，数据点的数量至少应当是模型中自适应参数数量的若干倍（例如 5 或 10）。然而，正如我们将在第 3 章中看到的，参数的数量并不一定是衡量模型复杂度的最合适指标。

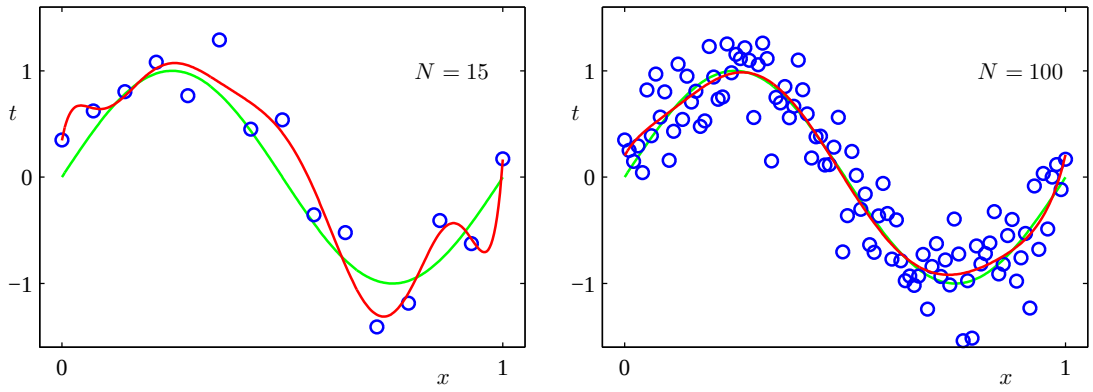

图 1.6 通过最小化平方和误差函数所得到的 $M=9$ 多项式解的曲线图：左图为 $N=15$ 个数据点的情况，右图为 $N=100$ 个数据点的情况。可以看到，增大数据集的规模能够减轻过拟合问题。

此外，将模型中的参数数量限制为与可用训练集的规模相适应，这种做法多少让人感到不够理想。更合理的方式似乎是根据所要解决问题的复杂性来选择模型的复杂度。我们将看到，用最小二乘法来求解模型参数可以看作是最大似然（maximum likelihood，见 1.2.5 节）的一个特例，而过拟合问题可以理解为最大似然的一种普遍性质。通过采用贝叶斯方法，可以避免过拟合问题。我们还将看到，从贝叶斯的角度来看，使用参数数量远大于数据点数量的模型并不存在困难。事实上，在贝叶斯模型中，有效参数数量会自动适应数据集的规模。

不过，目前继续采用当前的方法仍然是有启发性的，我们将考虑在实际中如何将其应用到规模有限的数据集上，同时又希望使用相对复杂和灵活的模型。在这种情况下，一个常用来控制过拟合现象的技术是正则化（regularization），其做法是在误差函

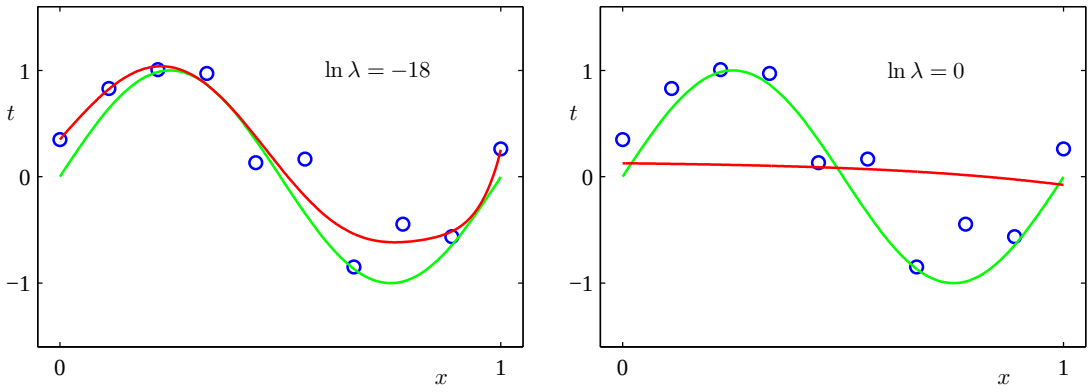

图 1.7 利用正则化误差函数 (1.4) 将 $M = 9$ 的多项式拟合到图 1.2 所示数据集的结果曲线图，其中正则化参数 λ 分别取 $\ln\lambda = -18$ 和 $\ln\lambda = 0$。无正则化的情况，即 $\lambda = 0$（对应 $\ln\lambda = -\infty$），见图 1.4 的右下角。

数 (1.2) 中加入一个惩罚项，以抑制系数取过大的值。最简单的惩罚项形式是所有系数平方和，从而得到修正后的误差函数：

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 \tag{1.4}$$

其中，$|\mathbf{w}|^2 \equiv \mathbf{w}^T\mathbf{w} = w_0^2 + w_1^2 + \cdots + w_M^2$，系数 λ 控制正则化项相对于平方和误差项的重要性。需要注意的是，w_0 常常不包含在正则化项中，因为将其包括进去会导致结果依赖于目标变量的原点选择（Hastie 等，2001），或者它也可以被包括进去，但赋予一个单独的正则化系数（这一问题将在 5.5.1 节中更详细讨论）。同样，(1.4) 式中的误差函数可以通过闭式形式精确地最小化。这类技术在统计学文献中被称为收缩方法（shrinkage methods），因为它们会减小系数的数值。二次正则化的特殊情形被称为岭回归（ridge regression，Hoerl 和 Kennard，1970）。在神经网络的背景下，这种方法被称为权重衰减（weight decay）。

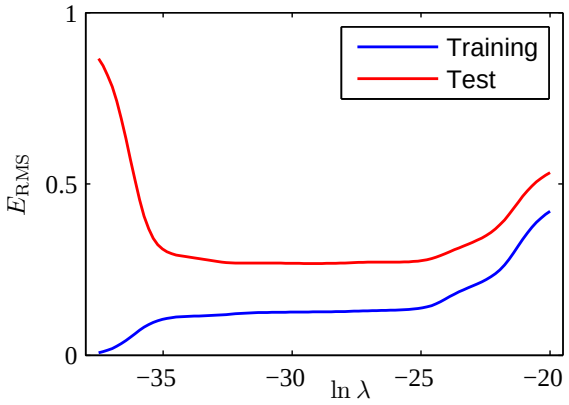

图 1.8 $M = 9$ 多项式情况下，均方根误差（公式 (1.3)）随 $\ln\lambda$ 变化的曲线图。

图 1.7 展示了将阶数 $M = 9$ 的多项式拟合到与之前相同的数据集上，但这次使用的是正则化误差函数 (1.4) 的结果。可以看到，当 $\ln\lambda = -18$ 时，过拟合现象得到了抑

表 1.2 $M = 9$ 多项式在不同正则化参数 λ 下的系数 $\mathbf{w}^*$ 表格。需要注意的是，$\ln\lambda = -\infty$ 对应于无正则化的模型，即图 1.4 右下角的情况。可以看到，随着 λ 的增大，系数的典型大小逐渐减小。

	$\ln\lambda = -\infty$	$\ln\lambda = -18$	$\ln\lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.92	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

制，此时得到的曲线更接近潜在函数 $\sin(2\pi x)$。然而，如果 λ 的取值过大，则同样会得到较差的拟合效果，如图 1.7 中 $\ln\lambda = 0$ 的情况所示。表 1.2 给出了对应拟合多项式的系数，结果显示正则化确实达到了减小系数数值大小的效果。

通过绘制训练集和测试集的均方根误差 (1.3) 随 $\ln\lambda$ 变化的曲线（如图 1.8 所示），可以看到正则化项对泛化误差的影响。结果表明，λ 实际上控制了模型的有效复杂度，从而决定了过拟合的程度。

模型复杂度的问题非常重要，将在 1.3 节中进行详细讨论。这里我们只需注意，如果要用这种最小化误差函数的方法来解决一个实际应用，就必须找到确定合适模型复杂度的方法。前面的结果给出了一个简单的思路，即将可用数据划分为训练集和验证集（也称为留出集 hold-out set）：训练集用于确定系数 $\mathbf{w}$，验证集用于优化模型复杂度（无论是 M 还是 λ）。然而，在许多情况下，这种方法会过于浪费宝贵的训练数据，因此我们需要寻找更复杂的方法。

到目前为止，我们对多项式曲线拟合的讨论主要依赖直觉。现在我们转向概率论，以寻求一种更有原则的方法来解决模式识别中的问题。概率论不仅为本书后续几乎所有的发展奠定基础，同时还能帮助我们在多项式曲线拟合的背景下，对已介绍的概念获得更深入的理解，并使我们能够将这些概念扩展到更复杂的情形中。

1.2 概率论

模式识别领域的一个关键概念是不确定性。它既来源于测量中的噪声，也来源于数据集规模的有限性。概率论提供了一个一致的框架来刻画和处理不确定性，并构成模式识别的核心基础之一。当它与决策论（见 1.5 节）结合时，即使在信息不完整或含糊的情况下，也能使我们基于所有可用信息做出最优预测。

我们将通过一个简单的例子来介绍概率论的基本概念。设想有两个盒子，一个红色，一个蓝色：红色盒子里有 2 个苹果和 6 个橙子，蓝色盒子里有 3 个苹果和 1 个橙子，如图 1.9 所示。现在假设我们随机选择一个盒子，然后从该盒子中随机取出一个水

果，观察它的种类后再将其放回原来的盒子中。我们可以想象将这一过程重复很多次。假设在这个过程中，选择红色盒子的概率为 40%，选择蓝色盒子的概率为 60%；并且当我们从某个盒子里取水果时，盒子里的每个水果被选中的可能性相同。

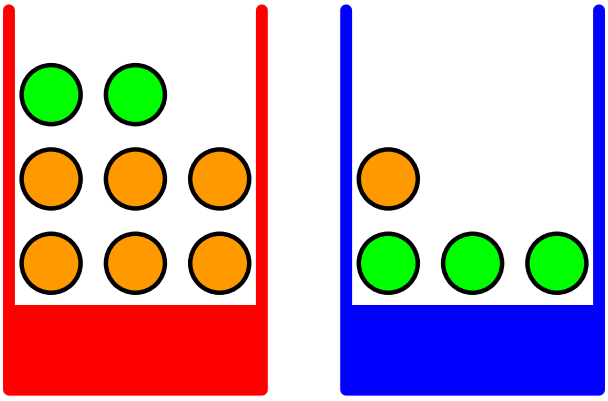

图 1.9 我们用一个简单的例子来引入概率（probability）的基本概念：两个有颜色的盒子，每个盒子里都装有水果（绿色表示苹果，橙色表示橙子）。

在这个例子中，被选中的盒子的身份是一个随机变量，记作 B。该随机变量可以取两个可能的值：r（对应红色盒子）或 b（对应蓝色盒子）。类似地，水果的身份也是一个随机变量，记作 F，它可以取值 a（表示苹果）或 o（表示橙子）。

首先，我们将事件的概率定义为：当试验次数趋于无穷大时，该事件发生的频率。由此可得，选择红色盒子的概率为 $4/10$，选择蓝色盒子的概率为 $6/10$。我们将这些概率写作 $p(B=r)=\frac{4}{10}$，　$p(B=b)=\frac{6}{10}$。需要注意的是，按照定义，概率必须落在区间 $[0,1]$ 内。此外，如果事件是互斥的并且包含了所有可能结果（例如在本例中，盒子必须是红色或蓝色），那么这些事件的概率之和必须为 1。

现在我们可以提出这样的问题：“选出的水果是苹果的总体概率是多少？”或者“在已知选中的是橙子的情况下，被选择的盒子是蓝色的概率是多少？”一旦掌握了概率的两个基本法则，即加法法则和乘法法则，我们就能回答这些问题，甚至是与模式识别相关的更复杂的问题。在介绍完这些法则之后，我们将回到水果盒子的例子。

为了推导概率法则，考虑图 1.10 所示的一个更一般的例子，其中包含两个随机变量 X 和 Y（例如可以对应上面提到的盒子和水果变量）。假设 X 可以取值 x_i，其中 $i=1,\ldots,M$，而 Y 可以取值 y_j，其中 $j=1,\ldots,L$。设进行 N 次试验，每次同时对变量 X 和 Y 取样。在 N 次试验中，$X=x_i$ 且 $Y=y_j$ 的次数记作 n_{ij}。同时，X 取值为 x_i 的次数（不论 Y 取何值）记作 c_i；类似地，Y 取值为 y_j 的次数记作 r_j。

X 取值为 x_i 且 Y 取值为 y_j 的概率记作 $p(X=x_i,Y=y_j)$，称为 $X=x_i$ 与 $Y=y_j$ 的联合概率 =。它由落在单元格 (i,j) 中的点数占总点数的比例给出，因此有

$$p(X=x_i,Y=y_j)=\frac{n_{ij}}{N} \tag{1.5}$$

这里我们隐含地考虑极限 $N\to\infty$。类似地，X 取值为 x_i 的概率（不论 Y 取何值）记作 $p(X=x_i)$，它由第 i 列中点数占总点数的比例给出，因此有

$$p(X=x_i)=\frac{c_i}{N} \tag{1.6}$$

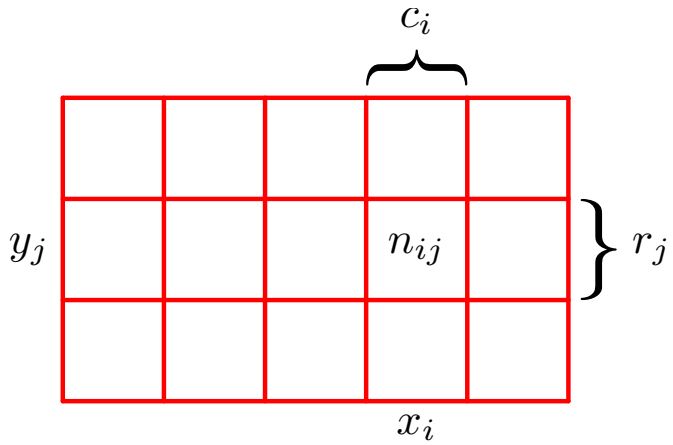

图 1.10 我们可以通过考虑两个随机变量来推导概率的加法法则和乘法法则。设随机变量 X 的取值为 x_i，其中 $i = 1, \ldots, M$；随机变量 Y 的取值为 y_j，其中 $j = 1, \ldots, L$。在这个例子中，我们取 $M = 5$，$L = 3$。如果考虑这两个变量的总实例数为 N，那么当 $X = x_i$ 且 $Y = y_j$ 时的实例数记作 n_{ij}，它表示数组对应单元格中的点数。第 i 列的点数（对应 $X = x_i$）记作 c_i，第 j 行的点数（对应 $Y = y_j$）记作 r_j。

因为图 1.10 中第 i 列的实例数正好是该列中各个单元格实例数的总和，即 $c_i = \sum_j n_{ij}$, 因此由 (1.5) 和 (1.6) 可得

$$p(X = x_i) = \sum_j p(X = x_i, Y = y_j) \tag{1.7}$$

这就是概率的加法法则。需要注意，$p(X = x_i)$ 有时也称为边缘概率，因为它是通过对其他变量（在本例中为 Y）边缘化，即求和消去后得到的。

如果我们只考虑 $X = x_i$ 的那些实例，那么其中 $Y = y_j$ 的比例记作 $p(Y = y_j \mid X = x_i)$，称为在 $X = x_i$ 条件下 $Y = y_j$ 的条件概率。它由第 i 列中落在单元格 (i, j) 的点数占该列总点数的比例给出，因此有

$$p(Y = y_j \mid X = x_i) = \frac{n_{ij}}{c_i} \tag{1.8}$$

由 (1.5)、(1.6) 和 (1.8) 可推得如下关系

$$\begin{aligned} p\left(X = x_i, Y = y_j\right) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\ &= p\left(Y = y_j \mid X = x_i\right) p\left(X = x_i\right) \end{aligned} \tag{1.9}$$

这就是概率的乘法法则（product rule）。

到目前为止，我们一直严格区分随机变量与其可能取的值。例如在水果的例子中，盒子 B 是随机变量，而它的一个取值可以是 r（表示红色盒子）。因此，B 取值为 r 的概率记作 $p(B = r)$。虽然这种写法有助于避免歧义，但符号会显得冗长，而在许多情况下并不需要如此繁琐。通常我们可以直接写 $p(B)$ 表示随机变量 B 的分布，或者写 $p(r)$ 表示分布在特定取值 r 处的值，只要上下文能保证含义清晰即可。

采用这种更简洁的符号，我们可以将概率论的两个基本法则写成如下形式：

概率法则

$$\text{求和法则} \quad p(X) = \sum_{Y} p(X, Y) \tag{1.10}$$

$$\text{乘法法则} \quad p(X, Y) = p(Y \mid X)\, p(X) \tag{1.11}$$

这里 $p(X, Y)$ 是联合概率，读作“X 和 Y 的概率”。类似地，$p(Y \mid X)$ 是条件概率），读作“在 X 已知的条件下 Y 的概率”；而 $p(X)$ 是边缘概率，即“X 的概率”。这两个简单的法则构成了本书中所使用的全部概率方法的基础。

由乘法法则及对称性 $p(X, Y) = p(Y, X)$，我们立刻可以得到条件概率之间的如下关系：

$$p(Y \mid X) = \frac{p(X \mid Y)\, p(Y)}{p(X)} \tag{1.12}$$

这被称为贝叶斯定理，它在模式识别和机器学习中起着核心作用。利用加法法则，贝叶斯定理中分母可以用分子中出现的量来表示：

$$p(Y \mid X) = \frac{p(X \mid Y)\, p(Y)}{\sum_{Y} p(X \mid Y)\, p(Y)} \tag{1.13}$$

在贝叶斯定理中，分母可以看作是归一化常数，其作用是确保 (1.12) 左边条件概率对所有 Y 的取值求和等于 1。

在图 1.11 中，我们展示了一个关于两个变量的联合分布的简单例子，用来说明边缘分布和条件分布的概念。这里从联合分布中抽取了 $N = 60$ 个数据点，显示在左上角。右上角给出了 Y 取两个不同值时数据点比例的直方图。根据概率的定义，当 $N \to \infty$ 时，这些比例将等于相应的概率 $p(Y)$。我们可以将直方图视为一种简单的方法，用有限个从分布中抽取的点来建模概率分布。从数据建模分布是统计模式识别的核心问题，并将在本书中深入探讨。图 1.11 中剩下的两幅图展示了 $p(X)$ 和 $p(X \mid Y = 1)$ 的直方图估计。

现在让我们回到水果盒子的例子。此时我们再次明确地区分随机变量及其取值。我们已经看到，选择红色盒子或蓝色盒子的概率分别为

$$p(B = r) = 4/10 \tag{1.14}$$

$$p(B = b) = 6/10 \tag{1.15}$$

分别如上所示。需要注意，它们满足 $p(B = r) + p(B = b) = 1$。

现在假设我们随机选择了一个盒子，结果发现是蓝色盒子。那么选择到苹果的概率就是蓝色盒子中苹果的比例，即 $3/4$，因此 $p(F = a \mid B = b) = 3/4$ 。事实上，我们可

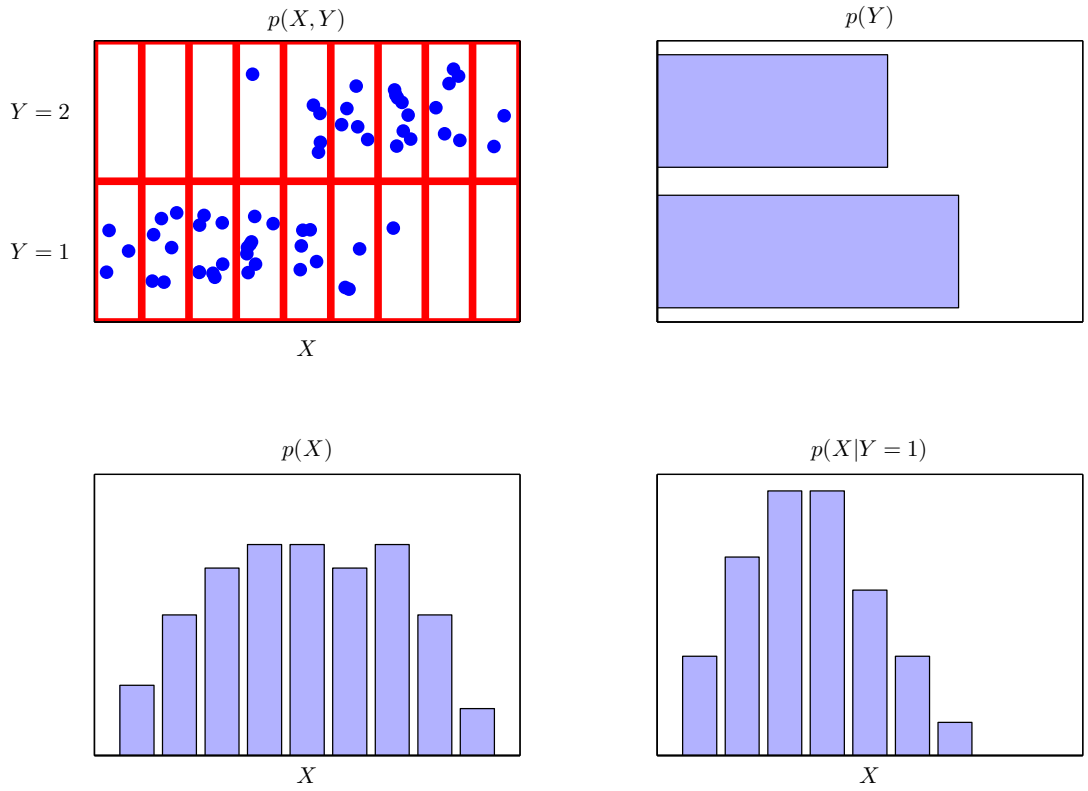

图 1.11 关于两个变量分布的示意图：变量 X 取 9 个可能值，变量 Y 取 2 个可能值。左上图展示了从这两个变量的联合概率分布中抽取的 60 个样本点。其余图则展示了边缘分布 $p(X)$ 和 $p(Y)$ 的直方图估计，以及条件分布 $p(X|Y=1)$，该条件分布对应于左上图的最下一行。

以写出在给定选中盒子的情况下，水果种类的四个条件概率：

$$p(F=a \mid B=r)=1/4 \tag{1.16}$$

$$p(F=o \mid B=r)=3/4 \tag{1.17}$$

$$p(F=a \mid B=b)=3/4 \tag{1.18}$$

$$p(F=o \mid B=b)=1/4 \tag{1.19}$$

同样需要注意，这些概率是归一化的，因此满足

$$p(F=a \mid B=r)+p(F=o \mid B=r)=1 \tag{1.20}$$

$$p(F=a \mid B=b)+p(F=o \mid B=b)=1 \tag{1.21}$$

现在我们可以利用概率的加法法则和乘法法则来计算选择苹果的总体概率：

$$\begin{aligned} p(F=a) &= p(F=a \mid B=r)\,p(B=r)+p(F=a \mid B=b)\,p(B=b) \\ &= \frac{1}{4}\cdot\frac{4}{10}+\frac{3}{4}\cdot\frac{6}{10}=\frac{1}{10}+\frac{9}{20}=\frac{11}{20} \end{aligned} \tag{1.22}$$

由此可得，利用加法法则（sum rule），橙子的概率为 $p(F=o)=1-\frac{11}{20}=\frac{9}{20}$。

现在假设相反的情形：我们只被告知选出了一件水果，并且它是橙子，我们想知道它来自哪个盒子。这就需要评估在给定水果种类的条件下盒子的概率分布（probability

distribution），而 (1.16)–(1.19) 给出的则是在给定盒子的条件下水果的概率分布。我们可以利用贝叶斯定理（Bayes' theorem）来求解这种条件概率的反转，得到

$$p(B=r|F=o)=\frac{p(F=o|B=r)p(B=r)}{p(F=o)}=\frac{3}{4}\times\frac{4}{10}\times\frac{20}{9}=\frac{2}{3} \tag{1.23}$$

由加法法则（sum rule）可得 $p(B=b\mid F=o)=1-\frac{2}{3}=\frac{1}{3}$。

我们可以对贝叶斯定理给出一个重要的解释。如果在得知选出的水果种类之前，我们被问到选择的是哪个盒子，那么我们能利用的最完整信息就是概率 $p(B)$。我们称其为先验概率，因为它是在观察水果种类之前可用的概率。而一旦得知选出的水果是橙子，我们就可以利用贝叶斯定理计算 $p(B\mid F)$，我们称其为后验概率，因为它是在观察到 F 之后得到的概率。需要注意的是，在这个例子中，选择红色盒子的先验概率是 $4/10$，因此在未观察水果前，更可能选择的是蓝色盒子。然而，当我们观察到选出的水果是橙子后，红色盒子的后验概率变为 $2/3$，此时更可能选中的是红色盒子。这一结果与直觉一致，因为红色盒子中橙子的比例远高于蓝色盒子，因此观察到橙子为红色盒子提供了强有力的支持。实际上，这一证据强到足以推翻先验，使得红色盒子被选中的可能性超过蓝色盒子。

最后需要指出，如果两个变量的联合分布可以分解为各自边缘分布的乘积，即 $p(X,Y)=p(X)\,p(Y)$，那么称 X 与 Y 相互独立。由乘法法则可知，此时 $p(Y\mid X)=p(Y)$，因此在给定 X 的条件下，Y 的条件分布确实与 X 的取值无关。举例来说，在水果盒子的例子中，如果每个盒子中苹果和橙子的比例相同，那么就有 $p(F\mid B)=p(F)$，于是选择苹果的概率与选择哪个盒子无关。

1.2.1 概率密度

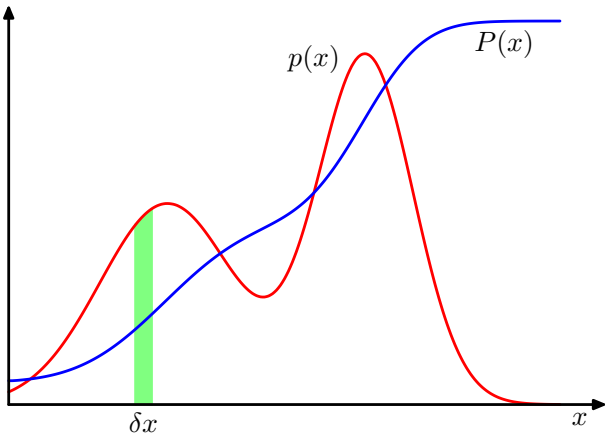

图 1.12 离散变量的概率概念可以扩展为连续变量 x 上的概率密度 $p(x)$，此时 x 落在区间 $(x,x+\delta x)$ 内的概率为 $p(x)\delta x$，当 $\delta x\to 0$。概率密度可以表示为累积分布函数 $P(x)$ 的导数。

除了考虑定义在离散事件集合上的概率之外，我们还希望研究关于连续变量的概率。这里我们将仅作相对非正式的讨论。若实值变量 x 落在区间 $(x,x+\delta x)$ 内的概率在 $\delta x\to 0$ 时由 $p(x)\delta x$ 给出，那么 $p(x)$ 就称为 x 上的概率密度。这一点在图 1.12 中有所

展示。变量 x 落在区间 (a, b) 内的概率为

$$p(a \in (a, b)) = \int_a^b p(x)\, dx \tag{1.24}$$

由于概率必须为非负数，并且 x 的取值必然落在实数轴上的某处，概率密度 $p(x)$ 必须满足以下两个条件：

$$p(x) \geq 0 \tag{1.25}$$

$$\int_{-\infty}^{\infty} p(x)\, dx = 1 \tag{1.26}$$

在非线性变量变换下，概率密度的变化方式不同于普通函数，这是由于雅可比因子的存在。例如，若我们考虑变量变换 $x = g(y)$，那么一个函数 $f(x)$ 会变为 $\bar{f}(y) = f(g(y))$。现在考虑一个概率密度 $p_x(x)$，它对应于新变量 y 下的概率密度 $p_y(y)$，这里下标表示 $p_x(x)$ 和 $p_y(y)$ 是不同的密度。落在区间 $(x, x + \delta x)$ 内的观测点，在 δx 很小时，会被映射到区间 $(y, y + \delta y)$，并且满足 $p_x(x)\, \delta x \simeq p_y(y)\, \delta y$，因此有

$$\begin{aligned} p_y(y) &= p_x(x) \left| \frac{dx}{dy} \right| \\ &= p_x(g(y))\, |g'(y)| \end{aligned} \tag{1.27}$$

这一性质的一个结果是：概率密度的最大值这一概念依赖于变量的选择。

变量 x 落在区间 $(-\infty, z)$ 内的概率由累积分布函数定义如下：

$$P(z) = \int_{-\infty}^{z} p(x)\, \mathrm{d}x \tag{1.28}$$

它满足 $P'(x) = p(x)$, 如图 1.12 所示。

如果我们有多个连续变量 $x\_1, \ldots, x\_D$，并将它们统称为向量 $\mathbf{x}$，则可以定义联合概率密度（joint probability density）$p(\mathbf{x}) = p(x_1, \ldots, x_D)$, 使得 $\mathbf{x}$ 落在包含点 $\mathbf{x}$ 的无穷小体积 $\delta\mathbf{x}$ 内的概率为 $p(\mathbf{x}), \delta\mathbf{x}$。这种多元概率密度必须满足

$$p(\mathbf{x}) \geq 0 \tag{1.29}$$

$$\int p(\mathbf{x})\, d\mathbf{x} = 1 \tag{1.30}$$

其中积分在整个 $\mathbf{x}$ 空间上进行。我们也可以考虑同时包含离散变量和连续变量的联合概率分布。需要注意的是，如果 x 是离散变量，那么 $p(x)$ 有时被称为概率质量函数，因为它可以看作是集中在 x 的允许取值上的一组“概率质量”。

概率的加法法则和乘法法则，以及贝叶斯定理，同样适用于概率密度，或者离散与连续变量的组合。例如，如果 x 和 y 是两个实值变量，那么加法法则和乘法法则形式为

$$p(x) = \int p(x, y)\, \mathrm{d}y \tag{1.31}$$

$$p(x, y) = p(y \mid x)\, p(x) \tag{1.32}$$

对连续变量的加法法则和乘法法则的严格证明（Feller, 1966）需要用到数学中的测度论，这超出了本书的讨论范围。不过，可以通过一种非正式的方式理解其合理性：将每个实值变量划分为宽度为 Δ 的区间，并考虑这些区间上的离散概率分布。当取极限 $\Delta \to 0$ 时，求和就转化为积分，从而得到所需的结果。

1.2.2 期望与协方差

概率运算中最重要的操作之一是求函数的加权平均值。某个函数 $f(x)$ 在概率分布 $p(x)$ 下的平均值称为 $f(x)$ 的期望，记作 $\mathbb{E}[f]$。对于离散分布，其定义为

$$\mathbb{E}[f] = \sum_{x} p(x)\, f(x) \tag{1.33}$$

因此，平均值是由不同 x 取值的相对概率加权得到的。对于连续变量，期望可以通过对相应概率密度的积分来表示：

$$\mathbb{E}[f] = \int p(x)\, f(x)\, dx \tag{1.34}$$

在这两种情况下，如果我们给定从概率分布或概率密度中抽取的有限个 N 个样本点，那么期望可以用这些点的有限求和来近似表示：

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n) \tag{1.35}$$

在第 11 章讨论采样方法时，我们将广泛使用这一结果。当 $N \to \infty$ 时，(1.35) 式中的近似就变为精确结果。

有时我们会考虑多变量函数的期望，这种情况下可以用下标来指示对哪个变量取平均。例如：

$$\mathbb{E}_x[f(x, y)] \tag{1.36}$$

这表示函数 $f(x, y)$ 关于 x 的分布取平均。需要注意，$\mathbb{E}_x[f(x, y)]$ 将是 y 的一个函数。我们还可以考虑关于条件分布的条件期望，其形式为

$$\mathbb{E}[f \mid y] = \int p(x \mid y)\, f(x)\, \mathrm{d}x, \tag{1.37}$$

在连续变量情况下有类似的定义。函数 $f(x)$ 的方差定义为

$$\mathrm{var}[f] = \mathbb{E}\big[(f(x) - \mathbb{E}[f(x)])^2\big] \tag{1.38}$$

它刻画了函数 $f(x)$ 围绕其均值 $\mathbb{E}[f(x)]$ 的波动程度。将平方项展开后，可以看到方差也可以用 $f(x)$ 和 $f(x)^2$ 的期望来表示：

$$\mathrm{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \tag{1.39}$$

特别地，变量 x 本身的方差可以写作

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 \tag{1.40}$$

对于两个随机变量 x 和 y，协方差定义为

$$\begin{aligned}\mathrm{cov}[x,y] &= \mathbb{E}_{x,y}[\{x-\mathbb{E}[x]\}\{y-\mathbb{E}[y]\}] \\ &= \mathbb{E}_{x,y}[xy]-\mathbb{E}[x]\mathbb{E}[y]\end{aligned} \tag{1.41}$$

它表示 x 和 y 共同变化的程度。如果 x 和 y 相互独立，那么它们的协方差为零。

对于两个随机向量 x 和 y，协方差是一个矩阵：

$$\begin{aligned}\mathrm{cov}[\mathbf{x},\mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x}-\mathbb{E}[\mathbf{x}]\}\left\{\mathbf{y}^T-\mathbb{E}[\mathbf{y}^T]\right\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^T]-\mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^T].\end{aligned} \tag{1.42}$$

如果我们考虑向量 x 各个分量之间的协方差，那么我们使用一个稍微简单的记号 $\mathrm{cov}[\mathbf{x}] \equiv \mathrm{cov}[\mathbf{x},\mathbf{x}]$。

1.2.3 贝叶斯概率

到目前为止，在本章中我们将概率看作是随机、可重复事件出现频率的角度来理解的。我们称这种解释为经典或频率派的概率解释。现在我们转向更一般的贝叶斯观点，在这种观点下，概率被用来刻画不确定性的大小。

考虑一个不确定的事件，例如月球是否曾经独自绕太阳公转，或者在本世纪末北极冰盖是否会消失。这些事件无法像之前水果盒子的例子那样，通过大量重复试验来定义概率。尽管如此，我们通常还是会有一些判断，比如我们认为极地冰层融化的速度有多快。如果我们获得了新的证据，例如来自新的地球观测卫星的全新诊断信息，我们可能会修正自己对冰层消融速度的看法。对这些问题的评估会影响我们所采取的行动，例如我们会在多大程度上努力减少温室气体的排放。在这种情况下，我们希望能够量化这种不确定性的表述，并在新的证据出现时对不确定性进行精确修正，同时因此能够采取最优的行动或决策。这一切都可以通过优雅而又非常通用的贝叶斯概率解释来实现。

然而，将概率用于表示不确定性并不是一种临时的随意选择，而是在我们希望在进行理性且自洽的推理时尊重常识的必然结果。举例来说，Cox (1946) 证明了：如果用数值来表示信念的程度，那么只要满足一组简单的公理，这些公理编码了这种信念应当符合的常识性性质，就会唯一地导出一套操作信念程度的法则，而这套法则恰好等价于概率的加法和乘法法则。这为概率论可以被看作布尔逻辑在涉及不确定性情形下的扩展提供了第一个严格的证明 (Jaynes, 2003)。许多其他作者也提出了不同的不确定性度量应当满足的性质或公理体系 (Ramsey, 1931; Good, 1950; Savage, 1961; de Finetti, 1970; Lindley, 1982)。在每种情况下，所得到的数值量都严格遵循概率的法则。因此，把这些数值称为（贝叶斯）概率是非常自然的。

在模式识别领域中，拥有一个更一般化的概率概念同样是有益的。以第 1.1 节讨论的多项式曲线拟合为例，将频率派的概率观念应用到观测变量 t_n 的随机取值似乎是合理的。然而，我们还希望能够处理并量化围绕模型参数 $\mathbf{w}$ 的合适选择所带来的不确定性。我们将看到，从贝叶斯 (Bayesian) 的角度出发，可以利用概率论的工具来刻画模型参数（例如 w）的不确定性，甚至是模型选择本身所带来的不确定性。

贝叶斯定理在这里获得了新的意义。回忆水果盒子的例子：对水果种类的观测提供了相关信息，从而改变了所选盒子是红色盒子的概率。在那个例子中，贝叶斯定理被用来将先验概率转化为后验概率，其方式是结合观测数据所提供的证据。正如我们稍后将详细看到的，在对某些量进行推断时（例如多项式曲线拟合中的参数 **w**），我们可以采取类似的方法。在观测数据之前，我们通过一个先验概率分布 $p(\mathbf{w})$ 来表达对 **w** 的假设。观测到的数据 $\mathcal{D} = \{t_1, \ldots, t_N\}$ 的作用则通过条件概率 $p(\mathcal{D}|\mathbf{w})$ 来体现，而我们将在第 1.2.5 节看到如何将其显式地表示出来。贝叶斯定理的形式为

$$p(\mathbf{w}|D) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \tag{1.43}$$

然后，这一定理就使我们能够在观测到数据 $\mathcal{D}$ 之后，以后验概率 $p(\mathbf{w}|\mathcal{D})$ 的形式来评估参数 w 的不确定性。

在贝叶斯定理右边的项中，$p(\mathcal{D}|\mathbf{w})$ 是针对观测数据集 $\mathcal{D}$ 计算的，并且可以被看作是参数向量 **w** 的函数，这时它被称为似然函数。它刻画了在不同的参数向量 **w** 取值下，观测到数据集的可能性有多大。需要注意的是，似然函数并不是关于 **w** 的概率分布，它对 **w** 的积分并不（必然）等于 1。

有了这种可能性的定义，我们可以用文字来表述贝叶斯定理

$$\text{后验分布} \propto \text{似然函数} \times \text{先验分布} \tag{1.44}$$

这里所有的量都被看作是关于 **w** 的函数。(1.43) 式中的分母是归一化常数，它保证左边的后验分布是一个有效的概率密度函数，并且积分为 1。实际上，对 (1.43) 式两边关于 **w** 积分，我们可以将贝叶斯定理中的分母表示为先验分布和似然函数的形式：

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})\, p(\mathbf{w})\, \mathrm{d}\mathbf{w} \tag{1.45}$$

在贝叶斯和频率派这两种范式中，似然函数 $p(\mathcal{D}|\mathbf{w})$ 都扮演着核心角色。然而，两者使用它的方式在根本上是不同的。在频率派的设定下，**w** 被视为一个固定的参数，其值通过某种“估计量”来确定，而该估计值的不确定性（误差区间）是通过考虑所有可能的数据集 $\mathcal{D}$ 的分布来获得的。相比之下，从贝叶斯的角度看，数据集 $\mathcal{D}$ 只有一个（也就是实际观测到的那个），参数的不确定性则通过关于 **w** 的概率分布来表达。

一种被广泛使用的频率派估计量是最大似然，其中参数 $mathbfw$ 被设定为使似然函数 $p(\mathcal{D}|mathbfw)$ 取最大值的那个值。这等价于选择能够使观测数据集概率最大化的参数 $mathbfw$。在机器学习文献中，似然函数的负对数被称为误差函数。由于负对数函数是单调递减的，最大化似然等价于最小化误差。

一种确定频率派误差区间的方法是自助法 (Efron, 1979; Hastie et al., 2001)。其过程如下：假设我们的原始数据集包含 N 个数据点 $\mathbf{X} = \{x_1, \ldots, x_N\}$。我们可以通过从 $\mathbf{X}$ 中有放回地随机抽取 N 个点来构造一个新的数据集 $\mathbf{X}_B$。这样，X 中的一些点可能在 $\mathbf{X}_B$ 中被重复出现，而另一些点可能没有出现在 $\mathbf{X}S_B$ 中。这个过程可以重复 L 次，从而生成 L 个大小为 N 的数据集，每个数据集都是从原始数据集 X 抽样得到的。随后，可以通过考察不同自助数据集之间预测结果的差异性，来评估参数估计的统计精度。

贝叶斯观点的一个优势在于，先验知识的引入是自然而然的。举个例子，假设一枚看起来公平的硬币被抛掷三次，并且每次都朝上。如果采用经典的最大似然方法来估计正面朝上的概率，结果会得到 1，这意味着所有未来的抛掷都会是正面！相比之下，贝叶斯方法结合任意一个合理的先验，都会得出一个远不如此极端的结论。

关于频率派与贝叶斯这两种范式的相对优劣，一直存在大量的争议和辩论，而这种情况更加复杂的原因在于，既不存在唯一的频率派观点，也不存在唯一的贝叶斯观点。例如，对贝叶斯方法的一个常见批评是：先验分布往往是基于数学上的方便性而选择的，而不是对任何真实先验信念的反映。甚至，由于结论依赖于先验的选择而具有主观性，这一点也被一些人视为问题所在。减少对先验依赖性的需求，正是所谓非信息先验的动机之一。然而，这类方法在模型比较时会带来困难；事实上，如果先验选择不当，基于贝叶斯方法的推断可能会在高置信度下给出很差的结果。频率派的评估方法在一定程度上为此类问题提供了一些保护，而诸如交叉验证等技术在模型比较等方面仍然是非常有用的。

本书重点强调贝叶斯观点，这反映了近年来贝叶斯方法在实际应用中的重要性正迅速增长，同时也会在需要的地方讨论一些有用的频率派概念。

尽管贝叶斯框架起源于 18 世纪，但在很长一段时间里，贝叶斯方法的实际应用都受到严重限制，其原因在于难以完整地执行贝叶斯过程，尤其是需要在整个参数空间上进行边缘化 (求和或积分)。正如我们将看到的，这一步在进行预测或比较不同模型时是必需的。采样方法的发展，例如马尔可夫链蒙特卡罗 (第 11 章将讨论)，以及计算机在速度和存储容量上的巨大提升，使得贝叶斯技术在广泛的问题领域中得以实际应用。蒙特卡罗方法非常灵活，可以应用于多种模型。然而，它们计算量巨大，因此主要被用于小规模问题。

近年来，高效的确定性近似方法得到了发展，例如变分贝叶斯和期望传播 (第 10 章将讨论)。这些方法为采样方法提供了互补的替代方案，并使得贝叶斯技术能够应用于大规模的问题 (Blei et al., 2003)。

1.2.4 高斯分布

我们将在第 2 章中专门研究各种概率分布及其关键性质。然而，在这里先介绍一种最重要的连续型概率分布——正态分布或高斯分布——是很方便的。在本章的后续部分乃至本书的大部分内容中，我们都会广泛使用这一分布。

对于单个实值变量 x，高斯分布定义为

$$\mathcal{N}(x|\mu,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{1.46}$$

该分布由两个参数控制：μ，称为均值；以及 σ^2，称为方差。方差的平方根记为 σ，称为标准差；方差的倒数记为 $\beta=1/\sigma^2$，称为精度。我们很快将看到这些术语的动机。图 1.13 展示了高斯分布的曲线图。

由 (1.46) 式的形式可以看出，高斯分布满足

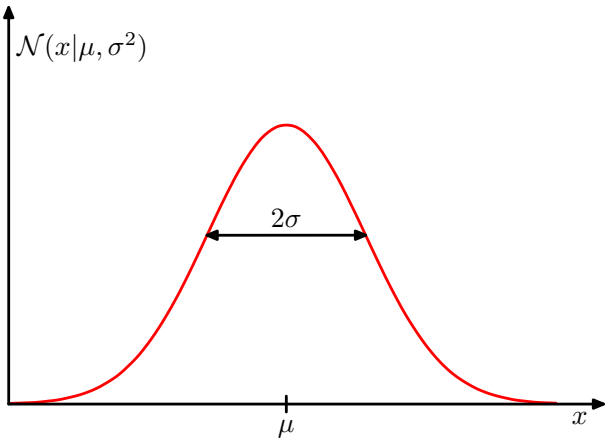

图 1.13 一维高斯分布的示意图，展示了均值 μ 和标准差 σ。

$$\mathcal{N}(x|\mu,\sigma^2) > 0 \tag{1.47}$$

此外，很容易证明高斯分布是归一化的，因此

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)\,\mathrm{d}x = 1 \tag{1.48}$$

因此，(1.46) 式满足一个有效概率密度的两个要求。

我们可以很容易地求出在高斯分布下函数 x 的期望。特别地，x 的平均值由下式给出：

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} x\,\mathcal{N}(x|\mu,\sigma^2)\,\mathrm{d}x = \mu \tag{1.49}$$

由于参数 μ 表示在该分布下 x 的平均值，因此称其为均值。类似地，对于二阶矩：

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)\,dx = \mu^2 + \sigma^2 \tag{1.50}$$

由 (1.49) 和 (1.50) 可得，x 的方差为

$$\mathrm{var}[x] = \mathbb{E}[x^2] - (\mathbb{E}[x])^2 = \sigma^2 \tag{1.51}$$

因此，σ^2 被称为方差参数。分布的最大值称为众数 (mode)。对于高斯分布，众数与均值是一致的。

我们同样关心定义在 D 维连续变量向量 $\mathbf{x}$ 上的高斯分布，其形式为

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{1.52}$$

其中，D 维向量 $\boldsymbol{\mu}$ 称为均值，$D \times D$ 的矩阵 Σ 称为协方差，而 $|\Sigma|$ 表示 Σ 的行列式。在本章中我们会简单使用多元高斯分布，而它的性质将在第 2.3 节中进行详细研究。

现在假设我们有一个观测数据集 $\mathbf{x} = (x_1,\ldots,x_N)^{\mathrm{T}}$，表示对标量变量 x 的 N 次观测。需要注意的是，我们使用字体 $\mathbf{x}$ 来与单个向量值变量的观测 $(x_1,\ldots,x_D)^{\mathrm{T}}$ 区分开，后者记为 $\mathbf{x}$。我们假设这些观测是从一个均值 μ 和方差 σ^2 都未知的高斯分布中独立抽取的，而我们的目标是根据数据集来确定这些参数。如果数据点是从同一分布中独立抽

取的，就称它们为独立同分布 (i.i.d.)。我们已经看到，两个独立事件的联合概率等于它们各自边缘概率的乘积。由于数据集 $\mathbf{x}$ 是 i.i.d. 的，因此在给定 μ 和 σ^2 的条件下，数据集的概率可以写成：

$$p(\mathbf{x}|\mu,\sigma^2)=\prod_{n=1}^{N}\mathcal{N}(x_n|\mu,\sigma^2) \tag{1.53}$$

当把它视为 μ 和 σ^2 的函数时，它就是高斯分布的似然函数。图 1.14 对其作了示意性的解释。

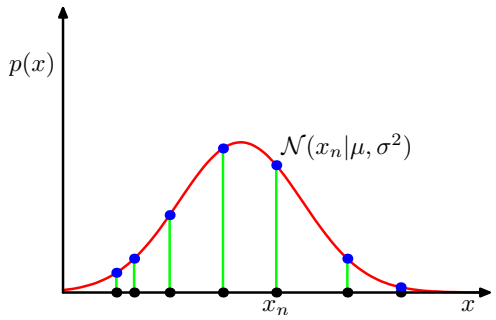

图 1.14 高斯分布似然函数的示意图由红色曲线表示。黑点表示数据集 x_n，公式 1.53 给出的似然函数对应于这些蓝色数值的乘积。最大化似然的过程就是通过调整高斯分布的均值和方差，使得这个乘积最大化。

利用观测数据集来确定概率分布参数的一个常见准则是，寻找能使似然函数最大化的参数值。乍一看，这似乎是一个奇怪的准则，因为根据我们之前对概率论的讨论，更自然的做法似乎是最大化在给定数据条件下参数的概率，而不是在给定参数条件下数据的概率。事实上，这两个准则是相关的，我们将在曲线拟合的背景下对此进行讨论。

目前，我们将通过最大化似然函数 (1.53) 来确定高斯分布中未知参数 μ 和 σ^2 的取值。在实际操作中，更方便的方法是最大化似然函数的对数。由于对数函数是其自变量的单调递增函数，最大化函数的对数等价于最大化函数本身。

取对数不仅简化了后续的数学分析，同时在数值计算上也更有利，因为大量小概率值的乘积很容易导致计算机的数值精度下溢，而改为计算对数概率的和则可以避免这一问题。

由 (1.46) 和 (1.53) 可得，对数似然函数可以写成如下形式：

$$\ln p(\mathbf{x}|\mu,\sigma^2)=-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2-\frac{N}{2}\ln\sigma^2-\frac{N}{2}\ln(2\pi) \tag{1.54}$$

对 (1.54) 式关于 μ 最大化，可得最大似然解：

$$\mu_{\mathrm{ML}}=\frac{1}{N}\sum_{n=1}^{N}x_n \tag{1.55}$$

这就是样本均值 (sample mean)，即观测值 $\{x_n\}$ 的平均值。类似地，对 (1.54) 式关于 σ^2 最大化，可得方差的最大似然解为

$$\sigma^2_{\mathrm{ML}}=\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2 \tag{1.56}$$

这就是关于样本均值 μ_{ML} 计算的样本方差。需要注意的是，我们实际上是对 (1.54) 式关于 μ 和 σ^2 进行联合最大化，但在高斯分布的情形下，μ 的解与 σ^2 的解是解耦的，因此我们可以先计算 (1.55)，然后再利用该结果去计算 (1.56)。

在本章后续部分以及接下来的章节中，我们将强调最大似然方法的显著局限性。这里我们先通过单变量高斯分布参数最大似然解的情境来说明这一问题。特别地，我们将展示最大似然方法会系统性地低估分布的方差。这是一种称为偏差的现象，并且与多项式曲线拟合中遇到的过拟合问题相关。首先注意到，最大似然解 μ_{ML} 和 σ^2_{ML} 都是数据集取值 $x_1, \ldots, x_N$ 的函数。而这些数据集本身是从一个参数为 μ 和 σ^2 的高斯分布中抽取的。考虑这些量相对于数据集取值的期望，可以很容易地证明：

$$\mathbb{E}[\mu_{\text{ML}}] = \mu \tag{1.57}$$

$$\mathbb{E}[\sigma^2_{\text{ML}}] = \frac{N-1}{N}\sigma^2 \tag{1.58}$$

因此，平均而言，最大似然估计能够得到正确的均值，但会以一个 $(N-1)/N$ 的因子低估真实方差。其直观解释如图 1.15 所示。

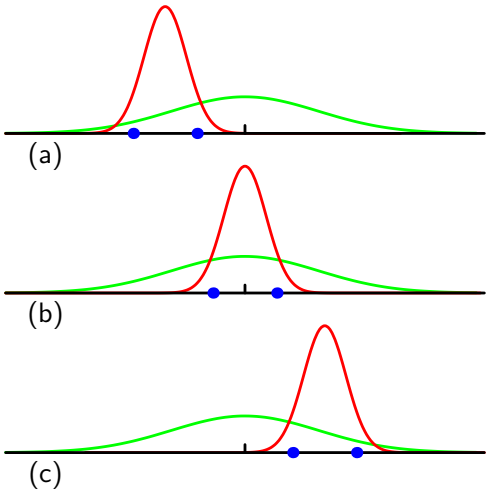

图 1.15 示意图展示了在使用最大似然方法确定高斯分布方差时偏差产生的原因。绿色曲线表示生成数据的真实高斯分布，三条红色曲线表示分别拟合三个数据集得到的高斯分布，每个数据集包含两个蓝色数据点，拟合采用最大似然结果 1.55 和 1.56。在三个数据集上取平均后，均值是正确的，但方差被系统性低估，因为它是相对于样本均值而不是相对于真实均值来计算的。

由 (1.58) 可得，下列方差参数的估计是无偏的：

$$\tilde{\sigma}^2 = \frac{1}{N-1}\sum_{n=1}^{N}\sigma^2_{ML} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{\text{ML}})^2 \tag{1.59}$$

在第 10.1.3 节中，我们将看到，当采用贝叶斯方法时，这一结果会自然而然地出现。

需要注意的是，随着数据点数量 N 的增加，最大似然解的偏差会逐渐减小，当 $N \to \infty$ 时，方差的最大似然解将等于生成数据的分布的真实方差。在实际应用中，除了在 N 很小的情况下，这种偏差通常不会成为严重问题。然而，在本书中我们将关注更复杂的、多参数的模型，在这种情形下，与最大似然相关的偏差问题会更加严重。事

实上，正如我们将看到的，最大似然中的偏差问题正是之前在多项式曲线拟合中所遇到的过拟合问题的根源所在。

1.2.5 曲线拟合再探

我们已经看到，多项式曲线拟合问题可以用误差最小化来表达。这里我们重新回到曲线拟合的例子，从概率的角度来考察它，从而加深对误差函数和正则化的理解，并逐步走向完整的贝叶斯处理。

在曲线拟合问题中，我们的目标是：基于包含 N 个输入值的数据集 $\mathbf{x} = (x_1, \ldots, x_N)^{\mathrm{T}}$ 及其对应的目标值 $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$，能够在给定新的输入变量 x 时，对目标变量 t 做出预测。我们可以通过一个概率分布来表达对目标变量取值的不确定性。为此，我们假设：在给定 x 的条件下，对应的 t 服从高斯分布，其均值等于由 (1.1)式给出的多项式曲线 $y(x, \mathbf{w})$ 的值。于是有：

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t \mid y(x, \mathbf{w}), \beta^{-1}\right) \tag{1.60}$$

在这里，为了与后续章节的记号保持一致，我们定义了一个精度参数 β，它对应于分布方差的倒数。这在图1.16中作了示意性的说明。

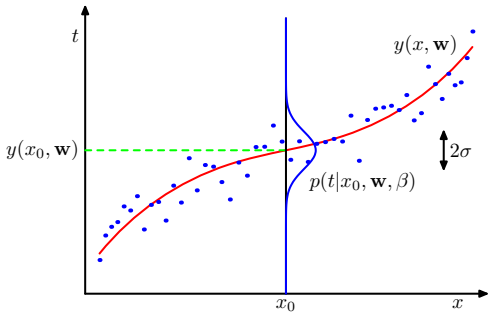

图 1.16 高斯条件分布的示意图，表示在给定 x 的情况下 t 的分布（公式 (1.60)）。其中，均值由多项式函数 $y(x, \mathbf{w})$ 给出，精度由参数 β 决定，并且与方差的关系为 $\beta^{-1} = \sigma^2$

我们现在利用训练数据 $\{\mathbf{x}, \mathbf{t}\}$ 来通过最大似然的方法确定未知参数 $\mathbf{w}$ 和 β 的取值。若假设数据是从分布 (1.60) 中独立抽取的，则其似然函数为

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid y(x_n, \mathbf{w}), \beta^{-1}\right) \tag{1.61}$$

正如之前在简单高斯分布情形中所做的那样，这里最大化似然函数的对数会更加方便。将 (1.46) 给出的高斯分布形式代入后，可以得到对数似然函数为

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi). \tag{1.62}$$

首先考虑多项式系数的最大似然解，记为 $\mathbf{w}_{\mathrm{ML}}$。这些系数通过对 (1.62) 式关于 $\mathbf{w}$ 最大化得到。为此，我们可以忽略 (1.62) 式右边最后两项，因为它们与 $\mathbf{w}$ 无关。同时需要注意的是，将对数似然乘上一个正的常数系数并不会改变其关于 $\mathbf{w}$ 的极大值位置，因

此我们可以把系数 $\beta/2$ 替换为 $1/2$。最后，与其最大化对数似然，我们也可以等价地最小化负对数似然。因此，就确定 $\mathbf{w}$ 而言，最大化似然等价于最小化由 (1.2) 定义的平方和误差函数。由此可见，平方和误差函数的出现，正是高斯噪声分布假设下最大化似然的结果。

我们同样可以利用最大似然来确定高斯条件分布的精度参数 β。对 (1.62) 式关于 β 最大化可得：

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\}^2 \tag{1.63}$$

同样地，我们可以先确定控制均值的参数向量 $\mathbf{w}_{\mathrm{ML}}$，随后再利用这一结果去求解精度 β_{ML}，这与简单高斯分布的情形是相同的。

在确定了参数 $\mathbf{w}$ 和 β 之后，我们就可以对新的 x 值进行预测。由于我们现在有了一个概率模型，因此预测结果用预测分布来表示，即给出 t 的概率分布，而不仅仅是一个点估计。通过将最大似然参数代入 (1.60)，得到：

$$p(t|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\big(t \mid y(x, \mathbf{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1}\big) \tag{1.64}$$

现在让我们向更加贝叶斯化的方法迈进一步，在多项式系数 $\mathbf{w}$ 上引入一个先验分布。为简单起见，我们考虑如下形式的高斯分布：

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right\} \tag{1.65}$$

其中，α 是该分布的精度，而对于 M 次多项式，向量 $\mathbf{w}$ 中的元素总数为 $M+1$。像 α 这样控制模型参数分布的变量称为超参数。利用贝叶斯定理，$\mathbf{w}$ 的后验分布正比于先验分布与似然函数的乘积：

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)\, p(\mathbf{w}|\alpha) \tag{1.66}$$

我们现在可以通过在给定数据的条件下寻找 $\mathbf{w}$ 的最可能取值来确定参数 $\mathbf{w}$，换句话说，就是最大化后验分布。这种方法称为最大后验估计。

对 (1.66) 取负对数，并结合 (1.62) 和 (1.65)，可以得到后验分布最大值对应于以下表达式的最小值：

$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \tag{1.67}$$

因此我们看到，最大化后验分布等价于最小化之前在 (1.4) 中出现过的正则化平方和误差函数，其正则化参数为 $\lambda = \alpha/\beta$ 。

1.2.6 贝叶斯曲线拟合

尽管我们已经引入了先验分布 $p(\mathbf{w}|\alpha)$，但到目前为止我们仍然是在对 $\mathbf{w}$ 做点估计，因此这还不能算作真正的贝叶斯处理。在完全的贝叶斯方法中，我们应当严格一致地

应用概率的加法和乘法法则，这就要求我们（正如稍后将看到的）需要对所有可能的 **w** 取值进行积分。这种边缘化正是贝叶斯方法在模式识别中的核心所在。

在曲线拟合问题中，我们给定了训练数据 **x** 和 **t**，以及一个新的测试点 x，我们的目标是预测对应的 t 值。因此，我们希望计算预测分布 $p(t|x, \mathbf{x}, \mathbf{t})$。在这里，我们假设参数 α 和 β 是固定且已知的（在后续章节中，我们将讨论如何在贝叶斯框架下从数据中推断这些参数）。

贝叶斯处理只是严格一致地应用概率的加法和乘法法则，这使得预测分布可以写成如下形式：

$$p(t|x, \mathbf{x}, \mathbf{t}, \alpha, \beta) = \int p(t|x, \mathbf{w}, \beta)\, p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta)\, \mathrm{d}\mathbf{w} \tag{1.68}$$

这里，$p(t|x, \mathbf{w})$ 由 (1.60) 给出，而我们为了简化记号省略了对 α 和 β 的依赖。$p(\mathbf{w}|\mathbf{x}, \mathbf{t})$ 是参数的后验分布，可以通过对 (1.66)右边进行归一化得到。

我们将在第 3.3 节中看到，对于诸如曲线拟合这样的例子，其后验分布是一个高斯分布，并且可以解析地求解。类似地，(1.68) 中的积分也可以解析地计算，其结果是预测分布本身也是一个高斯分布，其形式为：

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}\big(t \mid m(x), s^2(x)\big) \tag{1.69}$$

其中，均值和方差分别为

$$m(x) = \beta \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n)\, t_n \tag{1.70}$$

$$s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \boldsymbol{\phi}(x). \tag{1.71}$$

这里矩阵 **S** 给定为

$$\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x)^{\mathrm{T}} \tag{1.72}$$

其中，**I** 是单位矩阵，并且我们定义向量 $\boldsymbol{\phi}(x)$，其元素为 $\phi_i(x) = x^i,\ i = 0, \ldots, M$。

我们可以看到，在 (1.69) 中预测分布的方差和均值一样，都是依赖于 x 的。式 1.71 中的第一项表示目标变量上的噪声所导致的对 t 预测值的不确定性，这在最大似然预测分布 (1.64) 中已经通过 β_{ML}^{-1} 表达出来。然而，第二项来源于参数 **w** 的不确定性，这是贝叶斯处理的结果。图 1.17 展示了合成正弦回归问题的预测分布。

1.3 模型选择

在利用最小二乘法进行多项式曲线拟合的例子中，我们看到存在一个最优的多项式阶数，它能带来最佳的泛化性能。多项式的阶数控制了模型中自由参数的数量，从而决定了模型的复杂度。在正则化最小二乘法中，正则化系数 λ 同样控制了模型的有效复杂度；而在更复杂的模型中，例如混合分布或神经网络，则可能有多个参数共同决定复杂度。在实际应用中，我们需要确定这些参数的取值，其主要目标通常是使模型在新数据上的预测性能达到最佳。此外，除了在给定模型中找到合适的复杂度参数取值之外，我们还可能希望考虑多种不同类型的模型，以便为特定的应用找到最优的模型。

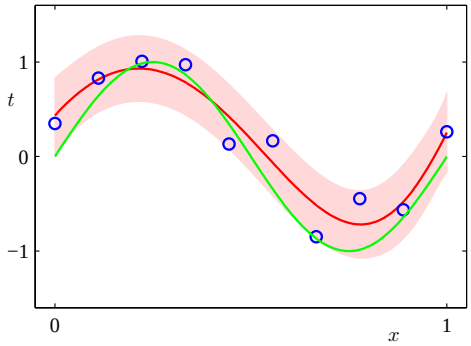

图 1.17 基于贝叶斯方法对 $M = 9$ 多项式进行曲线拟合所得到的预测分布（predictive distribution），其固定参数为 $\alpha = 5 \times 10^{-3}$ 和 $\beta = 11.1$（对应已知的噪声方差）。红色曲线表示预测分布的均值，红色区域表示均值上下一个标准差的范围。

我们已经看到，在最大似然方法中，由于过拟合问题，训练集上的性能并不能很好地反映在未见数据上的预测性能。如果数据量充足，一种方法是将部分可用数据用于训练一系列模型，或者在给定模型下尝试不同复杂度参数的取值，然后在独立数据（有时称为验证集）上对这些模型进行比较，并选择预测性能最优的那个。如果在数据有限的情况下多次迭代模型设计，就可能对验证集产生一定的过拟合，因此通常需要保留第三个独立的测试集，用于对最终选定的模型进行性能评估。

然而，在许多应用中，可用于训练和测试的数据往往有限，而为了建立一个良好的模型，我们希望尽可能多地利用已有数据进行训练。但是，如果验证集太小，它将只能给出一个带有较大噪声的预测性能估计。解决这一困境的一个方法是使用交叉验证，如图 1.18 所示。它允许将可用数据中的 $(S-1)/S$ 部分用于训练，同时利用所有数据来评估性能。当数据极其稀少时，可以考虑取 $S = N$，其中 N 是数据点的总数，这就得到了留一法技术。

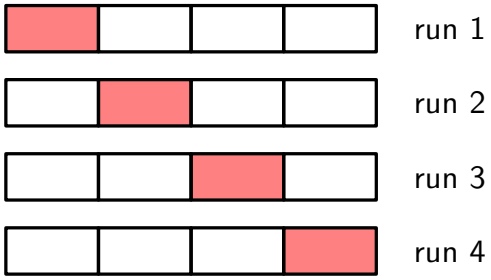

图 1.18 S 折交叉验证的方法，这里以 $S = 4$ 的情况为例进行说明。该方法将可用数据划分为 S 个组（在最简单的情况下，这些组大小相等）。然后使用其中 $S-1$ 个组来训练模型，并在剩下的一个组上进行评估。这个过程会对所有 S 种可能的留出组重复进行，这里用红色方块表示。最终将 S 次运行得到的性能指标取平均。

交叉验证的一个主要缺点是：所需的训练次数增加了 S 倍，对于训练过程本身计算开销很大的模型来说，这可能会成为严重问题。进一步的问题在于，像交叉验证这样依赖独立数据评估性能的方法，如果一个模型包含多个复杂度参数（例如可能存在多个正则化参数），那么探索这些参数组合的过程在最坏情况下可能需要指数级数量的训练次数。显然，我们需要一种更好的方法。理想情况下，这种方法应当只依赖于训练数

据，并且能够在一次训练过程中同时比较多个超参数和不同类型的模型。换句话说，我们需要找到一种仅依赖训练数据的性能度量方式，并且这种方式不会因为过拟合而产生偏差。

在历史上，人们提出了多种“信息准则”，试图通过在最大似然上加入惩罚项来修正其偏差，从而补偿复杂模型的过拟合。例如，赤池信息准则 (Akaike information criterion, AIC) (Akaike, 1974) 选择使下式最大的模型：

$$\ln p(\mathcal{D}|\mathbf{w}_{\mathrm{ML}}) - M \tag{1.73}$$

这里，$p(D|\mathbf{w}_{\mathrm{ML}})$ 表示最佳拟合下的对数似然 (log likelihood)，而 M 是模型中可调参数的个数。该量的一个变体称为贝叶斯信息准则，将在第 4.4.1 节中讨论。然而，这类准则并没有考虑模型参数中的不确定性，因此在实际中往往倾向于选择过于简单的模型。于是，在第 3.4 节中我们将转向一种完全贝叶斯化的处理方法，在那里我们会看到复杂度惩罚是如何以一种自然且有原则的方式出现的。

1.4 维度灾难

在多项式曲线拟合的例子中，我们只有一个输入变量 x。然而，在模式识别的实际应用中，我们必须处理包含多个输入变量的高维空间。正如我们将要讨论的那样，这会带来一些严重的挑战，并且是影响模式识别技术设计的重要因素。

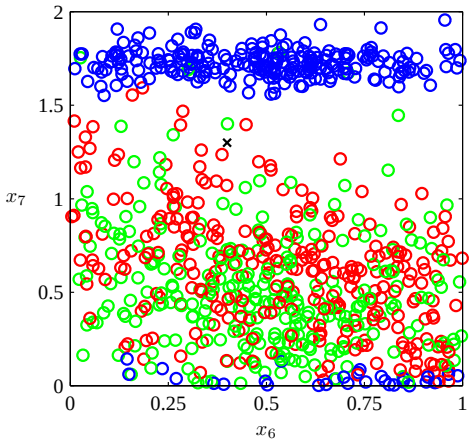

图 1.19 油流数据中输入变量 x_6 和 x_7 的散点图，其中红色表示“homogenous”类别，绿色表示“annular”类别，蓝色表示“laminar”类别。我们的目标是对标记为“×”的新测试点进行分类。

为了说明这个问题，我们考虑一个合成生成的数据集，该数据集代表了从一个包含油、水和气体混合物的管道中获得的测量结果 (Bishop and James, 1993)。这三种物质可以存在于三种不同的几何结构中，分别称为“均匀型”、“环状型”和“层流型”，并且三种物质的比例也可以变化。每个数据点由一个 12 维输入向量组成，这些向量来自伽马射线密度计的测量，该仪器测量伽马射线穿过管道时的衰减情况。该数据集的详细描述见附录 A。图 1.19 展示了该数据集中 100 个点，它们绘制在二维平面上，横纵坐标为其中的两个测量值 x_6 和 x_7（其余 10 个输入值在这里被忽略）。每个数据点都根据其所

属的三种几何类别之一进行了标注。我们的目标是利用这些数据作为训练集，以便能够对新的观测点 (x_6, x_7) 进行分类，例如图 1.19 中用“叉号”标记的那个点。我们可以看到，叉号周围有许多红色点，因此我们可能会猜测它属于红色类别。然而，附近也存在不少绿色点，因此它也可能属于绿色类别。它属于蓝色类别的可能性似乎很小。这里的直觉是：该观测点的类别应当更多地由训练集中邻近点决定，而较远的点影响则较小。事实上，这种直觉是合理的，我们将在后续章节中对其进行更全面的讨论。

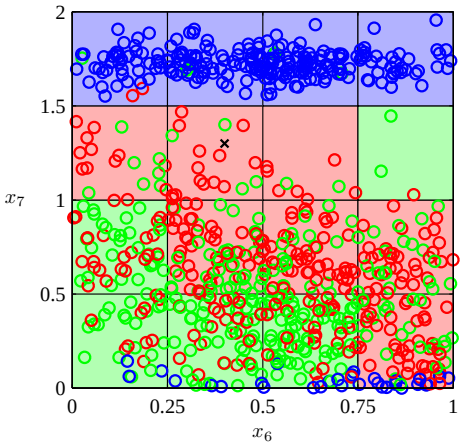

图 1.20 分类问题的一种简单解法示意图：将输入空间划分为若干单元格，并将新的测试点分配到与其所在单元格中占多数的类别。正如我们很快将看到的，这种过于简单的方法存在严重缺陷。

我们如何把这种直觉转化为一个学习算法？一种非常简单的方法是将输入空间划分为规则的网格单元，如图 1.20 所示。给定一个测试点并希望预测其类别时，首先判断它落在哪个单元格中，然后找到所有落在同一单元格内的训练数据点。将测试点的类别预测为：在该单元格中出现数量最多的那一类（若出现并列，则随机打破平局）。

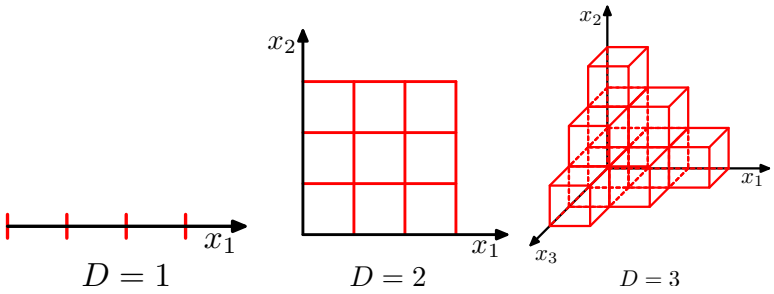

图 1.21 维度灾难的示意图，展示了规则网格的区域数量如何随着空间维度 D 的增加而指数式增长。为了清晰起见，在 $D = 3$ 的情况下仅展示了部分立方区域。

这种朴素方法存在许多问题，其中最严重的一点在把它扩展到具有更多输入变量、也就是更高维的输入空间时会变得显而易见。问题的根源如图1.21所示：如果我们把空间中的一个区域划分为规则的网格单元，那么此类单元的数量会随空间维度呈指数式增长。单元数量指数级增长所带来的问题是：为了确保这些单元不至于为空，我们将需要同样指数级庞大的训练数据量。显然，在超过少数几个变量的空间中，我们不可能采

用这种技术，因此必须寻找更为复杂的方法。

通过回到多项式曲线拟合的例子，并思考如何将其扩展到具有多个输入变量的情形，我们可以进一步理解高维空间中的问题。如果我们有 D 个输入变量，那么一个最高为三次的多项式的一般形式为：

$$y(\mathbf{x}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} w_{ijk} x_i x_j x_k \tag{1.74}$$

随着 D 的增加，自由系数的数量也随之增长（由于变量 x 之间的交换对称性，并非所有系数都是独立的），其数量大致与 D^3 成正比。在实际问题中，如果要捕捉数据中的复杂依赖关系，可能需要使用更高阶的多项式。对于一个 M 阶多项式，其系数数量的增长大致像 D^M。虽然这是一种幂律增长，而不是指数增长，但它仍然表明该方法会很快变得难以处理，并且在实际应用中用途有限。

我们的几何直觉是在三维空间的生活经验中形成的，但当考虑高维空间时，这种直觉往往会失效。举一个简单的例子：考虑一个在 D 维空间中的半径为 $r = 1$ 的球体，并问球体体积中有多少部分落在半径区间 $[r = 1 - \epsilon, r = 1]$ 内。我们可以通过注意到：在 D 维空间中，半径为 r 的球体体积随 r^D 缩放，来计算这个比例。因此可以写作：

$$V_D(r) = K_D r^D \tag{1.75}$$

其中常数 K_D 只依赖于维度 D。因此，所需的体积分数为

$$\frac{V_D(1) - V_D(1-\epsilon)}{V_D(1)} = 1 - (1-\epsilon)^D \tag{1.76}$$

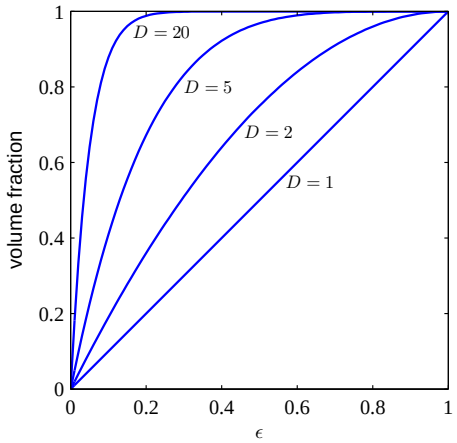

图 1.22 球体体积在半径范围 $r = 1 - \epsilon$ 到 $r = 1$ 内所占比例随空间维度 D 变化的曲线图。

这一结果在图 1.22 中展示为不同维度 D 下的函数曲线。可以看到，当 D 很大时，即使 ϵ 很小，该分数也会趋近于 1。因此，在高维空间中，球体的大部分体积实际上都集中在靠近表面的一个薄壳层中！

再举一个与模式识别直接相关的例子：考虑高维空间中高斯分布的行为。若我们将坐标从笛卡尔坐标系转换为极坐标系，并将方向变量积分掉，就可以得到密度 $p(r)$ 的

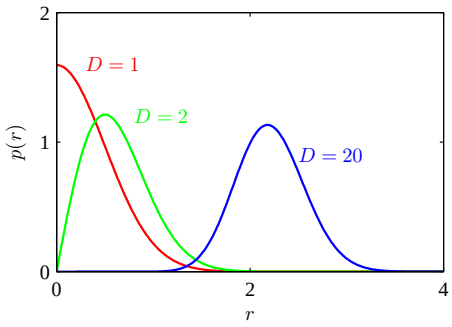

图 1.23 高斯分布关于半径 r 的概率密度随维度 D 变化的曲线图。在高维空间中，高斯分布的大部分概率质量集中在某一特定半径处的薄壳内。

表达式，它是关于到原点的半径 r 的函数。于是，$p(r)\,\delta r$ 表示位于半径 r、厚度为 δr 的薄壳中的概率质量。该分布在不同维度 D 下的曲线如图 1.23 所示，可以看到，当 D 很大时，高斯分布的概率质量同样集中在一个薄壳层中。

在多维空间中可能出现的严重困难有时被称为维度灾难 (Bellman, 1961)。在本书中，我们将广泛使用一维或二维输入空间的示例，因为这样能够更直观地用图形方式展示各种技术。然而，需要提醒读者的是，并非所有在低维空间中形成的直觉都可以推广到高维空间。

尽管维度灾难确实给模式识别应用带来了重要挑战，但它并不会阻止我们找到在高维空间中仍然有效的技术。原因主要有两个方面：首先，真实数据往往局限在空间中的某个较低有效维度的区域内，特别是目标变量发生重要变化的方向通常也仅限于该区域。其次，真实数据通常表现出某种光滑性特征（至少在局部上如此），也就是说输入变量的小变化通常只会引起目标变量的小变化。因此我们可以利用类似局部插值的方法，使得能够对新的输入值预测其对应的目标值。成功的模式识别技术往往利用了上述一种或两种性质。举个例子，在制造业的一个应用中，相机会对传送带上的相同平面物体拍摄图像，其目标是判断物体的朝向。每幅图像都对应高维空间中的一个点，这个空间的维度由像素数量决定。由于物体在图像中可能出现不同的位置和不同的方向，因此图像之间的变化实际上只具有三个自由度。于是，这组图像分布在高维空间中一个三维流形上。由于物体的位置或方向与像素强度之间存在复杂的关系，这个流形会是高度非线性的。如果我们的目标是学习一个模型，使得它能接收一幅输入图像并输出物体的朝向，而不受物体在图像中位置的影响，那么在这个流形中真正重要的自由度实际上只有一个。

1.5 决策理论

我们在第 1.2 节中已经看到，概率论为刻画和处理不确定性提供了一个一致的数学框架。接下来我们将讨论决策理论。当它与概率论结合时，就能够使我们在涉及不确定性的情形下（例如模式识别中遇到的情况）做出最优决策。

假设我们有一个输入向量 $\mathbf{x}$，以及与之对应的目标变量向量 $\mathbf{t}$，我们的目标是在给定新的 $\mathbf{x}$ 时预测 $\mathbf{t}$。对于回归问题，$\mathbf{t}$ 由连续变量组成；而对于分类问题，$\mathbf{t}$ 表示类别标

签。联合概率分布 $p(\mathbf{x}, \mathbf{t})$ 对这些变量相关的不确定性给出了完整的刻画。从训练数据集中确定 $p(\mathbf{x}, \mathbf{t})$ 是推断的一个例子，而这通常是一个非常困难的问题，也是本书的重要研究主题。然而，在实际应用中，我们往往必须对 $\mathbf{t}$ 的取值给出一个具体预测，或者更一般地，根据我们对 $\mathbf{t}$ 可能取值的理解采取某种具体行动。这一部分内容正是决策理论所研究的。

例如，考虑一个医学诊断问题：我们获得了一位病人的 X 光图像，希望判断该病人是否患有癌症。在这种情况下，输入向量 $\mathbf{x}$ 就是图像中的像素强度集合，而输出变量 t 表示癌症的有无。若病人患癌，则记为类别 C_1；若无癌症，则记为类别 C_2。我们可以令 t 为一个二值变量，例如 $t = 0$ 对应类别 C_1，而 $t = 1$ 对应类别 C_2。稍后我们会看到，这种标签取值的设定对于概率模型来说尤其方便。一般的推断问题是要确定联合分布 $p(\mathbf{x}, C_k)$，或等价地 $p(\mathbf{x}, t)$，它为我们提供了关于问题最完整的概率描述。尽管这个量非常有用且信息丰富，但最终我们必须决定是否对病人进行治疗，并且我们希望这一选择在某种合理意义下是最优的 (Duda and Hart, 1973)。这就是决策步骤，而决策理论的任务就是在给定合适的概率信息时告诉我们如何做出最优决策。我们将看到，一旦推断问题得到解决，决策阶段通常非常简单，甚至是微不足道的。

这里我们将介绍决策理论的核心思想，以便支撑本书其余部分的内容。更多的背景知识以及更详细的论述可以参考 Berger (1985) 和 Bather (2000)。

在进行更详细的分析之前，我们先非正式地思考一下概率在决策中可能发挥的作用。当我们获得一个新病人的 X 光图像 $\mathbf{x}$ 时，我们的目标是决定该图像应归入两个类别中的哪一个。我们关心的是在给定图像条件下两个类别的概率，即 $p(C_k|\mathbf{x})$。利用贝叶斯定理，这些概率可以写成如下形式：

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)\, p(C_k)}{p(\mathbf{x})} \tag{1.77}$$

需要注意的是，贝叶斯定理中出现的任意量都可以通过对联合分布 $p(\mathbf{x}, C_k)$ 进行适当的边缘化或条件化来得到。我们可以将 $p(C_k)$ 解释为类别 C_k 的先验概率而将 $p(C_k|\mathbf{x})$ 解释为对应的后验概率。例如，$p(C_1)$ 表示一个人在进行 X 光检查之前患癌的概率，而 $p(C_1|\mathbf{x})$ 则是在结合 X 光图像信息后，通过贝叶斯定理修正得到的对应概率。如果我们的目标是最小化将 $\mathbf{x}$ 归入错误类别的概率，那么直觉上我们应选择后验概率较大的那个类别。接下来我们将证明这一直觉确实正确，同时还会讨论更一般的决策准则。

1.5.1 最小化误分类率

假设我们的目标仅仅是尽可能减少误分类的数量。我们需要一个规则，将每个输入 $\mathbf{x}$ 分配到某个类别中。这样的规则会把输入空间划分为若干区域 R_k，称为决策区域，每个类别 C_k 对应一个决策区域，使得所有落在 R_k 内的点都被判为 C_k。决策区域之间的边界称为决策边界或决策曲面。需要注意的是，每个决策区域不必是连通的，它可能由若干不相连的子区域组成。我们将在后续章节中看到决策边界与决策区域的具体示例。

为了找到最优的决策规则，先考虑只有两个类别的情况，例如癌症诊断问题。当一

个属于类别 C_1 的输入向量被判为 C_2，或者反之，则会发生错误。该错误发生的概率为：

$$\begin{aligned} p(\text{mistake}) &= p(\mathrm{x}, \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathrm{x}, \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathrm{x}, \mathcal{C}_2)\,\mathrm{dx} \;+\; \int_{\mathcal{R}_2} p(\mathrm{x}, \mathcal{C}_1)\,\mathrm{dx} \end{aligned} \tag{1.78}$$

我们可以自由选择决策规则，将每个点 x 分配给两个类别中的一个。显然，为了最小化 $p(\text{mistake})$，我们应当使每个 x 被分配给在 (1.78) 积分中对应项较小的那个类别。也就是说，如果对于某个 **x**，有 $p(\mathrm{x}, \mathcal{C}_1) > p(\mathrm{x}, \mathcal{C}_2)$，那么就应当把该 x 判为类别 C_1。根据概率的乘法法则 (product rule)，我们有 $p(\mathrm{x}, \mathcal{C}_k) = p(\mathcal{C}_k|\mathrm{x})\, p(\mathrm{x})$。由于 $p(\mathrm{x})$ 在两个类别中是相同的，因此结果可以重新表述为：若希望误分类概率最小化，则应当把每个 x 判为其后验概率 $p(\mathcal{C}_k|\mathrm{x})$ 最大的类别。这一结论在两个类别、单个输入变量 x 的情形下如图 1.24 所示。

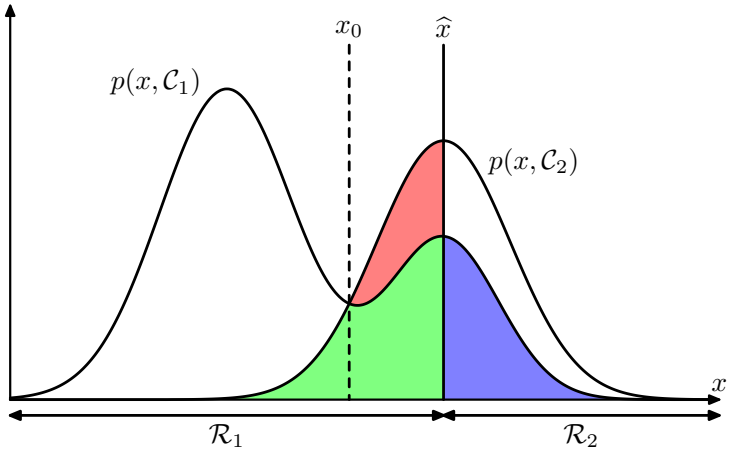

图 1.24 关于两个类别的联合概率 $p(x, \mathcal{C}_k)$ 随 x 变化的示意图，同时给出了决策边界 $x = \hat{x}$。当 $x \geq \hat{x}$ 时，被分类为类别 $\mathcal{C}_2$，因此属于决策区域 $\mathcal{R}_2$；而当 $x < \hat{x}$ 时，被分类为类别 $\mathcal{C}_1$，属于决策区域 R_1。错误来源于蓝色、绿色和红色区域：在 $x < \hat{x}$ 时，错误来自类别 C_2 的点被误分类为 $\mathcal{C}_1$（由红色和绿色区域表示）；而在 $x \geq \hat{x}$ 时，错误来自类别 $\mathcal{C}_1$ 的点被误分类为 $\mathcal{C}_2$（由蓝色区域表示）。当我们改变决策边界 $\hat{x}$ 的位置时，蓝色和绿色区域的总面积保持不变，而红色区域的大小发生变化。最优的 x_b 选择是 $p(x, \mathcal{C}_1)$ 与 $p(x, \mathcal{C}_2)$ 曲线的交点，即 $\hat{x} = x_0$，因为此时红色区域消失。这等价于最小误分类率的决策规则，即将每个 x 分配给后验概率 $p(\mathcal{C}_k|x)$ 较大的类别。

对于更一般的 K 个类别情形，处理起来稍微简单的方法是最大化判别正确的概率，其表达式为：

$$\begin{aligned} p(\text{correct}) &= \sum_{k=1}^{K} p(\mathrm{x} \in \mathcal{R}_k, \mathcal{C}_k) \\ &= \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathrm{x}, \mathcal{C}_k)\,\mathrm{dx} \end{aligned} \tag{1.79}$$

当决策区域 $\mathcal{R}_k$ 的选择使得每个 x 都被分配给满足 $p(\mathrm{x}, \mathcal{C}_k)$ 最大的那个类别时，上式达到最大值。再次利用概率的乘法法则：$p(\mathrm{x}, \mathcal{C}_k) = p(\mathcal{C}_k|\mathrm{x})\, p(\mathrm{x})$，并注意到因子 $p(\mathrm{x})$ 对所有项都是相同的，于是可以得到：每个 x 应该被判为其后验概率 $p(\mathcal{C}_k|\mathbf{x})$ 最大的类

别。

1.5.2 最小化期望损失

在许多应用中，我们的目标会比单纯最小化误分类数量更为复杂。再考虑一次医学诊断问题：如果一个没有癌症的病人被错误地诊断为患癌，后果可能只是带来一些心理负担以及进一步检查的需要；而如果一个真正患癌的病人被诊断为健康，那么后果可能是由于缺乏治疗而导致的过早死亡。因此，这两类错误所带来的后果可能截然不同。显然，更好的做法是尽量减少第二类错误，即使这意味着要容忍更多第一类错误。

我们可以通过引入损失函数（也称为代价函数）来形式化地处理这类问题。损失函数给出了在采取某个具体决策或行动时所造成的总体损失度量。我们的目标就是最小化总损失。需要注意的是，有些作者使用效用函数，并以最大化效用为目标。如果我们把效用定义为损失的相反数，这两种方法是等价的。在本书中，我们采用损失函数的约定。设对于某个新的 x，其真实类别为 $\mathcal{C}_k$，而我们将其判为类别 $\mathcal{C}_j$（其中 j 可以等于或不等于 k）。这样我们会产生某种程度的损失，记为 L_{kj}，它可以看作是损失矩阵的第 (k, j) 元素。例如，在癌症诊断问题中，我们可能会有如下形式的损失矩阵（如图 1.25 所示）。如果决策正确，则损失为 0；如果一个健康患者被诊断为癌症，则损失为 1；如果一个患癌患者被诊断为健康，则损失为 1000。

$$\begin{array}{c} \\ \text{cancer} \\ \text{noemal} \end{array} \begin{array}{c} \begin{array}{cc} \text{} & \text{} \end{array} \\ \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} \end{array}$$

图 1.25 癌症治疗问题中的一个损失矩阵示例，其元素为 L_{kj}。矩阵的行对应真实类别，而列对应于我们的决策准则所做的类别判定。

最优解应当是使损失函数最小化的那个决策。然而，损失函数依赖于真实类别，而真实类别是未知的。对于给定的输入向量 x，我们对真实类别的不确定性由联合分布 $p(\mathrm{x}, C_k)$ 表达。因此，我们寻求最小化的是平均损失 (average loss)，其计算是相对于该分布的期望，形式为：

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p\left(\mathrm{x}, \mathcal{C}_k\right) \mathrm{dx} \tag{1.80}$$

对于每个 x，选择区域 $\mathcal{R}\_j$ 的分配可以独立进行。我们的目标意味着，为了最小化期望损失 (1.80)，对于每个 x 应该最小化 $\sum_k L_{kj}\, p(\mathrm{x}, \mathcal{C}_k)$。如前所述，我们可以利用乘法规则 $p(\mathrm{x}, \mathcal{C}_k) = p(\mathcal{C}_k \mid \mathrm{x}), p(\mathrm{x})$ 消去公共因子 $p(\mathrm{x})$。因此，能够最小化期望损失的判别规则是：将每个新的 x 分配到使下式最小的类别 j，其最优类别就是使上述量达到最小值的那个类别。显然，一旦我们知道了后验类别概率 $p(\mathcal{C}_k|\mathrm{x})$，这一决策就变得非常简单。

$$\sum_k L_{kj}\, p(\mathcal{C}_k \mid \mathrm{x}) \tag{1.81}$$

1.5.3 拒绝选项

我们已经看到，分类错误主要发生在输入空间的某些区域，这些区域中最大的后验概率 $p(\mathcal{C}_k|\mathbf{x})$ 显著小于 1，或者等价地说，此时各联合分布 $p(\mathbf{x},\mathcal{C}_k)$ 的数值相差不大。这些区域正是类别归属存在较大不确定性的地方。在一些应用中，面对这类难以判别的情况，避免强行做出决策反而更合适，这样可以在真正进行分类的样本上获得更低的错误率。这被称为拒绝选项。例如，在医学诊断的例子中，自动系统可以用于判别那些类别几乎没有疑问的 X 光图像，而将更模糊的病例留给人类专家判断。实现这一点的方法是引入一个阈值 θ，当最大的后验概率 $p(\mathcal{C}_k|\mathbf{x}) \leq \theta$ 时，就拒绝对输入 $\mathbf{x}$ 进行分类。图 1.26 展示了两个类别、单个连续输入变量 x 的情况。需要注意的是：当 $\theta = 1$ 时，所有样本都会被拒绝；若类别数为 K，则当 $\theta < 1/K$ 时，没有样本会被拒绝。因此，被拒绝样本的比例是由阈值 θ 控制的。

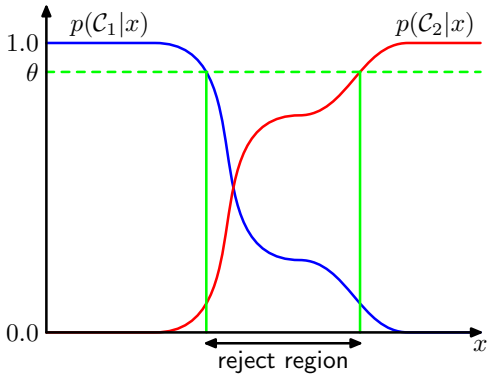

图 1.26 拒绝选项的示意图：对于输入 x，如果两类后验概率中较大的一个小于或等于某个阈值 θ，则该输入将被拒绝分类。

我们可以很容易地将拒绝准则扩展到最小化期望损失，当给定一个损失矩阵时，可以考虑在作出拒绝决策时所带来的损失。

1.5.4 推断与决策

我们将分类问题分解为两个独立的阶段：推断阶段，在此阶段我们利用训练数据学习一个关于 $p(\mathcal{C}_k|\mathbf{x})$ 的模型；以及随后的决策阶段，在此阶段我们使用这些后验概率来做出最优的类别分配。另一种可能性是将这两个问题一起解决，直接学习一个将输入 $\mathbf{x}$ 映射为决策的函数。这样的函数称为判别函数。

事实上，我们可以识别出三种解决决策问题的不同方法，这些方法都已经在实际应用中得到使用。按照复杂性递减的顺序，它们是：

(a) 首先解决推断问题，即分别确定每个类别 $\mathcal{C}_k$ 的类条件密度 $p(\mathbf{x}|\mathcal{C}_k)$。同时单独推断先验类别概率 $p(\mathcal{C}_k)$。然后使用贝叶斯定理，其形式为

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} \tag{1.82}$$

以此来求解后验类别概率 $p(\mathcal{C}_k|\mathbf{x})$。像往常一样，贝叶斯定理中的分母可以通过

分子中出现的量来表示，因为

$$p(\mathrm{x}) = \sum_j p(\mathrm{x}|\mathcal{C}_j)p(\mathcal{C}_j) \tag{1.83}$$

等价地，我们也可以直接对联合分布 $p(\mathrm{x}, \mathcal{C}_k)$ 建模，然后通过归一化得到后验概率。得到后验概率之后，我们利用决策理论来确定每个新输入 x 的类别归属。那些显式或隐式地同时对输入和输出分布建模的方法被称为生成模型，因为通过从它们中采样，可以在输入空间中生成合成的数据点。

(b) 首先解决推断问题，即确定后验类别概率 $p(\mathcal{C}_k|\mathrm{x})$，然后再利用决策理论将每个新的 x 分配到某个类别中。直接对后验概率建模的方法被称为判别模型。

(c) 找到一个函数 $f(\mathrm{x})$，称为判别函数，它将每个输入 x 直接映射到一个类别标签上。举例来说，在二分类问题中，函数 $f(\cdot)$ 可以是一个二值函数，例如 $f = 0$ 表示类别 $\mathcal{C}_1$，而 $f = 1$ 表示类别 $\mathcal{C}_2$。在这种情况下，概率不起作用。

让我们考虑这三种方法的相对优缺点。方法 (a) 是最具挑战性的，因为它涉及到寻找 x 和 $\mathcal{C}_k$ 的联合分布。对于许多应用来说，x 将具有高维性，因此我们可能需要一个很大的训练集，才能以合理的精度确定类条件密度 $p(\mathrm{x}|\mathcal{C}_k)$。注意，类别先验概率 $p(\mathcal{C}_k)$ 通常可以简单地通过训练集中各类别数据点的比例来估计。然而，方法 (a) 的一个优点是它还能通过公式 (1.83) 确定数据的边缘密度 $p(\mathrm{x})$。这对于检测那些在模型下概率很低、因而预测精度可能较低的新数据点是有用的，这被称为异常点检测或新奇检测（Bishop, 1994; Tarassenko, 1995）。

然而，如果我们只希望做分类决策，那么在实际上只需要后验概率 $p(\mathcal{C}_k|\mathrm{x})$ 的情况下去寻找联合分布 $p(\mathrm{x}, \mathcal{C}_k)$，就可能是对计算资源的浪费，并且对数据的要求也过高。而通过方法 (b) 就可以直接获得所需的后验概率。事实上，类条件密度中可能包含大量结构，但这些结构对后验概率几乎没有影响，如图 1.27 所示。对于生成方法和判别方法在机器学习中的相对优缺点，以及如何将它们结合起来，已经引起了广泛的研究兴趣（Jebara, 2004; Lasserre 等, 2006）。

一个更简单的方法是 (c)，在该方法中我们利用训练数据找到一个判别函数 $f(\mathrm{x})$，它将每个 x 直接映射到一个类别标签上，从而把推断阶段和决策阶段合并为一个单一的学习问题。在图 1.27 的例子中，这对应于找到由绿色竖线表示的 x 值，因为这就是给出最小误分类概率的决策边界。

然而，在选择方法 (c) 时，我们将无法再获得后验概率 $p(\mathcal{C}_k|\mathrm{x})$。即使最终是用它们来做决策，仍然有许多重要的理由需要计算后验概率，其中包括：

最小化风险。 考虑这样一个问题：损失矩阵的元素会不时地被修订（例如在金融应用中可能会发生这种情况）。如果我们知道后验概率，就可以通过适当修改公式 (1.81)，轻而易举地修订最小风险决策准则。而如果我们只有一个判别函数，那么损失矩阵的任何变化都将要求我们回到训练数据，重新解决分类问题。

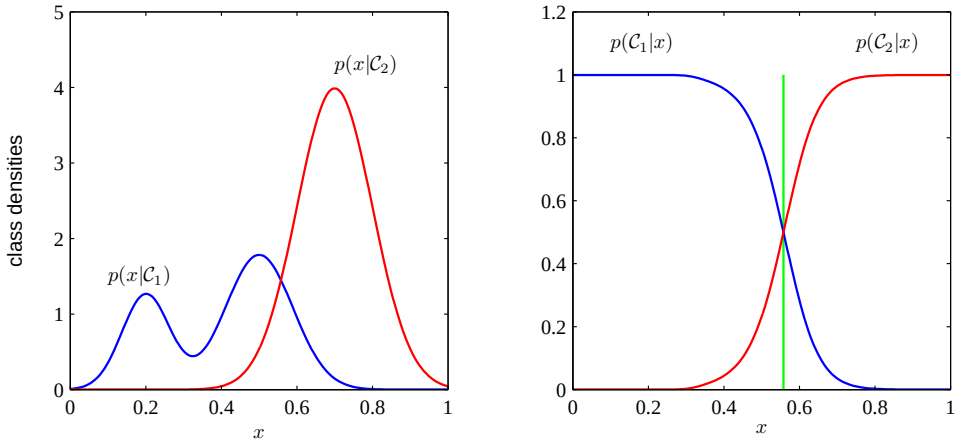

图 1.27 两个类别在单一输入变量 x 下的类条件密度示例（左图），以及对应的后验概率（右图）。需要注意的是，左图中蓝色表示的类条件密度 $p(\mathrm{x}|\mathcal{C}_1)$ 的左侧峰值对后验概率没有影响。右图中的竖直绿色线表示在 x 上给出最小误分类率的决策边界。

拒绝选项。 后验概率使我们能够确定一个拒绝准则，该准则可以在给定比例的被拒绝数据点下，最小化误分类率，或者更一般地，最小化期望损失。

补偿类别先验。 再来考虑我们的医学 X 光问题。假设我们从普通人群中收集了大量 X 光图像作为训练数据，用于构建一个自动化筛查系统。由于癌症在普通人群中很少见，我们可能会发现，例如每 1,000 个样本中只有 1 个对应于癌症。如果我们用这样的数据集去训练一个自适应模型，就可能会遇到严重的问题，因为癌症类别的比例太小。例如，一个把所有点都判为正常类别的分类器就已经可以达到 99.9% 的准确率，而很难避免这种平凡的解。此外，即使是一个很大的数据集，也只会包含极少数对应癌症的 X 光图像，因此学习算法无法接触到这类图像的广泛样例，从而不太可能有良好的泛化能力。一个解决方案是构造一个平衡的数据集，在其中为每个类别选取相等数量的样本，这样可以帮助我们得到一个更准确的模型。然而，在这种情况下，我们必须对训练数据的修改进行补偿。假设我们使用了这样的平衡数据集并得到了后验概率的模型。根据贝叶斯定理 (1.82)，我们知道后验概率与先验概率成正比，而先验概率可以解释为每个类别中的样本比例。因此，我们只需将从人工平衡数据集中得到的后验概率，先除以该数据集中的类别比例，再乘以目标人群中的类别比例。最后，还需要进行归一化，以确保新的后验概率之和为 1。需要注意的是，如果我们直接学习的是判别函数，而不是后验概率，那么这一补偿过程就无法进行。

模型结合。 对于复杂的应用，我们可能希望将问题分解为若干较小的子问题，每个子问题由一个单独的模块来解决。例如，在假设的医学诊断问题中，我们可能同时拥有来自血液检测和 X 光图像的信息。与其把这些异质信息合并到一个庞大的输入空间中，不如分别构建一个系统来解释 X 光图像，另一个系统来解释血液数据。只要这两个模型都能给出类别的后验概率，我们就可以利用概率论的规则系统地

结合它们的输出。一个简单的方法是假设在每个类别下，X 光图像的输入分布（记作 $\mathrm{x_I}$）与血液数据的输入分布（记作 $\mathrm{x_B}$）相互独立，于是有

$$p(\mathrm{x_I}, \mathrm{x_B}|\mathcal{C}_k) = p(\mathrm{x_I}|\mathcal{C}_k)\, p(\mathrm{x_B}|\mathcal{C}_k) \tag{1.84}$$

这是条件独立性质的一个例子，因为独立性在分布以类别 $\mathcal{C}_k$ 为条件时成立。于是，在同时给定 X 光数据和血液数据的情况下，后验概率可以表示为

$$\begin{aligned} p(\mathcal{C}_k|\mathbf{x}_\mathrm{I}, \mathbf{x}_\mathrm{B}) &\propto p(\mathbf{x}_\mathrm{I}, \mathbf{x}_\mathrm{B}|\mathcal{C}_k)p(\mathcal{C}_k) \\ &\propto p(\mathbf{x}_\mathrm{I}|\mathcal{C}_k)p(\mathbf{x}_\mathrm{B}|\mathcal{C}_k)p(\mathcal{C}_k) \\ &\propto \frac{p(\mathcal{C}_k|\mathbf{x}_\mathrm{I})p(\mathcal{C}_k|\mathbf{x}_\mathrm{B})}{p(\mathcal{C}_k)} \end{aligned} \tag{1.85}$$

因此，我们需要类别先验概率 $p(\mathcal{C}_k)$，它可以很容易地通过各类别数据点的比例来估计。接着，我们需要对得到的后验概率进行归一化，使其和为 1。上述特定的条件独立假设 (1.84) 是朴素贝叶斯模型（naive Bayes model）的一个例子。需要注意的是，在该模型下，联合边缘分布 $p(\mathrm{x_I}, \mathrm{x_B})$ 通常并不会因式分解。我们将在后续章节中看到如何构建不依赖于条件独立假设 (1.84) 的数据结合模型。

1.5.5 回归的损失函数

到目前为止，我们在分类问题的背景下讨论了决策理论。现在我们转向回归问题，例如之前讨论过的曲线拟合问题。在回归中，决策阶段就是为每个输入 x 选择一个关于 t 的具体估计 $y(\mathrm{x})$。假设这样做会带来一个损失 $L(t, y(\mathrm{x}))$。那么平均（期望）损失可以表示为

$$\mathbb{E}[L] = \int\int L(t, y(\mathrm{x}))\, p(\mathrm{x}, t)\, d\mathrm{x}\, dt \tag{1.86}$$

在回归问题中，一个常见的损失函数选择是平方损失（squared loss），其形式为 $L(t, y(x)) = \{y(x) - t\}^2$ 。在这种情况下，期望损失可以写作

$$\mathbb{E}[L] = \int\int \{y(x) - t\}^2\, p(x, t)\, dx\, dt \tag{1.87}$$

我们的目标是选择 $y(\mathbf{x})$ 以最小化 $\mathbb{E}[L]$。如果我们假设 $y(\mathbf{x})$ 是完全灵活的函数，我们可以形式化地使用变分法得到

$$\frac{\delta \mathbb{E}[L]}{\delta y(\mathbf{x})} = 2\int \{y(\mathbf{x}) - t\} p(\mathbf{x}, t) \mathrm{d}t = 0 \tag{1.88}$$

解出 $y(x)$，并结合概率的求和法则和乘法法则，我们得到这是在给定 x 条件下 t 的条件均值，被称为回归函数（regression function）。这一结果在图 1.28 中有所展示。它可以很容易地扩展到多个目标变量的情形，此时目标变量由向量 $\mathbf{t}$ 表示，而最优解是条件均值 $y(x) = \mathbb{E}_\mathbf{t}[\mathbf{t} \mid x]$。

$$y(\mathbf{x}) = \frac{\int t p(\mathbf{x}, t) \mathrm{d}t}{p(\mathbf{x})} = \int t p(t \mid \mathbf{x}) \mathrm{d}t = \mathbb{E}_t[t \mid \mathbf{x}] \tag{1.89}$$

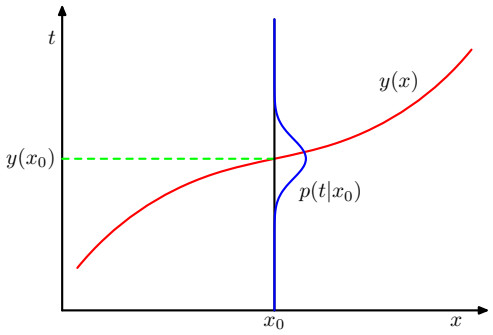

图 1.28 使期望平方损失最小的回归函数 $y(x)$ 由条件分布 $p(t|x)$ 的均值给出。

我们也可以用稍微不同的方法推导这一结果，这也有助于理解回归问题的本质。既然我们已知最优解是条件期望，就可以如下展开平方项：

$$\begin{aligned}\{y(\mathbf{x}) - t\}^2 &= \{y(\mathbf{x}) - \mathbb{E}[t \mid \mathbf{x}] + \mathbb{E}[t \mid \mathbf{x}] - t\}^2 \\ &= \{y(\mathbf{x}) - \mathbb{E}[t \mid \mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t \mid \mathbf{x}]\}\{\mathbb{E}[t \mid \mathbf{x}] - t\} + \{\mathbb{E}[t \mid \mathbf{x}] - t\}^2\end{aligned}$$

为使记号更简洁，我们以 $\mathbb{E}[t \mid \mathbf{x}]$ 表示 $\mathbb{E}_t[t \mid \mathbf{x}]$。将其代回损失函数并对 t 积分，可见交叉项消失，从而得到损失函数的形式

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t \mid \mathbf{x}]\}^2 p(\mathbf{x}) \mathrm{d}\mathbf{x} + \int \{\mathbb{E}[t \mid \mathbf{x}] - t\}^2 p(\mathbf{x}) \mathrm{d}\mathbf{x} \tag{1.90}$$

我们要求解的函数 $y(\mathbf{x})$ 只出现在第一项中。当 $y(\mathbf{x}) = \mathbb{E}[t \mid \mathbf{x}]$ 时，这一项将被最小化并消失。这正是我们之前推导出的结果，表明最优的最小二乘预测器就是条件均值。第二项是 t 的分布的方差，对 $\mathbf{x}$ 取平均。它表示目标数据的内在变异性，可以视作噪声。由于它与 $y(\mathbf{x})$ 无关，因此它对应于损失函数不可约的最小值。

与分类问题类似，我们可以先确定相应的概率，再利用这些概率做出最优决策；或者我们也可以直接构建模型来做决策。实际上，我们可以识别出三种解决回归问题的不同方法，按照复杂性递减的顺序为：

(a) 首先解决推断问题，即确定联合密度 $p(\mathbf{x}, t)$。随后通过归一化得到条件密度 $p(t \mid \mathbf{x})$，最后对其进行边缘化以得到由式 (1.89) 给出的条件均值。

(b) 首先解决推断问题，即确定条件密度 $p(t \mid \mathbf{x})$，随后对其进行边缘化以得到由式 (1.89) 给出的条件均值。

(c) 直接从训练数据中求得回归函数 $y(\mathbf{x})$。

这三种方法的相对优缺点与前面分类问题中的情况类似。

平方损失并不是回归中唯一可能的损失函数选择。实际上，在某些情况下，平方损失可能会导致非常糟糕的结果，这时我们需要更复杂的方法。一个重要的例子是条件分布 $p(t \mid \mathbf{x})$ 为多峰分布的情形，这在求解逆问题时经常出现。这里我们简要讨论平方损失的一种简单推广形式，称为 Minkowski 损失，其期望形式为

$$\mathbb{E}[L_q] = \iint |\, y(\mathbf{x}) - t \,|^q \, p(\mathbf{x}, t) \, d\mathbf{x} \, dt, \tag{1.91}$$

当 $q = 2$ 时，这个形式就退化为期望平方损失。图 1.29 中绘制了不同 q 值下函数 $|y-t|^q$ 随 $y-t$ 的变化曲线。期望损失 $\mathbb{E}[L_q]$ 的最小值在以下情形下分别由不同的条件统计量给出：当 $q = 2$ 时是条件均值；当 $q = 1$ 时是条件中位数；而当 $q \to 0$ 时则是条件众数。

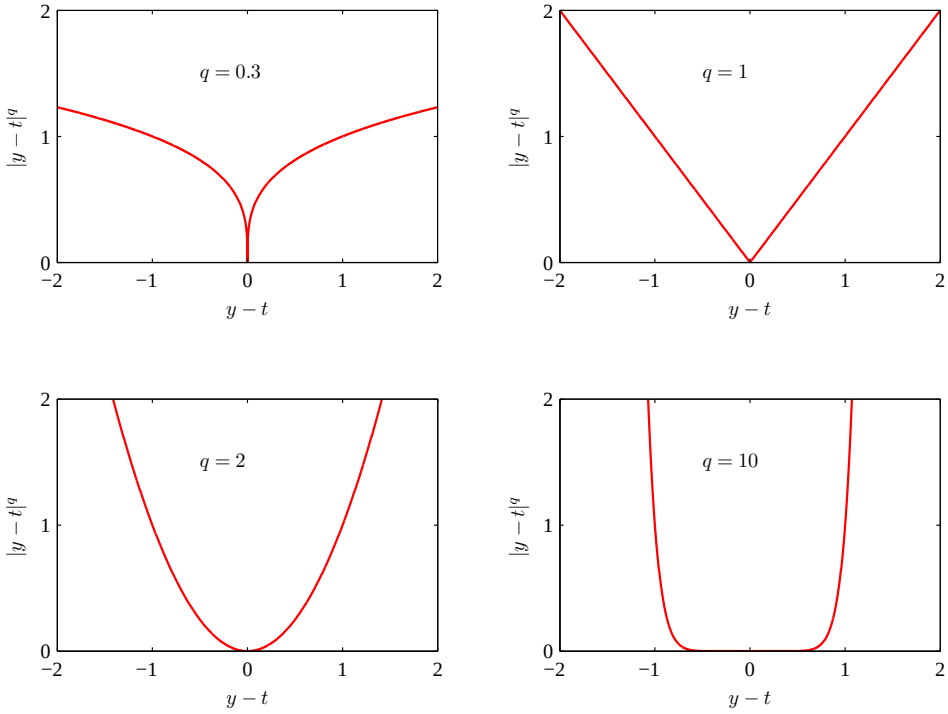

图 1.29 不同 q 取值下量 $L_q = |y-t|^q$ 的曲线图。

1.6 信息论

在本章中，我们讨论了概率论和决策理论中的一系列概念，它们将构成本书后续内容的重要基础。最后，我们将引入一些来自信息论（information theory）领域的附加概念，这些概念同样将在模式识别与机器学习技术的发展中发挥重要作用。这里我们依旧只关注关键概念，更多的细节性讨论可参考其他文献（Viterbi and Omura, 1979; Cover and Thomas, 1991; MacKay, 2003）。

我们先考虑一个离散随机变量 x，并提出问题：当我们观测到该变量的一个具体取值时，接收到多少信息？信息量可以被理解为在得知 x 的取值后产生的“惊讶程度”。如果我们被告知一个极不可能发生的事件已经发生，那么我们接收到的信息量要比得知一个极有可能发生的事件时更多；而如果事件本来就是必然发生的，那么我们将不会获得任何信息。因此，我们对信息量的度量应依赖于概率分布 $p(x)$，并且我们寻找一个量 $h(x)$，它是概率 $p(x)$ 的单调函数，用来表达信息内容。$h(\cdot)$ 的形式可以通过以下思路得到：如果我们有两个不相关的事件 x 和 y，那么观测到它们的总信息量应当等于分别观测到它们的信息量之和，即 $h(x, y) = h(x) + h(y)$。对于不相关事件，它们统计上是独立的，因此 $p(x, y) = p(x)p(y)$。由这两个关系可以容易证明，$h(x)$ 必须是 $p(x)$ 的

对数函数，因此我们有

$$h(x) = -\log_2 p(x) \tag{1.92}$$

其中负号确保了信息量是非负的。需要注意的是，低概率事件 x 对应于较高的信息量。对数的底数选择是任意的，在此我们采用信息论中常见的约定，即使用以 2 为底的对数。这样一来，正如我们很快会看到的，$h(x)$ 的单位就是比特（bits，意为“二进制数字”）。

现在假设一个发送者希望将一个随机变量的取值传递给接收者。在这一过程中传递的信息的平均量，可以通过对式 (1.92) 在分布 $p(x)$ 下取期望来获得，其形式为

$$H[x] = \mathbb{E}[h(x)] = -\sum_x p(x)\log_2 p(x) \tag{1.93}$$

这个重要的量被称为随机变量 x 的熵（entropy）。注意到 $\lim_{p\to 0} p\ln p = 0$，因此，当某个取值满足 $p(x) = 0$ 时，我们约定 $p(x)\ln p(x) = 0$。

到目前为止，我们对信息定义 (1.92) 以及相应的熵 (1.93) 给出了一个比较直观的动机。现在我们来展示这些定义确实具有有用的性质。考虑一个随机变量 x，它有 8 个可能的状态，并且每个状态出现的概率都相等。为了将 x 的取值传递给接收者，我们需要发送一条长度为 3 比特的信息。注意到这个随机变量的熵为

$$H[x] = -\sum_{i=1}^{8}\frac{1}{8}\log_2\frac{1}{8} = -8\cdot\frac{1}{8}\log_2\frac{1}{8} = 3\text{bits}$$

现在考虑一个例子（Cover and Thomas, 1991），随机变量有 8 个可能的状态 $\{a,b,c,d,e,f,g,h\}$，其对应的概率分别为 $\left(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{16},\frac{1}{64},\frac{1}{64},\frac{1}{64},\frac{1}{64}\right)$。在这种情况下，熵为

$$H[x] = -\tfrac{1}{2}\log_2\tfrac{1}{2} - \tfrac{1}{4}\log_2\tfrac{1}{4} - \tfrac{1}{8}\log_2\tfrac{1}{8} - \tfrac{1}{16}\log_2\tfrac{1}{16} - \tfrac{4}{64}\log_2\tfrac{1}{64} = 2\text{bits}$$

我们看到，非均匀分布的熵比均匀分布要小，这一点在稍后讨论熵在“无序”意义下的解释时会更清晰。现在先考虑如何将该变量的状态传递给接收者。像之前一样，我们可以用一个 3 比特的数字来表示。但我们也可以利用分布的非均匀性：对概率较大的事件使用更短的编码，而对概率较小的事件使用更长的编码，从而希望获得更短的平均码长。例如，可以用如下编码方案表示状态 $\{a,b,c,d,e,f,g,h\}$：0, 10, 110, 1110, 111100, 111101, 1
此时所需传输的平均码长为

$$L = \tfrac{1}{2}\cdot 1 + \tfrac{1}{4}\cdot 2 + \tfrac{1}{8}\cdot 3 + \tfrac{1}{16}\cdot 4 + 4\cdot\tfrac{1}{64}\cdot 6 = 2\text{bits}$$

这再次等于该随机变量的熵。需要注意的是，不能使用更短的编码串，因为必须能够将这些编码串的拼接唯一地分解为其组成部分。例如，序列 11001110 可以被唯一解码为状态序列 c, a, d。

熵与最短编码长度之间的这种关系是普遍成立的。无噪声编码定理（noiseless coding theorem，Shannon, 1948）指出：熵是传输一个随机变量状态所需比特数的下界。

从现在开始，我们将在熵的定义中改用自然对数，这样可以更方便地与本书其他部分的思想建立联系。在这种情况下，熵的度量单位是“纳特”（nats），而不是比特（bits），两者仅相差一个 $\ln 2$ 的因子。

我们已经从“指定随机变量状态所需信息的平均量”角度引入了熵的概念。事实上，熵在物理学中起源更早，它最初是在平衡热力学的背景下提出的，后来在统计力学的发展中被赋予了“无序度量”的更深层解释。为了理解熵的这种另一种观点，考虑这样一个问题：有 N 个相同的物体要分配到若干个盒子中，其中第 i 个盒子里有 n_i 个物体。我们来计算不同分配方式的数目。分配第一个物体有 N 种方式，第二个物体有 $(N-1)$ 种方式，依此类推，总共有 $N!$ 种方式将 N 个物体全部分配到盒子中，其中 $N!$ 表示乘积 $N \times (N-1) \times \cdots \times 2 \times 1$。然而，我们并不想区分同一盒子内物体的不同排列。在第 i 个盒子中，存在 $n_i!$ 种方式重新排列这些物体。因此，N 个物体分配到各个盒子的不同方式总数为

$$W = \frac{N!}{\prod_i n_i!} \tag{1.94}$$

这被称为多重度（multiplicity）。熵在这里定义为多重度的对数，并乘以一个适当的常数：

$$\mathrm{H} = \frac{1}{N} \ln W = \frac{1}{N} \ln N! - \frac{1}{N} \sum_i \ln n_i! \tag{1.95}$$

现在我们考虑极限 $N \to \infty$，在此过程中保持各个比例 n_i/N 不变，并应用 Stirling 近似：

$$\ln N! \simeq N \ln N - N \tag{1.96}$$

由此可得

$$\mathrm{H} = -\lim_{N\to\infty} \sum_i \left(\frac{n_i}{N}\right) \ln\left(\frac{n_i}{N}\right) = -\sum_i p_i \ln p_i \tag{1.97}$$

这里我们利用了 $\sum_i n_i = N$。其中 $p_i = \lim_{N\to\infty} \left(\frac{n_i}{N}\right)$ 表示一个物体被分配到第 i 个盒子的概率。在物理学术语中，物体在盒子中的具体排列称为微观态，而通过比例 n_i/N 表示的总体占据数分布称为宏观态。多重度 W 也被称为宏观态的权重。

我们可以把这些盒子理解为离散随机变量 X 的状态 x_i，其中 $p(X = x_i) = p_i$。于是随机变量 X 的熵为

$$\mathrm{H}[p] = -\sum_i p(x_i) \ln p(x_i) \tag{1.98}$$

如果分布 $p(x_i)$ 在少数几个取值附近高度集中，那么其熵就会相对较低；而如果分布在多个取值上比较均匀地展开，那么熵就会更高，如图 1.30 所示。由于 $0 \le p_i \le 1$，熵总是非负的。当某个 $p_i = 1$ 且所有其他 $p_{j\neq i} = 0$ 时，熵取最小值 0。最大熵配置可以通过在概率的归一化约束下最大化 H 来找到。为此，我们使用拉格朗日乘子方法：

$$\tilde{\mathrm{H}} = -\sum_i p(x_i) \ln p(x_i) + \lambda \left(\sum_i (x_i) - 1 \right) \tag{1.99}$$

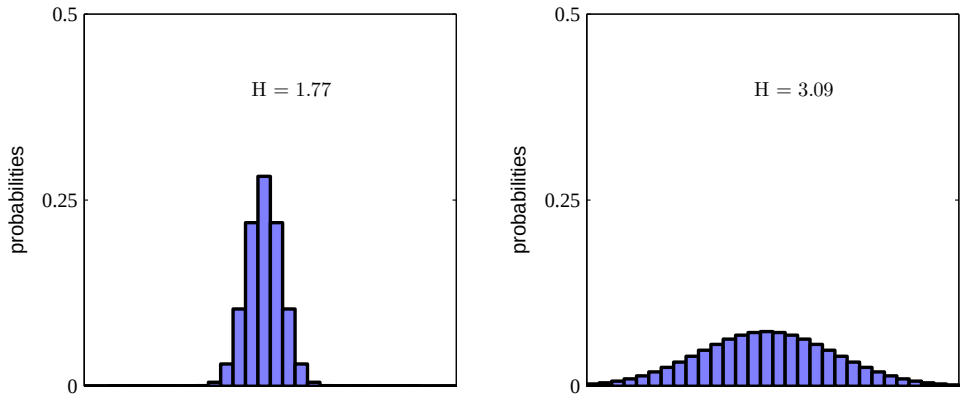

图 1.30 两个在 30 个区间上的概率分布的直方图，用来说明更宽的分布具有更大的熵 H。最大熵出现在均匀分布的情况下，此时 $H = -\ln\left(\frac{1}{30}\right) = 3.40$。

由此我们得到，所有的 $p(x_i)$ 都相等，即 $p(x_i) = \frac{1}{M}$，其中 M 是状态 x_i 的总数。对应的熵为 $\mathrm{H} = \ln M$。这一结果也可以通过 Jensen 不等式（Jensen's inequality，将在后面讨论）推导出来。为了验证该驻点确实是最大值，我们可以计算熵的二阶导数，得到

$$\frac{\partial \tilde{\mathrm{H}}}{\partial p(x_i)\partial p(x_j)} = -I_{ij}\frac{1}{p_i} \tag{1.100}$$

这里的 I_{ij} 是单位矩阵的元素。

我们可以将熵的定义扩展到连续变量 x 上的分布 $p(x)$。方法是：首先把 x 划分为宽度为 Δ 的小区间。然后，在假设 $p(x)$ 连续的条件下，根据平均值定理（Weisstein, 1999），对于每个区间，必定存在一个值 x_i，使得

$$\int_{i\Delta}^{(i+1)\Delta} p(x)\,\mathrm{d}x \;=\; p(x_i)\,\Delta \tag{1.101}$$

现在，我们可以通过量化连续变量 x 来处理：当 x 落入第 i 个区间时，就将其对应到值 x_i。于是观测到值 x_i 的概率为 $p(x_i)\Delta$。这样我们得到一个离散分布，其熵为

$$H_\Delta = -\sum_i p(x_i)\Delta \ln(p(x_i)\Delta) = -\sum_i p(x_i)\Delta \ln p(x_i) - \ln \Delta \tag{1.102}$$

这里我们利用了关系 $\sum_i p(x_i)\Delta = 1$，这由式 (1.101) 可得。现在我们在式 (1.102) 的右边省略第二项 $-\ln\Delta$，然后考虑极限 $\Delta \to 0$。在这个极限下，右边的第一项将趋于积分形式 $\int p(x)\ln p(x)\,dx$，于是有

$$\lim_{\Delta\to 0}\left\{\sum_i p(x_i)\Delta \ln p(x_i)\right\} = -\int p(x)\ln p(x)\mathrm{d}x \tag{1.103}$$

右边的这个量被称为微分熵。我们可以看到，离散形式和连续形式的熵相差一个 $\ln\Delta$ 项，而当 $\Delta \to 0$ 时该项发散。这反映了一个事实：要非常精确地描述一个连续变量，需要大量的比特。

对于定义在多个连续变量上的密度（这些变量统称为向量 $\mathbf{x}$），其微分熵为

$$H[\mathbf{x}] = -\int p(\mathbf{x}) \ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \tag{1.104}$$

在离散分布的情形中，我们看到最大熵配置对应于在所有可能状态上均匀分布的概率。现在我们来考虑连续变量的最大熵配置。为了使这个最大值有明确意义，我们需要在保持归一化约束的同时，对 $p(x)$ 的一阶和二阶矩加以限制。

因此，我们在以下三个约束条件下最大化微分熵：

$$\int p(x) \, \mathrm{d}x = 1 \tag{1.105}$$

$$\int x \, p(x) \, \mathrm{d}x = \mu \tag{1.106}$$

$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x) \, \mathrm{d}x = \sigma^2 \tag{1.107}$$

这个带约束的最大化问题可以通过拉格朗日乘子法来解决，因此我们需要对 $p(x)$ 最大化如下泛函：

$$\begin{aligned} &-\int_{-\infty}^{\infty} p(x) \ln p(x) \mathrm{d}x + \lambda_1 \left(\int_{-\infty}^{\infty} p(x) \mathrm{d}x - 1 \right) \\ &+\lambda_2 \left(\int_{-\infty}^{\infty} x p(x) \mathrm{d}x - \mu \right) + \lambda_3 \left(\int_{-\infty}^{\infty} (x-\mu)^2 p(x) \mathrm{d}x - \sigma^2 \right) \end{aligned}$$

使用变分法，令该泛函对 $p(x)$ 的变分导数为零，可得

$$p(x) = \exp\left\{ -1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 \right\} \tag{1.108}$$

将上述结果代回三个约束条件中求解拉格朗日乘子，最终得到的分布为

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) \tag{1.109}$$

因此，能够最大化微分熵的分布就是高斯分布。需要注意的是，在最大化熵的过程中，我们并没有显式地约束分布必须为非负。然而，由于最终得到的分布本身确实是非负的，所以事后可以看出，这样的约束并不是必要的。

如果我们计算高斯分布的微分熵，可以得到

$$H[x] = \frac{1}{2} \{ 1 + \ln(2\pi\sigma^2) \} \tag{1.110}$$

因此我们再次看到，随着分布变得更宽（即方差 σ^2 增大），熵也随之增加。这个结果还表明，微分熵与离散熵不同，它可以为负值，因为当 $\sigma^2 < 1/(2\pi e)$ 时，由式 (1.110) 可得 $H(x) < 0$。

假设我们有一个联合分布 $p(\mathrm{x}, \mathrm{y})$，从中抽取 (x, y) 成对的值。如果 x 的取值已经已知，那么确定相应 y 所需的额外信息量为 $-\ln p(\mathrm{y} \mid \mathrm{x})$。因此，确定 y 所需的平均额外信息量可以写为

$$H[\mathrm{y}|\mathrm{x}] = -\iint p(\mathrm{y}, \mathrm{x}) \, \ln p(\mathrm{y} \mid \mathrm{x}) \, \mathrm{d}\mathrm{x} \, \mathrm{d}\mathrm{y} \tag{1.111}$$

这被称为在给定 x 的条件下的条件熵 y。利用乘法法则可以很容易看出，条件熵满足以下关系：

$$H[\mathrm{x},\mathrm{y}] = H[\mathrm{y}|\mathrm{x}] + H[\mathrm{x}] \tag{1.112}$$

其中 $H[\mathrm{x},\mathrm{y}]$ 是联合分布 $p(\mathrm{x},\mathrm{y})$ 的微分熵，而 $H[\mathrm{x}]$ 是边缘分布 $p(\mathrm{x})$ 的微分熵。由此可见，描述 x 与 y 所需的信息量，等于单独描述 x 所需的信息量，加上在已知 x 的条件下进一步描述 y 所需的额外信息量。

1.6.1 相对熵与互信息

到目前为止，本节我们已经介绍了若干信息论的概念，其中最核心的是熵。现在我们开始把这些思想与模式识别联系起来。考虑某个未知分布 $p(\mathbf{x})$，假设我们用一个近似分布 $q(\mathbf{x})$ 来对其建模。如果我们基于 $q\mathbf{x})$ 构造一个编码方案，用于将 $\mathbf{x}$ 的取值传输给接收者，那么由于使用 $q(\mathbf{x})$ 而不是真实分布 $p(\mathbf{x})$，在指定 $\mathbf{x}$ 的值时（假设选择了最优编码方案）所需的额外平均信息量（以 nats 为单位）为

$$\begin{aligned} KL(p \parallel q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x}) \mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x}) \mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x}) \ln \left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} \mathrm{d}\mathbf{x} \end{aligned} \tag{1.113}$$

这被称为相对熵或 Kullback–Leibler 散度（简称 KL 散度）（Kullback and Leibler, 1951），用于度量分布 $p(\mathbf{x})$ 与 $q(\mathbf{x})$ 之间的差异。需要注意的是，它不是对称的，即 $KL(p\|q) \neq KL(q\|p)$。

我们现在来证明 KL 散度满足 $KL(p\|q) \geq 0$，且当且仅当 $p(\mathbf{x}) = q(\mathbf{x})$ 时取等号。为此，首先引入凸函数的概念。若一个函数 $f(\mathbf{x})$ 具有如下性质，则称其为凸函数：它的任意一条弦都位于函数图像的上方或重合，如图 1.31 所示。区间 $[a, b]$ 内的任意一点 $\mathbf{x}$ 可以写成 $\mathbf{x} = \lambda a + (1-\lambda)b, \quad 0 \leq \lambda \leq 1$。相应的弦上的点为 $\lambda f(a) + (1-\lambda) f(b)$，而函数在该点的取值为 $f\big(\lambda a + (1-\lambda)b\big)$。凸性意味着

$$f\big(\lambda a + (1-\lambda)b\big) \;\leq\; \lambda f(a) + (1-\lambda) f(b) \tag{1.114}$$

这等价于要求函数的二阶导数在其定义域内处处为正。凸函数的例子包括 $x \ln x$（当 $x > 0$ 时）以及 x^2。如果等号仅在 $\lambda = 0$ 和 $\lambda = 1$ 时成立，则称该函数为严格凸。与之相反，如果函数的任意一条弦都位于函数图像的下方或重合，则称该函数为凹函数，并有相应的严格凹的定义。如果函数 $f(x)$ 是凸的，那么 $-f(x)$ 就是凹的。

利用数学归纳法可由式 (1.114) 推出：对于任意非负权重 $\{\lambda_i\}$（满足 $\sum_i \lambda_i = 1$）与任意点集 $\{x_i\}$，凸函数 $f(x)$ 满足

$$f\left(\sum_{i=1}^{M} \lambda_i x_i\right) \leqq \sum_{i=1}^{M} \lambda_i f(x_i) \tag{1.115}$$

其中 $\lambda_i \geq 0$ 且 $\sum_i \lambda_i = 1$，适用于任意点集 $\{x_i\}$。结果 (1.115) 被称为詹森不等式（Jensen’s inequality）。如果我们把 λ_i 解释为离散随机变量 x 取值 $\{x_i\}$ 上的概率分布，

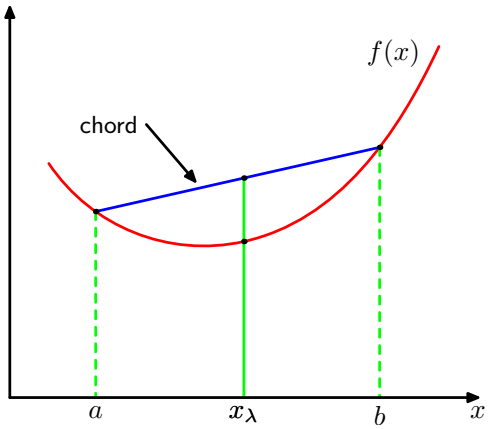

图 1.31 凸函数 $f(x)$ 的定义是：任意弦（蓝色部分）都位于函数曲线（红色部分）的上方或重合。

那么式 (1.115) 可以写作

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)] \tag{1.116}$$

其中 $\mathbb{E}[\cdot]$ 表示期望。对于连续变量，詹森不等式形式为

$$f\left(\int \mathbf{x}p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right) \leq \int f(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{1.117}$$

我们可以将詹森不等式 (1.117) 应用于 KL 散度 (1.113)，得到

$$KL(p||q) = -\int p(\mathbf{x}) \ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} d\mathbf{x} \geq -\ln \int q(\mathbf{x})\mathrm{d}\mathbf{x} = 0 \tag{1.118}$$

这里我们利用了 $-\ln x$ 是凸函数这一事实，以及归一化条件 $\int q(\mathbf{x})\,d\mathbf{x} = 1$。实际上，$-\ln x$ 是严格凸函数，因此仅当 $q(\mathbf{x}) = p(\mathbf{x})$ 对所有 $\mathbf{x}$ 都成立时才会取等号。由此我们可以将 KL 散度理解为度量分布 $p(\mathbf{x})$ 与 $q(\mathbf{x})$ 之间差异性的指标。

我们可以看到，数据压缩与密度估计（即对未知概率分布建模问题）之间有着紧密的联系。因为只有在知道真实分布的情况下，才能实现最有效的压缩。如果使用的分布与真实分布不同，那么编码必然会变得低效，平均而言所需传输的额外信息量（至少）等于这两个分布之间的 KL 散度。

假设数据是由某个未知分布 $p(\mathbf{x})$ 生成的，而我们希望对其进行建模。可以尝试用某个参数化分布 $q(\mathbf{x}|\boldsymbol{\theta})$ 来近似它，其中 $\boldsymbol{\theta}$ 是一组可调参数，例如多元高斯分布。确定 θ 的一种方法是最小化 $p(\mathbf{x})$ 与 $q(\mathbf{x}|\boldsymbol{\theta})$ 之间的 KL 散度。由于我们并不知道 $p(\mathbf{x})$，因此无法直接进行这种最小化。然而，假设我们观测到了一个有限训练集 $\{x_n\}_{n=1}^N$，这些样本由 $p(\mathbf{x})$ 抽取。那么根据式 (1.35)，在 $p(\mathbf{x})$ 下的期望可以通过这些点的有限求和来近似，于是有

$$\mathrm{KL}(p\|q) \approx \sum_{n=1}^{N}\{-\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n)\} \tag{1.119}$$

在式 (1.119) 的右边，第二项与参数 θ 无关，而第一项正是分布 $q(\mathbf{x}|\theta)$ 在训练集上的负对数似然函数。因此，我们看到，最小化这个 KL 散度等价于最大化似然函数。

现在考虑两组变量 $\mathbf{x}$ 和 $\mathbf{y}$ 的联合分布 $p(\mathbf{x},\mathbf{y})$。如果这两组变量相互独立，那么它们的联合分布可以分解为边缘分布的乘积：$p(\mathbf{x},\mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$。如果它们不是独立的，我

们可以通过考察联合分布与边缘分布乘积之间的 KL 散度，来判断它们是否“接近”独立：

$$\begin{aligned} I[\mathbf{x}, \mathbf{y}] &\equiv \mathrm{KL}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x}) p(\mathbf{y})) \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x}) p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) \mathrm{d}\mathbf{x} \mathrm{d}\mathbf{y} \end{aligned} \tag{1.120}$$

这被称为变量 $\mathbf{x}$ 和 y 之间的互信息。根据 KL 散度的性质，我们知道 $I(\mathbf{x}, \mathbf{y}) \geq 0$，且当且仅当 $\mathbf{x}$ 和 $\mathbf{y}$ 独立时取等号。利用概率的求和法则和乘法法则，可以看出互信息与条件熵之间的关系为

$$I[\mathbf{x}, \mathbf{y}] = H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}] = H[\mathbf{y}] - H[\mathbf{y}|\mathbf{x}] \tag{1.121}$$

因此，我们可以把互信息理解为：由于获知了 $\mathbf{y}$ 的取值，从而对 $\mathbf{x}$ 的不确定性减少的量（反之亦然）。从贝叶斯的角度来看，可以将 $p(\mathbf{x})$ 视为关于 $\mathbf{x}$ 的先验分布，而在观测到新的数据 $\mathbf{y}$ 之后，$p(\mathbf{x}|\mathbf{y})$ 就是后验分布。于是，互信息就表示了由于新的观测 $\mathbf{y}$ 而导致的关于 $\mathbf{x}$ 的不确定性的减少。

2 概率分布

在第 1 章中，我们强调了概率论在解决模式识别问题中所起的核心作用。现在我们将探讨一些具体的概率分布及其性质。这些分布本身就具有重要的研究价值，同时它们也可以作为构建更复杂模型的基本单元，并将在全书中被广泛应用。本章介绍的分布还将发挥另一个重要作用，即为我们提供一个契机，在简单模型的背景下讨论一些关键的统计概念（例如贝叶斯推断），从而在后续更复杂的情形中遇到这些概念时能够更好地理解。

本章讨论的分布的一个作用，是在给定有限观测样本 $\mathbf{x}_1, \ldots, \mathbf{x}_N$ 的情况下，用来建模随机变量 $\mathbf{x}$ 的概率分布 $p(\mathbf{x})$。这个问题被称为密度估计。在本章中，我们假设数据点是独立同分布的（i.i.d.）。需要强调的是，密度估计问题在本质上是病态的，因为可能存在无穷多个概率分布都能够生成观测到的有限数据集。实际上，任何在数据点 $\mathbf{x}_1, \ldots, \mathbf{x}_N$ 处取非零值的分布 $p(\mathbf{x})$ 都是潜在候选解。选择合适分布的问题，本质上与模型选择相关，这个问题我们在第 1 章的多项式曲线拟合中已经遇到过，并且它也是模式识别中的核心问题之一。

我们首先考虑离散随机变量的二项分布和多项分布，以及连续随机变量的高斯分布。这些都是参数化分布的具体例子，之所以这样称呼，是因为它们由少量可调参数决定，例如高斯分布中的均值和方差。要将这类模型应用于密度估计问题，我们需要一种方法，在给定观测数据集的情况下确定合适的参数值。在频率学派的处理方法中，我们通过优化某个准则（例如似然函数）来选择参数的具体数值。与此相对，在贝叶斯学派的处理方法中，我们为参数引入先验分布，然后利用贝叶斯定理，在给定观测数据的情况下计算相应的后验分布。

我们将看到，共轭先验在其中起着重要作用。它们的特点是使得后验分布与先验分布具有相同的函数形式，从而大大简化了贝叶斯分析。例如，多项分布参数的共轭先验是狄利克雷分布，而高斯分布的均值参数的共轭先验则是另一个高斯分布。这些分布都是指数族分布的例子。指数族分布具有许多重要性质，我们将在后续部分进行详细讨论。

参数化方法的一个局限在于它假设了分布具有特定的函数形式，而这种形式在某些应用中可能并不合适。另一种方法是非参数密度估计），在这种方法中，分布的形式通常依赖于数据集的规模。这类模型依然包含参数，但这些参数控制的是模型的复杂度，而不是分布的形式。本章最后我们将讨论三种非参数方法，它们分别基于直方图、最近邻以及核函数。

2.1 二元变量

我们先考虑一个二元随机变量 $x \in \{0, 1\}$。例如，x 可以表示一次抛硬币的结果，其中 $x = 1$ 表示“正面”，而 $x = 0$ 表示“反面”。我们可以设想这是一枚损坏的硬币，因此出现正面的概率不一定等于反面的概率。记 $x = 1$ 的概率为参数 μ，于是有

$$p(x = 1 \mid \mu) = \mu \tag{2.1}$$

其中 $0 \leq \mu \leq 1$，由此可得 $p(x = 0 \mid \mu) = 1 - \mu$。因此，随机变量 x 的概率分布可以写成如下形式：

$$\text{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x} \tag{2.2}$$

这被称为伯努利分布。很容易验证该分布是归一化的，并且它的均值和方差分别为

$$\mathbb{E}[x] = \mu \tag{2.3}$$

$$\text{var}[x] = \mu(1 - \mu) \tag{2.4}$$

现在假设我们有一个观测数据集 $D = \{x_1, \ldots, x_N\}$。在假设这些观测是独立地从分布 $p(x \mid \mu)$ 中抽取的前提下，可以构造似然函数，它是 μ 的函数：

$$p(\mathcal{D} \mid \mu) = \prod_{n=1}^{N} p(x_n \mid \mu) = \prod_{n=1}^{N} \mu^{x_n} (1 - \mu)^{1-x_n} \tag{2.5}$$

在频率学派的框架下，我们可以通过最大化似然函数来估计 μ 的值，等价地，也可以最大化似然函数的对数。在伯努利分布的情况下，对数似然函数为

$$\ln p(\mathcal{D} \mid \mu) = \sum_{n=1}^{N} \ln p\left(x_n \mid \mu\right) = \sum_{n=1}^{N} \{x_n \ln \mu + (1 - x_n) \ln(1 - \mu)\} \tag{2.6}$$

此时值得注意的是，对数似然函数依赖于 N 个观测值 x_n 的方式，仅通过它们的和 $\sum_n x_n$。这个和就是该分布下数据的一个充分统计量的例子。我们将在后面更详细地研究充分统计量的重要作用。如果令 $\ln p(D|\mu)$ 对 μ 的导数为零，就得到最大似然估计

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n \tag{2.7}$$

它也被称为样本均值。若记数据集中观测到 $x = 1$（正面）的次数为 m，那么式 (2.7) 可以写成

$$\mu_{\text{ML}} = \frac{m}{N} \tag{2.8}$$

因此，在最大似然框架下，硬币落到正面的概率由数据集中正面出现的比例给出。

现在假设我们掷一枚硬币 3 次，恰好观察到 3 次都是正面。此时 ($N = m = 3$)，于是最大似然估计得到 ($\mu_{ML} = 1$)。在这种情况下，最大似然的结果会预测所有未来的观察都将得到正面。常识告诉我们，这显然是不合理的，而事实上，这正是最大似然方法所带来的过拟合的一个极端例子。我们将会看到，通过为 (μ) 引入一个先验分布，可以得到更加合理的结论。

我们还可以求出在给定数据集大小为 N 的情况下，观察到 $x = 1$ 的次数 m 的分布。这称为二项分布，从 (2.5) 式中我们看到它与 $\mu^m(1 - \mu)^{N-m}$ 成比例。为了得到归一化系数，我们注意到在 N 次投掷硬币中，我们必须把所有可能得到 m 次正面的方式相加，因此二项分布可以写成：

$$\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m} \tag{2.9}$$

其中

$$\binom{N}{m} \equiv \frac{N!}{m!(N-m)!} \tag{2.10}$$

是从总共 N 个相同对象中选择 m 个对象的方法数。图 2.1 展示了 $N = 10$ 且 $\mu = 0.25$ 时的二项分布图。

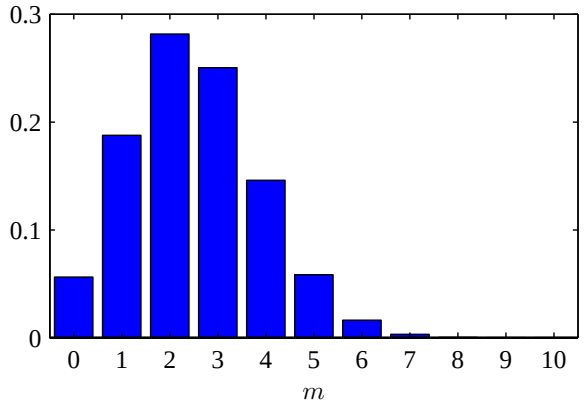

图 2.1 直方图绘制了二项分布 (2.9) 作为 m 的函数，其中 $N = 10$ 且 $\mu = 0.25$。

二项分布的均值和方差可以通过使用习题 1.10 的结果来求得，该习题表明对于独立事件，和的均值是均值的和，和的方差是方差的和。由于 $m = x_1 + \cdots + x_N$，并且对于每个观测，其均值和方差分别由 (2.3) 和 (2.4) 给出，因此我们有：

$$E[m] = \sum_{n=1}^{N} E[x_n] = \sum_{n=1}^{N} \mu = N\mu \tag{2.11}$$

$$\text{var}[m] = \sum_{n=1}^{N} \text{var}[x_n] = \sum_{n=1}^{N} \mu(1-\mu) = N\mu(1-\mu) \tag{2.12}$$

这些结果也可以直接使用微积分来证明。

2.1.1 贝塔分布

我们已经在 (2.8) 中看到，伯努利分布，以及相应的二项分布中参数 (μ) 的最大似然估计是数据集中 (x = 1) 的观测比例。正如我们已经指出的，对于较小的数据集，这会导致严重的过拟合。为了对该问题建立一个贝叶斯处理方法，我们需要在参数 (μ) 上引入一个先验分布 ($p(\mu)$)。在这里，我们考虑一种既有清晰解释又具有良好分析性质的先验分布。为说明这种先验的来源，注意到似然函数可以写成形如 ($\mu^x(1-\mu)^{1-x}$) 的多个因子的乘积。如果我们选择的先验在形式上与 (μ) 和 ($1-\mu$) 的幂成正比，那么后验分布就会与先验具有相同的函数形式，因为后验分布正比于先验与似然函数的乘积。这种性质称为共轭性，我们将在本章后续看到多个类似示例。因此，我们选择一种称为贝塔分布的先验，其形式为

$$\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{2.13}$$

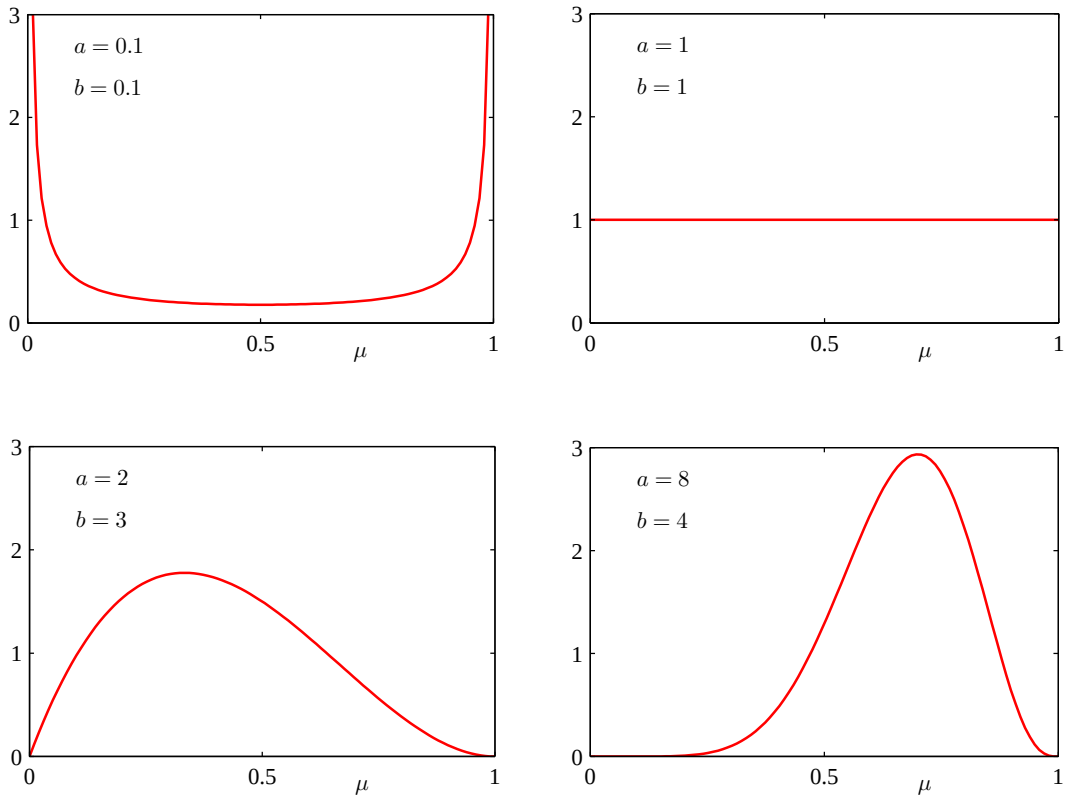

图 2.2 贝塔分布 ($\text{Beta}(\mu \mid a, b)$) 的曲线图，展示了式 (2.13) 所定义的分布随参数 (μ) 的变化情况，其中 (a) 和 (b) 取不同的超参数值。

其中 $\Gamma(x)$ 为伽马函数，其定义见公式 (1.141)。而式 (2.13)中的系数保证了贝塔分布是归一化的，即满足

$$\int_0^1 \text{Beta}(\mu \mid a, b), d\mu = 1. \tag{2.14}$$

贝塔分布的均值和方差为

$$\mathbb{E}[\mu] = \frac{a}{a+b} \tag{2.15}$$

$$\text{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}. \tag{2.16}$$

参数 a 和 b 通常被称为超参数，因为它们控制着参数 μ 的分布。图 2.2 展示了在不同超参数取值下的贝塔分布形状。

现在，通过将贝塔先验分布式 (2.13)与二项分布似然函数式 (2.9)相乘并进行归一化，可以得到参数 μ 的后验分布。保留仅依赖于 (μ) 的因子后，可以看到该后验分布具有如下形式：

$$p(\mu \mid m, l, a, b) \propto \mu^{m+a-1}(1-\mu)^{l+b-1} \tag{2.17}$$

其中 $l = N - m$ ，因此在掷硬币的例子中，l 对应于出现反面的次数。可以看到，式(2.17)与先验分布在 μ) 上具有相同的函数形式，这反映了该先验分布相对于似然函数的共轭性质。事实上，它依然是一个贝塔分布，其归一化系数可通过与式(2.13) 对比得

到：

$$p(\mu \mid m, l, a, b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1} \tag{2.18}$$

我们看到，观测到一个包含 m 个 $x=1$ 的样本和 l 个 $x=0$ 的样本，其效果是：在从先验分布到后验分布的过程中，将参数 a 的值增加了 m，并将参数 b 的值增加了 l。这使我们可以对先验分布中的超参数 a 和 b 给出一个简单的解释：它们分别可以看作是对 $x=1$ 和 $x=0$ 的“有效观测次数”。需要注意的是，a 和 b 不一定是整数。此外，如果我们随后又观测到了新的数据，那么后验分布还可以作为新的先验分布来使用。为说明这一点，我们可以设想依次逐个地获取观测值，并在每次观测之后，通过将当前的后验分布与新观测值的似然函数相乘，然后进行归一化，从而得到新的修正后验分布。在每一步中，后验分布都是一个 Beta 分布，其参数 a 和 b 分别代表（来自先验和实际观测的）$x=1$ 和 $x=0$ 的总观测次数。当加入一个新的 $x=1$ 的观测时，只需将 a 的值加 1；而当加入一个 $x=0$ 的观测时，只需将 b 的值加 1。图 2.3 展示了这一过程中的一个步骤。

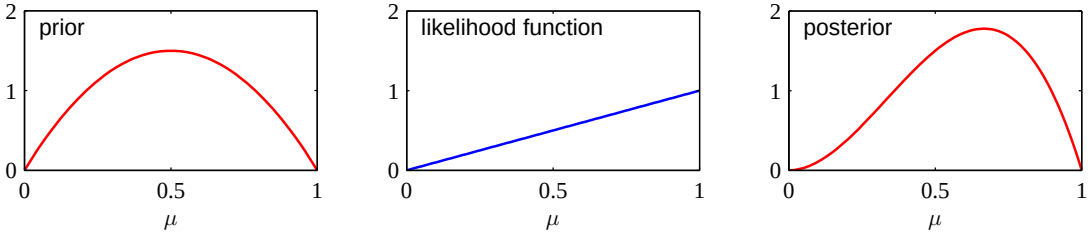

图 2.3 顺序贝叶斯推断过程中的单步示意图。先验分布由参数 $a=2$、$b=2$ 的 Beta 分布给出；似然函数由公式 (2.9) 给出，其中 $N=m=1$，对应于一次 $x=1$ 的观测，因此后验分布为参数 $a=3$、$b=2$ 的 Beta 分布。

我们可以看到，当采用贝叶斯观点时，这种顺序学习的方法自然地出现了。它与先验分布或似然函数的具体形式无关，只依赖于数据独立同分布（i.i.d.）的假设。顺序方法一次使用一个或一小批观测数据，然后在使用下一组数据之前丢弃之前的数据。这种方法可以应用于实时学习场景中，例如当数据以连续流的形式不断到来，而必须在尚未获得全部数据之前进行预测的情况。由于顺序方法不需要将整个数据集存储或加载到内存中，因此它们在处理大规模数据集时也非常有用。最大似然方法同样也可以被表述为一种顺序化的框架。

如果我们的目标是尽可能准确地预测下一次试验的结果，那么我们必须在给定观测数据集 $\mathcal{D}$ 的条件下，计算 x 的预测分布。根据概率的加法与乘法法则，这个分布可以写成如下形式：

$$p(x=1 \mid \mathcal{D}) = \int_0^1 p(x=1 \mid \mu)p(\mu \mid \mathcal{D})\mathrm{d}\mu = \int_0^1 \mu p(\mu \mid \mathcal{D})\mathrm{d}\mu = \mathbb{E}[\mu \mid \mathcal{D}]. \tag{2.19}$$

利用关于后验分布 $p(\mu \mid \mathcal{D})$ 的结果 (2.18)，以及关于 Beta 分布均值的结果 (2.15)，

我们得到

$$p(x=1 \mid \mathcal{D}) = \frac{m+a}{m+a+l+b} \tag{2.20}$$

这个结果可以作出一个简单的解释：它表示与 $x=1$ 相对应的观测（包括真实观测和先验中的虚拟观测）所占的总比例。需要注意的是，在数据集无限增大的极限下，即 $m,l \to \infty$ 时，结果 (2.10) 会退化为最大似然结果 (2.8)。正如我们将看到的那样，这是一个非常普遍的性质：在数据量无限大的情况下，贝叶斯结果与最大似然结果是一致的。对于有限的数据集，μ 的后验均值总是位于先验均值与由 (2.7) 给出的事件相对频率所对应的最大似然估计之间。

从图 2.2 可以看出，随着观测数量的增加，后验分布变得越来越尖锐。这一点也可以从 Beta 分布方差的结果 (2.16) 中看出，因为当 $a \to \infty$ 或 $b \to \infty$ 时，方差趋于零。事实上，我们可能会好奇：在贝叶斯学习中，是否普遍存在这样一种性质——当我们不断获取更多数据时，由后验分布所表示的不确定性会持续减小。

为了解答这一问题，我们可以从频率学派的角度来看待贝叶斯学习，并证明这种性质在平均意义下确实成立。考虑一个关于参数 θ 的一般贝叶斯推断问题，我们已经观测到一个数据集 D，其由联合分布 $p(\theta, D)$ 描述。接下来给出如下结果：

$$\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}] = \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta} \mid \mathcal{D}]\right] \tag{2.21}$$

其中

$$\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}] \equiv \int p(\boldsymbol{\theta})\boldsymbol{\theta}\mathrm{d}\boldsymbol{\theta} \tag{2.22}$$

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta} \mid \mathcal{D}]\right] \equiv \int \left\{ \int \boldsymbol{\theta} p(\boldsymbol{\theta} \mid \mathcal{D})\mathrm{d}\boldsymbol{\theta} \right\} p(\mathcal{D})\mathrm{d}\mathcal{D} \tag{2.23}$$

该结果表明：在生成数据的分布上取平均后，$\boldsymbol{\theta}$ 的后验均值等于 $\boldsymbol{\theta}$ 的先验均值。类似地，我们还可以证明

$$\mathrm{var}_{\boldsymbol{\theta}}[\boldsymbol{\theta}] = \mathbb{E}_{\mathcal{D}}\left[\mathrm{var}_{\boldsymbol{\theta}}[\boldsymbol{\theta} \mid \mathcal{D}]\right] + \mathrm{var}_{\mathcal{D}}\left[\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta} \mid \mathcal{D}]\right] \tag{2.24}$$

式 (2.24) 左边的项表示参数 θ 的先验方差。右边的第一项是 θ 的后验方差在数据生成分布上的平均值，第二项则衡量了 θ 的后验均值的方差。由于该方差项为正，这一结果表明：平均而言，θ 的后验方差小于其先验方差。而且，后验均值的方差越大，方差的减少幅度也就越大。但需要注意的是，这个结论只在平均意义下成立；对于某一个特定的观测数据集，后验方差有可能比先验方差更大。

2.2 多项式变量

二值变量用于描述只能取两种可能值的量。然而，在很多情况下，我们会遇到离散变量，它们可以取 K 个互斥的状态之一。虽然有多种方式可以表示这样的变量，但我们将很快看到，一种特别方便的表示方法是 1-of-K 编码。在这种表示中，变量被表示为一个 K 维向量 $\mathbf{x}$，其中某一个分量 $x_k = 1$，其余所有分量均为 0。例如，若某个变量可以取 $K=6$ 个状态，并且某次观测对应于第 3 个状态（即 $x_3 = 1$），那么该变量可以

表示为

$$\mathbf{x} = (0, 0, 1, 0, 0, 0)^{\mathrm{T}}. \tag{2.25}$$

注意，这样的向量满足 $\sum_{k=1}^{K} x_k = 1$. 如果我们用参数 μ_k 表示事件 $x_k = 1$ 的概率，则 $\mathbf{x}$ 的分布可以写为

$$p(\mathbf{x} \mid \boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}, \tag{2.26}$$

其中 $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}}$，并且参数满足约束条件 $\mu_k \geqslant 0$ 且 $\sum_k \mu_k = 1$, 因为它们表示概率。分布式 (2.26) 可以看作是伯努利分布）在多于两种结果时的推广。可以很容易验证该分布是归一化的：

$$\sum_{\mathbf{x}} p(\mathbf{x} \mid \boldsymbol{\mu}) = \sum_{k=1}^{K} \mu_k = 1, \tag{2.27}$$

并且有期望

$$\mathbb{E}[\mathbf{x} \mid \boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x} \mid \boldsymbol{\mu})\mathbf{x} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}} = \boldsymbol{\mu}. \tag{2.28}$$

现在考虑一个包含 N 个独立观测 $\mathbf{x}_1, \ldots, \mathbf{x}_N$ 的数据集 $\mathcal{D}$。其对应的似然函数为

$$p(\mathcal{D} \mid \boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{\left(\sum_n x_{nk}\right)} = \prod_{k=1}^{K} \mu_k^{m_k}. \tag{2.29}$$

我们可以看到，似然函数仅通过 K 个量 m_k 与 N 个数据点相关：

$$m_k = \sum_{n} x_{nk}, \tag{2.30}$$

其中 m_k 表示 $x_k = 1$ 被观测到的次数。这些量称为该分布的充分统计量。

为了求出 $\boldsymbol{\mu}$ 的最大似然解，我们需要在约束条件 $\sum_{k=1}^{K} \mu_k = 1$ 下，最大化 $\ln p(\mathcal{D} \mid \boldsymbol{\mu})$ 关于 μ_k 的值。为此，我们可以引入一个拉格朗日乘子 λ，并最大化以下函数：

$$\sum_{k=1}^{K} m_k \ln \mu_k + \lambda \left(\sum_{k=1}^{K} \mu_k - 1 \right). \tag{2.31}$$

对式 (2.31) 关于 μ_k 求导并令其为零，得到

$$\mu_k = -\frac{m_k}{\lambda}. \tag{2.32}$$

我们可以通过将式 (2.32) 代入约束条件 $\sum_k \mu_k = 1$ 来求解拉格朗日乘子 λ，从而得到 $\lambda = -N$。因此，最大似然解为

$$\mu_k^{\mathrm{ML}} = \frac{m_k}{N}, \tag{2.33}$$

这表示在 N 个观测中，事件 $x_k = 1$ 出现的比例。

我们还可以考虑在给定参数 $\boldsymbol{\mu}$ 和总观测数 N 的条件下，$m_1, \ldots, m_K$ 的联合分布。根据式 (2.29)，该分布可写为

$$\mathrm{Mult}(m_1, m_2, \ldots, m_K \mid \boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k}, \tag{2.34}$$

这称为多项式分布（multinomial distribution）。归一化系数表示将 N 个对象划分为大小分别为 $m_1, \ldots, m_K$ 的 K 个组的不同方式的数量，其形式为

$$\binom{N}{m_1 m_2 \ldots m_K} = \frac{N!}{m_1! m_2! \ldots m_K!}. \tag{2.35}$$

注意，变量 m_k 满足约束条件

$$\sum_{k=1}^{K} m_k = N. \tag{2.36}$$

2.2.1 狄利克雷分布

我们现在为多项式分布 (2.34) 的参数集合 μ_k 引入一族先验分布。根据多项式分布的形式可以看出，其共轭先验为

$$p(\boldsymbol{\mu} \mid \boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}, \tag{2.37}$$

其中 $0 \leqslant \mu_k \leqslant 1$ 且 $\sum_k \mu_k = 1$。这里的 $\alpha_1, \ldots, \alpha_K$ 是该分布的参数，记作 $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^{\mathrm{T}}$。由于存在求和约束，分布在 μ_k 空间中被限制在一个 $(K-1)$ 维的单纯形上，如图 2.4 中 $K = 3$ 的情况所示。

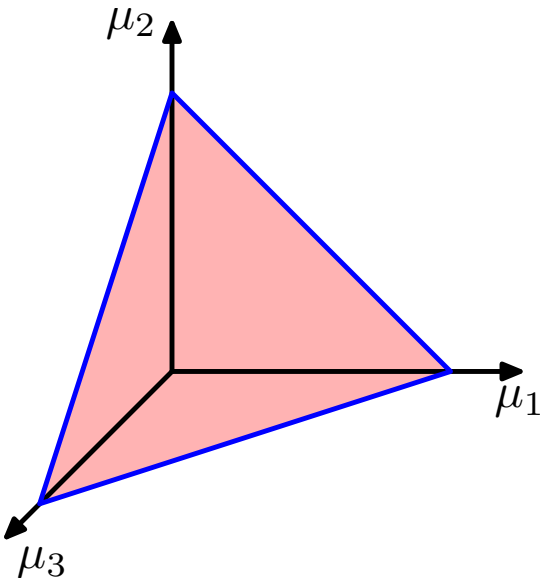

图 2.4 对三个变量 μ_1, μ_2, μ_3 的 Dirichlet 分布被限制在一个如图所示的单纯形（有界线性流形）中，这是由约束条件 $0 \leq \mu_k \leq 1$ 和 $\sum_k \mu_k = 1$ 所导致的。

该分布的归一化形式为

$$\mathrm{Dir}(\boldsymbol{\mu} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}, \tag{2.38}$$

这被称为狄利克雷分布。其中 $\Gamma(x)$ 是伽马函数，定义见式 (1.141)，并且

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k. \tag{2.39}$$

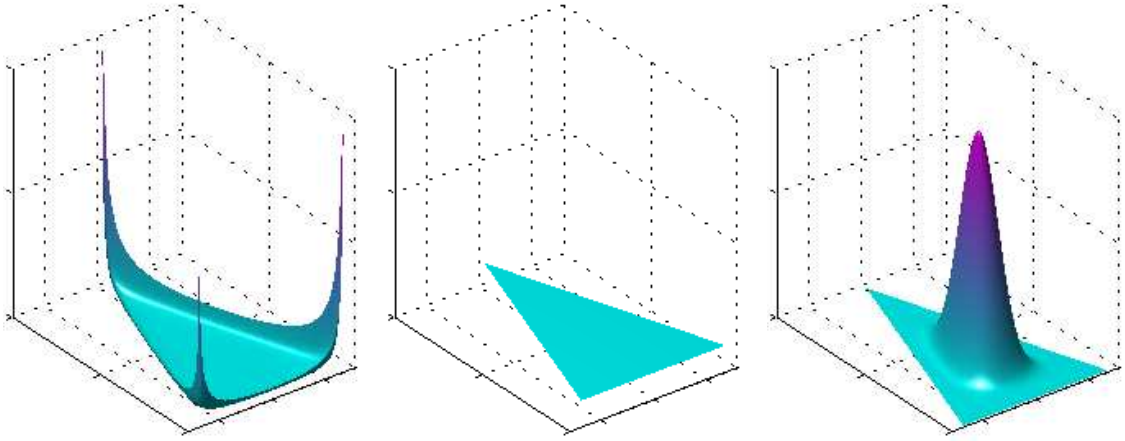

图 2.5 三元 Dirichlet 分布的图示中，两个水平坐标轴位于单纯形平面内，垂直轴表示密度值。左图中 $\{\alpha_k\} = 0.1$，中图中 $\{\alpha_k\} = 1$，右图中 $\{\alpha_k\} = 10$。

图 2.5 展示了在不同参数 α_k 设置下，狄利克雷分布在单纯形上的形状。

将先验分布 (2.38) 与似然函数 (2.34) 相乘，我们得到参数 μ_k 的后验分布：

$$p(\boldsymbol{\mu} \mid \mathcal{D}, \boldsymbol{\alpha}) \propto p(\mathcal{D} \mid \boldsymbol{\mu}), p(\boldsymbol{\mu} \mid \boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}. \tag{2.40}$$

可以看出，后验分布仍然是一个狄利克雷分布形式的分布，这验证了狄利克雷分布确实是多项式分布的共轭先验。通过与式 (2.38) 比较，我们可以确定其归一化系数，从而得到：

$$p(\boldsymbol{\mu} \mid \mathcal{D}, \boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\mu} \mid \boldsymbol{\alpha} + \mathbf{m}) \quad = \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1) \cdots \Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}, \tag{2.41}$$

其中 $\mathbf{m} = (m_1, \ldots, m_K)^{\mathrm{T}}$。与 Beta 分布作为二项分布的共轭先验的情形类似，我们可以将狄利克雷先验的参数 α_k 理解为对事件 $x_k = 1$ 的“有效观测次数”。

需要注意的是，二状态变量既可以表示为二值变量，并使用二项分布 (2.9) 建模；也可以表示为 1-of-2 变量，并使用 $K = 2$ 的多项式分布 (2.34) 建模。

2.3 高斯分布

高斯分布，又称正态分布，是一种广泛用于描述连续变量分布的模型。对于单个变量 x，高斯分布可以写为

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}, \tag{2.42}$$

其中，μ 为均值，σ^2 为方差。对于 D 维向量 $\mathbf{x}$，多元高斯分布的形式为

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\}, \tag{2.43}$$

其中，$\boldsymbol{\mu}$ 是 D 维均值向量，$\boldsymbol{\Sigma}$ 是 $D \times D$ 协方差矩阵，而 $|\boldsymbol{\Sigma}|$ 表示其行列式。

高斯分布在许多不同的情境中都会自然出现，可以从多种角度加以理解。例如，对于单个实数变量，能最大化熵的分布就是高斯分布。这一性质同样适用于多元高斯情形。

另一种出现高斯分布的情形是当我们考虑多个随机变量的和时。中心极限定理指出：在某些较弱的条件下，一组随机变量的和（本身也是一个随机变量）的分布会随着求和项数的增加而逐渐趋近于高斯分布（参见 Walker, 1969）。例如，假设有 N 个随机变量 $x_1, \ldots, x_N$，它们在区间 $[0, 1]$ 上均匀分布，我们关心它们的平均值 $\frac{x_1+x_2+\cdots+x_N}{N}$. 当 N 很大时，该平均值的分布会趋近于高斯分布，如图 2.6 所示。实际上，随着 N 的增大，向高斯分布的收敛速度非常快。这一结果的一个直接推论是：二项分布 (2.9)，即由对随机二值变量 x 进行 N 次观测后得到的 m 的分布，在 $N \to \infty$ 时也将趋近于高斯分布（见图 2.1 中 $N = 10$ 的情况）。

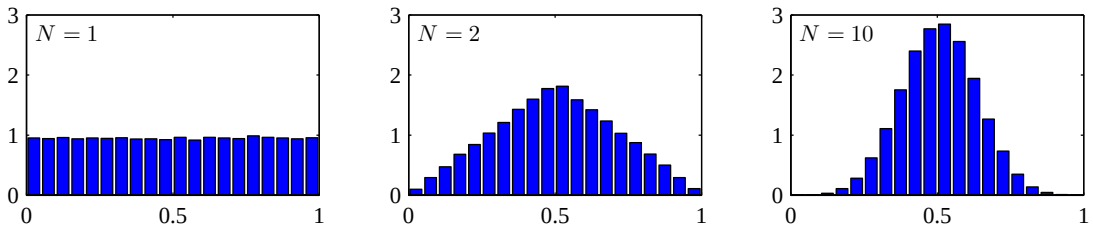

图 2.6 对于不同取值的 N，绘制由 N 个均匀分布数值的均值所构成的直方图。可以观察到，随着 N 的增大，该分布趋近于高斯分布。

高斯分布具有许多重要的解析性质，本节将详细讨论其中的一些。由于涉及较多的矩阵恒等式，本节的技术细节会比之前的章节更复杂。然而，掌握这里介绍的高斯分布操作技巧，将对理解后续章节中的复杂模型非常有帮助。

我们首先从几何角度来考察高斯分布的形式。高斯分布对 x 的函数依赖体现在如下二次型中：

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}), \tag{2.44}$$

其中量 Δ 被称为马氏距离，表示从 $\boldsymbol{\mu}$ 到 $\mathbf{x}$ 的距离。当协方差矩阵 $\boldsymbol{\Sigma}$ 为单位矩阵时，马氏距离退化为欧氏距离。因此，高斯分布在 $\mathbf{x}$ 空间中将保持常数于使该二次型取相同值的曲面上。

首先我们注意到，矩阵 $\boldsymbol{\Sigma}$ 可以在不失一般性的情况下视为对称矩阵，因为任何反对称部分在指数中都会消失。考虑协方差矩阵的特征向量方程：

$$\boldsymbol{\Sigma} \mathbf{u}_i = \lambda_i \mathbf{u}_i, \tag{2.45}$$

其中 $i = 1, \ldots, D$。由于 $\boldsymbol{\Sigma}$ 是实对称矩阵，其特征值 λ_i 均为实数，并且其特征向量 $\mathbf{u}_i$ 可以选择为正交归一（orthonormal）的一组向量，因此有

$$\mathbf{u}_i^{\mathrm{T}} \mathbf{u}_j = I_{ij}, \tag{2.46}$$

其中 I_{ij} 是单位矩阵的第 (i,j) 个元素，满足

$$I_{ij} = \begin{cases} 1, & 若i = j, \\ 0, & 否则. \end{cases} \tag{2.47}$$

协方差矩阵 $\boldsymbol{\Sigma}$ 可以用其特征向量展开为

$$\boldsymbol{\Sigma} = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}}, \tag{2.48}$$

而其逆矩阵 $\boldsymbol{\Sigma}^{-1}$ 则可写为

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}}. \tag{2.49}$$

将上述式子代入 (2.44)，可得该二次型为

$$\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}, \tag{2.50}$$

其中我们定义

$$y_i = \mathbf{u}_i^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}). \tag{2.51}$$

我们可以将 y_i 理解为由正交归一向量 $\mathbf{u}_i$ 定义的新坐标系，它相对于原坐标 x_i 进行了旋转和平移。形成向量 $\mathbf{y} = (y_1, \ldots, y_D)^{\mathrm{T}}$, 则有

$$\mathbf{y} = \mathbf{U}(\mathbf{x} - \boldsymbol{\mu}), \tag{2.52}$$

其中，矩阵 $\mathbf{U}$ 的各行由 $\mathbf{u}_i^{\mathrm{T}}$ 给出。由式 (2.46) 可知，$\mathbf{U}$ 是一个正交矩阵，即它满足 $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}$, 因此也有 $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$, 其中 $\mathbf{I}$ 为单位矩阵。

该二次型（以及相应的高斯密度）在式 (2.51) 的值保持不变的曲面上取常数。如果所有特征值 λ_i 都为正数，这些曲面就是椭球面，它们的中心位于 $\boldsymbol{\mu}$，主轴方向由特征向量 $\mathbf{u}_i$ 给定，而在每个方向上的伸缩尺度由 $\lambda_i^{1/2}$ 决定，如图 2.7 所示。

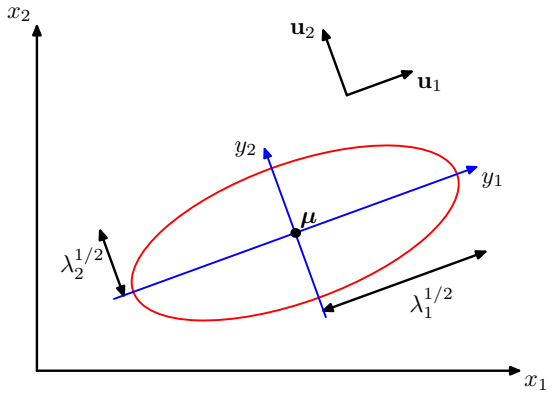

图 2.7 红色曲线表示二维空间 $\mathbf{x} = (x_1, x_2)$ 中高斯分布的等概率密度椭面，其密度为在 $\mathbf{x} = \mu$ 处取值的 $\exp(-1/2)$。该椭圆的主轴由协方差矩阵的特征向量 $\mathbf{u}_i$ 定义，且对应的特征值为 λ_i。

为了保证高斯分布是良好定义的，协方差矩阵的所有特征值 λ_i 必须严格为正，否则分布无法正确归一化。一个所有特征值都为正的矩阵称为正定矩阵。在第 12 章中，我

们将遇到一些特征值为零的高斯分布，这种情况下分布是奇异的，仅定义在低维子空间中。如果所有特征值非负（即允许为零），则称协方差矩阵是半正定矩阵。

现在考虑在由 y_i 定义的新坐标系下高斯分布的形式。从 $\mathbf{x}$ 到 $\mathbf{y}$ 坐标系的变换具有一个雅可比矩阵 $\mathbf{J}$，其元素为

$$J_{ij} = \frac{\partial x_i}{\partial y_j} = U_{ji}, \tag{2.53}$$

其中 U_{ji} 是矩阵 $\mathbf{U}^{\mathrm{T}}$ 的元素。利用 $\mathbf{U}$ 的正交性，可得雅可比矩阵行列式的平方为

$$|\mathbf{J}|^2 = |\mathbf{U}^{\mathrm{T}}|^2 = |\mathbf{U}^{\mathrm{T}}|, |\mathbf{U}| = |\mathbf{U}^{\mathrm{T}}\mathbf{U}| = |\mathbf{I}| = 1, \tag{2.54}$$

因此有 $|\mathbf{J}| = 1$。此外，协方差矩阵 $\mathbf{\Sigma}$ 的行列式可以表示为其特征值的乘积，因此有

$$|\mathbf{\Sigma}|^{1/2} = \prod_{j=1}^{D} \lambda_j^{1/2}. \tag{2.55}$$

因此，在 y_j 坐标系下，高斯分布可以写为

$$p(\mathbf{y}) = p(\mathbf{x})|\mathbf{J}| = \prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\left\{-\frac{y_j^2}{2\lambda_j}\right\}, \tag{2.56}$$

这表示为 D 个相互独立的单变量高斯分布的乘积。由此可见，特征向量 $\mathbf{u}_i$ 定义了一个经过平移与旋转的新坐标系，在该坐标系下，联合概率分布可分解为独立分布的乘积。在 $\mathbf{y}$ 坐标系中，对该分布进行积分得到

$$\int p(\mathbf{y}), \mathrm{d}\mathbf{y} = \prod_{j=1}^{D} \int_{-\infty}^{\infty} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\left\{-\frac{y_j^2}{2\lambda_j}\right\} \mathrm{d}y_j = 1, \tag{2.57}$$

其中使用了单变量高斯分布归一化的结果 (1.48)。这验证了多元高斯分布 (2.43) 确实是归一化的。

接下来我们来看高斯分布的矩，并由此解释参数 $\boldsymbol{\mu}$ 和 $\mathbf{\Sigma}$ 的意义。在高斯分布下，$\mathbf{x}$ 的期望为

$$\begin{aligned} \mathbb{E}[\mathbf{x}] &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \mathbf{x}, \mathrm{d}\mathbf{x} \\ &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{z}\right\} (\mathbf{z}+\boldsymbol{\mu}), \mathrm{d}\mathbf{z}, \end{aligned} \tag{2.58}$$

其中我们进行了变量替换 $\mathbf{z} = \mathbf{x} - \boldsymbol{\mu}$。注意到指数项是关于 $\mathbf{z}$ 各分量的偶函数，并且积分区间为 $(-\infty, \infty)$，因此在 $(\mathbf{z}+\boldsymbol{\mu})$ 中与 $\mathbf{z}$ 线性相关的项由于对称性消失。于是有

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}, \tag{2.59}$$

因此称 $\boldsymbol{\mu}$ 为高斯分布的均值。

现在考虑高斯分布的二阶矩。在一维情形下，我们研究的是 $\mathbb{E}[x^2]$；而在多维情况下，有 D^2 个二阶矩 $\mathbb{E}[x_i x_j]$，我们可以将它们组合成一个矩阵 $\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}]$。该矩阵可写为

$$\begin{aligned} \mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \mathbf{x}\mathbf{x}^{\mathrm{T}}, \mathrm{d}\mathbf{x} \\ &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{z}\right\} (\mathbf{z}+\boldsymbol{\mu})(\mathbf{z}+\boldsymbol{\mu})^{\mathrm{T}}, \mathrm{d}\mathbf{z}. \end{aligned}$$

这里我们再次进行了变量替换 $\mathbf{z} = \mathbf{x} - \boldsymbol{\mu}$。注意，其中涉及 $\boldsymbol{\mu}\mathbf{z}^{\mathrm{T}}$ 和 $\boldsymbol{\mu}^{\mathrm{T}}\mathbf{z}$ 的交叉项会由于对称性而消失。项 $\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}$ 是常数，可以从积分中提出，而该积分本身等于 1，因为高斯分布已经归一化。现在考虑涉及 $\mathbf{z}\mathbf{z}^{\mathrm{T}}$ 的项。我们再次利用协方差矩阵的特征向量展开式 (2.45)，并利用特征向量集合的完备性，有

$$\mathbf{z} = \sum_{j=1}^{D} y_j \mathbf{u}_j, \tag{2.60}$$

其中 $y_j = \mathbf{u}_j^{\mathrm{T}}\mathbf{z}$。因此可得

$$\begin{aligned} &\frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\} \mathbf{z}\mathbf{z}^{\mathrm{T}}, \mathrm{d}\mathbf{z} \\ &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \sum_{i=1}^{D}\sum_{j=1}^{D} \mathbf{u}_i\mathbf{u}_j^{\mathrm{T}} \int \exp\left\{-\sum_{k=1}^{D}\frac{y_k^2}{2\lambda_k}\right\} y_i y_j, \mathrm{d}\mathbf{y} \\ &= \sum_{i=1}^{D} \mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\lambda_i = \boldsymbol{\Sigma}. \end{aligned} \tag{2.61}$$

在这里我们使用了特征向量方程 (2.45)，并利用了中间行右侧的积分在 $i \neq j$ 时因对称性为零的事实。在最后一步，我们结合结果 (1.50)、(2.55) 以及 (2.48) 得到上式。于是

$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}. \tag{2.62}$$

在单变量情形下，我们通过减去均值再取二阶矩来定义方差。类似地，在多变量情况下，也可以通过减去均值来定义协方差，即

$$\mathrm{cov}[\mathbf{x}] = \mathbb{E}!\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}\right]. \tag{2.63}$$

对于高斯分布的特例，结合 $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$ 和上式结果 (2.62)，得到

$$\mathrm{cov}[\mathbf{x}] = \boldsymbol{\Sigma}. \tag{2.64}$$

因此，参数矩阵 $\boldsymbol{\Sigma}$ 控制了高斯分布下 $\mathbf{x}$ 的协方差，因此称其为协方差矩阵。

虽然高斯分布 (2.43) 被广泛用作密度模型，但它也存在显著的局限性。考虑其自由参数的数量：一般的对称协方差矩阵 $\boldsymbol{\Sigma}$ 具有 $\frac{D(D+1)}{2}$ 个独立参数，而均值向量 $\boldsymbol{\mu}$ 另有 D 个独立参数，因此总参数数量为 $\frac{D(D+3)}{2}$. 当维度 D 很大时，参数的总数会随 D 二次增长，从而使得在计算上操作和求逆大型矩阵的代价极高。为应对这一问题，可以使用一些受限形式的协方差矩阵。例如，若我们考虑对角协方差矩阵，即 $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_i^2)$, 那么模型中共有 $2D$ 个独立参数。此时的等密度曲线是坐标轴对齐的椭球面。进一步地，我们还可以将协方差矩阵约束为与单位矩阵成比例，即 $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, 这称为各向同性协方差，此时模型仅包含 $D+1$ 个独立参数，对应的等密度面为球面。一般协方差矩阵、对角协方差矩阵和各向同性协方差矩阵这三种情况如图 2.8 所示。不幸的是，虽然这些简化方法减少了分布的自由度，并使协方差矩阵求逆更加高效，但它们也极大地限制了分布的形状，从而削弱了模型捕捉数据中有趣相关结构的能力。

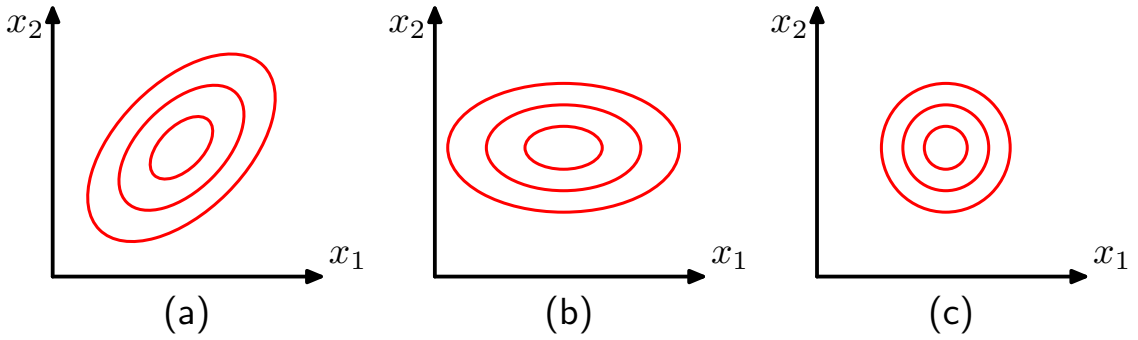

图 2.8 二维高斯分布的等概率密度轮廓：(a) 协方差矩阵为一般形式；(b) 协方差矩阵为对角形式，此时椭圆轮廓与坐标轴对齐；(c) 协方差矩阵与单位矩阵成比例，此时轮廓为同心圆。

此外，高斯分布本身存在一个根本限制：它固有地是单峰的，即仅有一个最大值，因此无法很好地逼近多峰分布。由此可见，高斯分布在某种意义上既“过于灵活”——参数太多；又“过于受限”——无法表示足够复杂的概率结构。后续我们将看到，通过引入潜变量，也称为隐变量或未观测变量，这两类问题都能得到解决。特别地，若引入离散潜变量，即可得到一类丰富的多峰分布——即高斯混合模型，这一内容将在第 2.3.9 节中讨论。而若引入连续潜变量（第 12 章将详细介绍），则可以构建一类模型，其自由参数数量可独立于数据空间维度 D 而被控制，同时仍能捕捉数据中的主要相关结构。事实上，这两种方法可以结合，并进一步扩展为一整套层次化模型，可适应各种实际应用场景。例如，高斯马尔可夫随机场（Gaussian Markov Random Field）是一种广泛用于图像建模的概率模型。它在像素强度的联合空间上是一个高斯分布，但由于在模型中引入了反映像素空间结构的约束，从而保持了可计算性。类似地，线性动态系统常用于时间序列建模（如目标跟踪），它也是在观测变量与潜变量的联合空间上的高斯分布，并且由于其结构化形式而保持可解性。描述这些复杂分布形式及其性质的强大框架是概率图模型，这将是第 8 章的主题。

2.3.1 条件高斯分布

高斯分布的一个重要性质是：如果两个变量集合联合服从多元高斯分布，那么其中一个变量在给定另一个变量条件下的条件分布仍然是高斯分布；同时，每个变量集合的边缘分布也都是高斯的。

首先考虑条件分布的情形。假设 $\mathbf{x}$ 是一个维度为 D 的随机向量，服从高斯分布 $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$，并将其划分为两个互不相交的子向量 $\mathbf{x}_a$ 和 $\mathbf{x}_b$。不失一般性地，可以令 $\mathbf{x}_a$ 为 $\mathbf{x}$ 的前 M 个分量，而 $\mathbf{x}_b$ 为剩下的 $D - M$ 个分量，于是有

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}. \tag{2.65}$$

相应地，均值向量 $\boldsymbol{\mu}$ 也被划分为

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \tag{2.66}$$

而协方差矩阵 $\boldsymbol{\Sigma}$ 被分块写为

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}. \tag{2.67}$$

由于协方差矩阵是对称的，即 $\boldsymbol{\Sigma}^{\mathrm{T}} = \boldsymbol{\Sigma}$，因此 $\boldsymbol{\Sigma}_{aa}$ 和 $\boldsymbol{\Sigma}_{bb}$ 都是对称矩阵，且有 $\boldsymbol{\Sigma}_{ba} = \boldsymbol{\Sigma}_{ab}^{\mathrm{T}}$.

在许多情况下，使用协方差矩阵的逆会更加方便，我们定义

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}, \tag{2.68}$$

称 $\boldsymbol{\Lambda}$ 为精度矩阵。事实上，高斯分布的一些性质用协方差矩阵表达最为自然，而另一些性质在用精度矩阵表示时更为简洁。因此，我们也将精度矩阵写成与 $\mathbf{x}$ 分块形式对应的结构：

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}. \tag{2.69}$$

由于对称矩阵的逆仍然是对称的，因此 $\boldsymbol{\Lambda}_{aa}$ 和 $\boldsymbol{\Lambda}_{bb}$ 也是对称矩阵，并且有 $\boldsymbol{\Lambda}_{ab}^{\mathrm{T}} = \boldsymbol{\Lambda}_{ba}$. 需要强调的是，$\boldsymbol{\Lambda}_{aa}$ 并不等于 $\boldsymbol{\Sigma}_{aa}$ 的逆。稍后我们将讨论分块矩阵的逆与其各分块之间的关系。

我们现在来求条件分布 $p(\mathbf{x}_a \mid \mathbf{x}_b)$ 的表达式。我们可以看到，条件分布可以直接从联合分布 $p(\mathbf{x}) = p(\mathbf{x}_a, \mathbf{x}_b)$ 中求得，只需将 $\mathbf{x}_b$ 固定为其观测值，然后对结果进行归一化，以得到关于 $\mathbf{x}_a$ 的有效概率分布。与其显式地执行这种归一化操作，不如更高效地通过考察高斯分布 (2.44) 指数中的二次型来求解，最后再补上归一化系数。利用分块形式 (2.65)、(2.66) 和 (2.69)，可得

$$\begin{aligned} -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) & \\ = & -\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) \\ & -\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b). \end{aligned} \tag{2.70}$$

可以看到，这个式子作为 $\mathbf{x}_a$ 的函数时仍然是一个二次型，因此相应的条件分布 $p(\mathbf{x}_a \mid \mathbf{x}_b)$ 也是一个高斯分布。由于高斯分布完全由其均值和协方差确定，我们的目标是通过观察式 (2.70) 来确定 $p(\mathbf{x}_a \mid \mathbf{x}_b)$ 的均值与协方差的表达式。

这种操作在高斯分布的推导中非常常见，被称为配方。在这种情况下，我们已知高斯分布指数中的二次型，需要从中确定对应的均值和协方差。对于一般高斯分布 $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$，其指数项可写为

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\mathrm{const}, \tag{2.71}$$

其中“const”表示与 $\mathbf{x}$ 无关的常数项，并且利用了 $\boldsymbol{\Sigma}$ 的对称性。因此，如果我们把一个给定的二次型写成右侧这种形式，就可以直接识别出：$\mathbf{x}$ 二次项的系数矩阵即为逆协方差矩阵 $\boldsymbol{\Sigma}^{-1}$；线性项的系数为 $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$，从而可以求出 $\boldsymbol{\mu}$。

现在将这一方法应用到条件高斯分布 $p(\mathbf{x}_a \mid \mathbf{x}_b)$ 上。对于式 (2.70) 的指数部分，设其均值与协方差分别为 $\boldsymbol{\mu}_{a|b}$ 和 $\boldsymbol{\Sigma}_{a|b}$。考虑其关于 $\mathbf{x}_a$ 的依赖形式，并把 $\mathbf{x}_b$ 看作常量。取出其中关于 $\mathbf{x}_a$ 的二次项，可得

$$-\frac{1}{2}\mathbf{x}_a^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}\mathbf{x}_a. \tag{2.72}$$

由此我们可以直接得到条件分布的协方差（或逆精度）为

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1}. \tag{2.73}$$

现在考虑式 (2.70) 中所有关于 $\mathbf{x}_a$ 的一次项

$$\mathbf{x}_a^{\mathrm{T}}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right)\right\}, \tag{2.74}$$

其中我们使用了 $\boldsymbol{\Lambda}_{ba}^{\mathrm{T}} = \boldsymbol{\Lambda}_{ab}$。根据对一般形式 (2.71) 的讨论，上式中 $\mathbf{x}_a$ 的系数必须等于 $\boldsymbol{\Sigma}_{a|b}^{-1}\boldsymbol{\mu}_{a|b}$，因此有

$$\begin{aligned}\boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right)\right\} \\ &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right)\end{aligned} \tag{2.75}$$

其中我们使用了式 (2.73)。

结果 (2.73) 和 (2.75) 是用原联合分布 $p(\mathbf{x}_a, \mathbf{x}_b)$ 的分块精度矩阵来表示的。我们也可以把这些结果用相应的分块协方差矩阵来表示。为此，使用分块矩阵求逆的如下恒等式

$$\begin{pmatrix}\mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D}\end{pmatrix}^{-1} = \begin{pmatrix}\mathbf{M} & -\mathbf{MBD}^{-1} \\ -\mathbf{D}^{-1}\mathbf{CM} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{CMBD}^{-1}\end{pmatrix}, \tag{2.76}$$

其中

$$\mathbf{M} = \left(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C}\right)^{-1}. \tag{2.77}$$

量 $\mathbf{M}^{-1}$ 称为左侧矩阵关于子矩阵 $\mathbf{D}$ 的 Schur 补。利用定义

$$\begin{pmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{pmatrix}^{-1} = \begin{pmatrix}\boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb}\end{pmatrix} \tag{2.78}$$

并结合 (2.76)，得到

$$\boldsymbol{\Lambda}_{aa} = \left(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}\right)^{-1}, \tag{2.79}$$

$$\boldsymbol{\Lambda}_{ab} = -\left(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}\right)^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}. \tag{2.80}$$

由此可得条件分布 $p(\mathbf{x}_a \mid \mathbf{x}_b)$ 的均值与协方差为

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right), \tag{2.81}$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}. \tag{2.82}$$

对比 (2.73) 与 (2.82) 可以看到，当用分块精度矩阵来表示时，条件分布 $p(\mathbf{x}_a \mid \mathbf{x}_b)$ 的形式更为简洁。注意，由 (2.81) 给出的条件分布的均值是 $\mathbf{x}_b$ 的线性函数，而由 (2.82) 给出的协方差与 $\mathbf{x}_a$ 无关。这就是线性—高斯模型的一个实例。

2.3.2 边缘高斯分布

我们已经看到，如果联合分布 $p(\mathbf{x}_a, \mathbf{x}_b)$ 是高斯分布，那么条件分布 $p(\mathbf{x}_a \mid \mathbf{x}_b)$ 仍将是高斯分布。现在我们转向讨论由下式给出的边缘分布：

$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b)\, \mathrm{d}\mathbf{x}_b \tag{2.83}$$

正如我们将要看到的，这个边缘分布也是高斯分布。同样，我们有效计算此分布的策略是：关注联合分布指数中的二次型，从而识别边缘分布 $p(\mathbf{x}_a)$ 的均值和协方差。

联合分布的二次型可以使用分块精度矩阵，以 (2.70) 式的形式表示。因为我们的目标是积分消去 $\mathbf{x}_b$，最容易的实现方式是：首先考虑涉及 $\mathbf{x}_b$ 的项，然后进行配方，以利于积分。挑出仅涉及 $\mathbf{x}_b$ 的项，我们得到

$$-\frac{1}{2}\mathbf{x}_b^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}\mathbf{x}_b + \mathbf{x}_b^{T}\mathbf{m} = -\frac{1}{2}\left(\mathbf{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m}\right)^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}\left(\mathbf{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m}\right) + \frac{1}{2}\mathbf{m}^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m} \tag{2.84}$$

其中我们定义了

$$\mathbf{m} = \boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{ba}\left(\mathbf{x}_a - \boldsymbol{\mu}_a\right) \tag{2.85}$$

我们看到，对 $\mathbf{x}_b$ 的依赖性已被转化为高斯分布的标准二次型形式（对应于 (2.84) 式右侧的第一项），再加上一个不依赖于 $\mathbf{x}_b$ 的项（但它确实依赖于 $\mathbf{x}_a$）。因此，当我们取这个二次型的指数时，我们看到 (2.83) 所要求的对 $\mathbf{x}_b$ 的积分将采取以下形式：

$$\int \exp\left\{-\frac{1}{2}\left(\mathbf{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m}\right)^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}\left(\mathbf{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m}\right)\right\}\mathrm{d}\mathbf{x}_b \tag{2.86}$$

通过注意到这是一个对未归一化高斯函数的积分，可以很容易地执行此积分，因此结果将是归一化系数的倒数。我们从 (2.43) 给出的归一化高斯函数形式可知，该系数与均值无关，仅取决于协方差矩阵的行列式。因此，通过对 $\mathbf{x}_b$ 进行配方，我们可以积分消去 $\mathbf{x}_b$，而 (2.84) 式左侧贡献中唯一剩余的且依赖于 $\mathbf{x}_a$ 的项是 (2.84) 式右侧的最后一项，其中 $\mathbf{m}$ 由 (2.85) 给出。将这一项与 (2.70) 式中剩余的依赖于 $\mathbf{x}_a$ 的项结合起来，我们得到

$$\begin{aligned}
&\tfrac{1}{2}\left[\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{ba}\left(\mathbf{x}_a - \boldsymbol{\mu}_a\right)\right]^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}^{-1}\left[\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{ba}\left(\mathbf{x}_a - \boldsymbol{\mu}_a\right)\right] \\
&-\tfrac{1}{2}\mathbf{x}_a^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}\mathbf{x}_a + \mathbf{x}_a^{\mathrm{T}}\left(\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b\right) + \text{const} \\
&= -\tfrac{1}{2}\mathbf{x}_a^{\mathrm{T}}\left(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\right)\mathbf{x}_a \\
&+\mathbf{x}_a^{\mathrm{T}}\left(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\right)^{-1}\boldsymbol{\mu}_a + \text{const}
\end{aligned} \tag{2.87}$$

其中“const”表示与 $\mathbf{x}_a$ 无关的量。同样，通过与 (2.71) 式比较，我们看到边缘分布 $p(\mathbf{x}_a)$ 的协方差由下式给出

$$\boldsymbol{\Sigma}_a = \left(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\right)^{-1} \tag{2.88}$$

类似地，均值由下式给出

$$\boldsymbol{\Sigma}_a\left(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\right)\boldsymbol{\mu}_a = \boldsymbol{\mu}_a \tag{2.89}$$

其中我们使用了 (2.88) 式。(2.88) 式中的协方差是根据 (2.69) 式给出的分块精度矩阵来表达的。我们可以将其改写成 (2.67) 式给出的协方差矩阵的相应分块形式，就像我们对条件分布所做的那样。这些分块矩阵之间的关系是

$$\begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \tag{2.90}$$

利用 (2.76) 式，我们得到

$$\left(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\right)^{-1} = \boldsymbol{\Sigma}_{aa} \tag{2.91}$$

因此我们得到了一个直观上令人满意的结果，即边缘分布 $p(\mathbf{x}_a)$ 的均值和协方差由下式给出

$$\mathbb{E}[\mathbf{x}_a] = \boldsymbol{\mu}_a \tag{2.92}$$

$$\operatorname{cov}[\mathbf{x}_a] = \boldsymbol{\Sigma}_{aa}. \tag{2.93}$$

我们看到，对于边缘分布，均值和协方差用分块协方差矩阵来表达最为简洁，这与条件分布形成了对比，条件分布的分块精度矩阵产生了更简洁的表达式。

我们对分块高斯分布的边缘分布和条件分布的结果总结如下。

分块高斯分布

给定联合高斯分布 $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$，其中 $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$ 以及

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \tag{2.94}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}. \tag{2.95}$$

条件分布：

$$p(\mathbf{x}_a \mid \mathbf{x}_b) = \mathcal{N}\left(\mathbf{x}_a \mid \boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}\right), \tag{2.96}$$

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right). \tag{2.97}$$

边缘分布：

$$p(\mathbf{x}_a) = \mathcal{N}\left(\mathbf{x}_a \mid \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}\right). \tag{2.98}$$

我们通过一个涉及两个变量的例子（如图 2.9 所示）来说明多元高斯分布中的条件分布与边缘分布的概念。

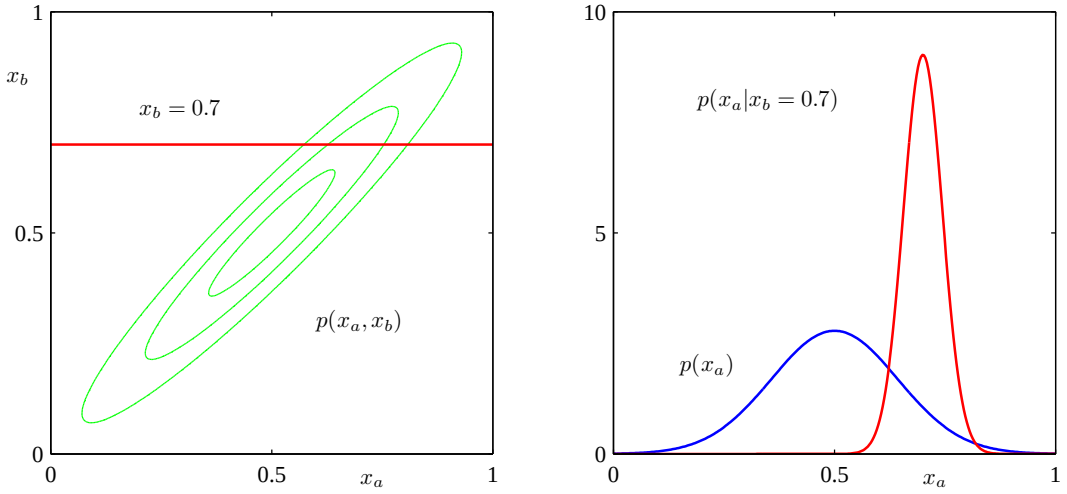

图 2.9 左图显示了关于两个变量的高斯分布 $p(x_a, x_b)$ 的等高线；右图显示了边缘分布 $p(x_a)$（蓝色曲线）以及在 $x_b = 0.7$ 时的条件分布 $p(x_a \mid x_b)$（红色曲线）。

2.3.3 高斯变量的贝叶斯定理

在 2.3.1 和 2.3.2 节中，我们考虑了一个高斯分布 $p(\mathbf{x})$，将向量 $\mathbf{x}$ 分割为两个子向量 $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)$，并求出了条件分布 $p(\mathbf{x}_a \mid \mathbf{x}_b)$ 和边缘分布 $p(\mathbf{x}_a)$ 的表达式。我们注意到，条件分布 $p(\mathbf{x}_a \mid \mathbf{x}_b)$ 的均值是 $\mathbf{x}_b$ 的线性函数。这里我们假设已知一个高斯边缘分布 $p(\mathbf{x})$ 和一个高斯条件分布 $p(\mathbf{y} \mid \mathbf{x})$，其中 $p(\mathbf{y} \mid \mathbf{x})$ 的均值是 $\mathbf{x}$ 的线性函数，且协方差与 $\mathbf{x}$ 无关。这是一个线性高斯模型（Roweis and Ghahramani, 1999）的例子，我们将在 8.1.4 节中更一般地研究它。我们希望求出边缘分布 $p(\mathbf{y})$ 和条件分布 $p(\mathbf{x} \mid \mathbf{y})$。这个问题在后续章节中会频繁出现，因此在这里推导出一般结果会很方便。

我们取边缘分布和条件分布分别为

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{2.99}$$

$$p(\mathbf{y} \mid \mathbf{x}) = \mathcal{N}(\mathbf{y} \mid \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{2.100}$$

其中 $\boldsymbol{\mu}$、$\mathbf{A}$ 和 $\mathbf{b}$ 是控制均值的参数，$\boldsymbol{\Lambda}$ 和 $\mathbf{L}$ 是精度矩阵。如果 $\mathbf{x}$ 的维数为 M，$\mathbf{y}$ 的维数为 D，则矩阵 $\mathbf{A}$ 的尺寸为 $D \times M$。

首先我们求 $\mathbf{x}$ 和 $\mathbf{y}$ 的联合分布表达式。为此，定义

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \tag{2.101}$$

然后考虑联合分布的对数

$$\begin{aligned} \ln p(\mathbf{z}) &= \ln p(\mathbf{x}) + \ln p(\mathbf{y} \mid \mathbf{x}) \\ &= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu}) \\ &\quad - \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^{\mathrm{T}} \mathbf{L} (\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}) + \text{const} \end{aligned} \tag{2.102}$$

其中“const”表示与 $\mathbf{x}$ 和 $\mathbf{y}$ 无关的项。同之前一样，我们看到这是 $\mathbf{z}$ 分量的二次型，因此 $p(\mathbf{z})$ 是高斯分布。为了求这个高斯分布的精度，我们考察（2.102）中的二次项，可写为

$$\begin{aligned} &-\tfrac{1}{2}\mathbf{x}^{\mathrm{T}}(\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})\mathbf{x}-\tfrac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y}+\tfrac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{A}\mathbf{x}+\tfrac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{y} \\ &=-\tfrac{1}{2}\binom{\mathbf{x}}{\mathbf{y}}^{\mathrm{T}}\begin{pmatrix}\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A} & -\mathbf{A}^{\mathrm{T}}\mathbf{L}\\ -\mathbf{L}\mathbf{A} & \mathbf{L}\end{pmatrix}\binom{\mathbf{x}}{\mathbf{y}}=-\tfrac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{R}\mathbf{z}\end{aligned} \tag{2.103}$$

因此 $\mathbf{z}$ 上的高斯分布的精度（逆协方差）矩阵为

$$\mathbf{R}=\begin{pmatrix}\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A} & -\mathbf{A}^{\mathrm{T}}\mathbf{L}\\ -\mathbf{L}\mathbf{A} & \mathbf{L}\end{pmatrix}. \tag{2.104}$$

协方差矩阵通过对精度矩阵求逆得到，利用矩阵求逆公式（2.76）可得

$$\operatorname{cov}[\mathbf{z}]=\mathbf{R}^{-1}=\begin{pmatrix}\boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\end{pmatrix}. \tag{2.105}$$

同样，我们可以通过识别（2.102）中的线性项来求 $\mathbf{z}$ 上高斯分布的均值，这些线性项为

$$\mathbf{x}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\mu}-\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{b}+\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{b}=\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu}-\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{b}\\ \mathbf{L}\mathbf{b}\end{pmatrix}. \tag{2.106}$$

利用我们之前通过对多元高斯二次型完成平方得到的结论（2.71），可得 $\mathbf{z}$ 的均值为

$$\mathbb{E}[\mathbf{z}]=\mathbf{R}^{-1}\begin{pmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu}-\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{b}\\ \mathbf{L}\mathbf{b}\end{pmatrix}. \tag{2.107}$$

再利用（2.105），得到

$$\mathbb{E}[\mathbf{z}]=\begin{pmatrix}\boldsymbol{\mu}\\ \mathbf{A}\boldsymbol{\mu}+\mathbf{b}\end{pmatrix}. \tag{2.108}$$

接下来我们求对 $\mathbf{x}$ 积分后得到的边缘分布 $p(\mathbf{y})$。回忆一下，当用分块协方差矩阵表示时，高斯随机向量部分分量的边缘分布具有特别简单的形式。具体地，其均值和协方差分别由（2.92）和（2.93）给出。利用（2.105）和（2.108），边缘分布 $p(\mathbf{y})$ 的均值和协方差为

$$\mathbb{E}[\mathbf{y}]=\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \tag{2.109}$$

$$\operatorname{cov}[\mathbf{y}]=\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}. \tag{2.110}$$

这一结果的一个特例是 $\mathbf{A}=\mathbf{I}$，此时它退化为两个高斯分布的卷积。我们看到，卷积的均值等于两个高斯均值之和，卷积的协方差等于两个协方差之和。

最后，我们求条件分布 $p(\mathbf{x}\mid\mathbf{y})$ 的表达式。回忆条件分布的结果用分块精度矩阵表示最为简便，即（2.73）和（2.75）。将这些结果应用于（2.105）和（2.108），得到条件分布 $p(\mathbf{x}\mid\mathbf{y})$ 的均值和协方差为

$$\mathbb{E}[\mathbf{x}\mid\mathbf{y}]=\left(\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A}\right)^{-1}\left\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \tag{2.111}$$

$$\operatorname{cov}[\mathbf{x}\mid\mathbf{y}]=\left(\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A}\right)^{-1}. \tag{2.112}$$

对该条件分布的求解可以视为贝叶斯定理的一个例子。我们可以将分布 $p(\mathbf{x})$ 解释为 $\mathbf{x}$ 上的先验分布。若变量 $\mathbf{y}$ 被观测到，则条件分布 $p(\mathbf{x} \mid \mathbf{y})$ 即为 $\mathbf{x}$ 上的相应后验分布。在求得边缘分布和条件分布后，我们实际上已将联合分布 $p(\mathbf{z}) = p(\mathbf{x})p(\mathbf{y} \mid \mathbf{x})$ 表示成了 $p(\mathbf{x} \mid \mathbf{y})p(\mathbf{y})$ 的形式。这些结果总结如下。

边缘与条件高斯分布

给定 $\mathbf{x}$ 的边缘高斯分布以及 $\mathbf{y}$ 在给定 $\mathbf{x}$ 下的条件高斯分布如下

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right) \tag{2.113}$$

$$p(\mathbf{y} \mid \mathbf{x}) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right) \tag{2.114}$$

则 $\mathbf{y}$ 的边缘分布以及给定 $\mathbf{y}$ 时 $\mathbf{x}$ 的条件分布分别为

$$p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\right) \tag{2.115}$$

$$p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\Sigma}\left\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}\right) \tag{2.116}$$

其中

$$\boldsymbol{\Sigma} = \left(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A}\right)^{-1} \tag{2.117}$$

2.3.4 高斯分布的最大似然估计

给定数据集 $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^{\mathrm{T}}$，其中观测值 $\{\mathbf{x}_n\}$ 被假定独立地从多元高斯分布中抽取，我们可以通过最大似然估计分布的参数。对数似然函数为

$$\ln p(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}). \tag{2.118}$$

通过简单重排可见，似然函数仅通过以下两个量依赖于数据集

$$\sum_{n=1}^{N}\mathbf{x}_n, \quad \sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}. \tag{2.119}$$

这两个量被称为高斯分布的充分统计量。利用（C.19），对数似然关于 $\boldsymbol{\mu}$ 的导数为

$$\frac{\partial}{\partial\boldsymbol{\mu}}\ln p(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}), \tag{2.120}$$

将该导数设为零，得到均值的最大似然估计解为

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \tag{2.121}$$

这是观测数据点的均值。对（2.118）关于 $\boldsymbol{\Sigma}$ 的最大化要复杂得多。最简单的方法是暂时忽略对称性约束，先求解，然后证明结果自然是对称的。关于显式施加对称性和正定

约束的其他推导，可参见 Magnus and Neudecker (1999)。结果符合预期，形式为

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}, \tag{2.122}$$

这里使用了 $\boldsymbol{\mu}_{\mathrm{ML}}$，是因为这是对 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 的联合最大化。注意，$\boldsymbol{\mu}_{\mathrm{ML}}$ 的解（2.121）不依赖于 $\boldsymbol{\Sigma}_{\mathrm{ML}}$，因此我们可以先计算 $\boldsymbol{\mu}_{\mathrm{ML}}$，再用它计算 $\boldsymbol{\Sigma}_{\mathrm{ML}}$。

如果我们在真实分布下对最大似然解取期望，得到

$$\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}] = \boldsymbol{\mu}, \tag{2.123}$$

$$\mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] = \frac{N-1}{N}\boldsymbol{\Sigma}. \tag{2.124}$$

可见，均值的最大似然估计的期望等于真实均值。然而，协方差的最大似然估计的期望小于真实值，因此是有偏的。我们可以通过定义一个不同的估计量来消除这种偏倚

$$\widetilde{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}. \tag{2.125}$$

显然，由（2.122）和（2.124）可知，$\widetilde{\boldsymbol{\Sigma}}$ 的期望等于 $\boldsymbol{\Sigma}$。

2.3.5 顺序估计

我们对高斯分布参数最大似然解的讨论，为我们提供了一个绝佳的机会，来更一般地讨论最大似然问题的顺序估计课题。顺序方法允许每次只处理一个数据点，处理完后即可丢弃，这在以下两种情形中尤为重要：在线应用，以及数据量巨大、一次性批量处理所有数据点不可行的情况。

考虑均值的最大似然估计式（2.121），当基于 N 个观测时我们记作 $\boldsymbol{\mu}_{\mathrm{ML}}^{(N)}$。如果把最后一个数据点 $\mathbf{x}_N$ 的贡献单独拆出来，可得

$$\begin{aligned}
\boldsymbol{\mu}_{\mathrm{ML}}^{(N)} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \\
&= \frac{1}{N}\mathbf{x}_N + \frac{1}{N}\sum_{n=1}^{N-1}\mathbf{x}_n \\
&= \frac{1}{N}\mathbf{x}_N + \frac{N-1}{N}\boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} \\
&= \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} + \frac{1}{N}\left(\mathbf{x}_N - \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)}\right).
\end{aligned} \tag{2.126}$$

这个结果有一个很直观的解释：观测了 $N-1$ 个数据点后，我们已用 $\boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)}$ 估计了 $\boldsymbol{\mu}$。现在又观测到新数据点 $\mathbf{x}_N$，我们就把旧估计值沿着“误差信号” $\left(\mathbf{x}_N - \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)}\right)$ 的方向移动一小步（步长为 $1/N$），得到新的估计 $\boldsymbol{\mu}_{\mathrm{ML}}^{(N)}$。注意，随着 N 增大，后续每个数据点的贡献越来越小。

式（2.126）显然与批量方法（2.121）给出完全相同的结果，因为两者等价。然而，并非所有问题都能通过这种方式直接导出顺序算法，因此我们需要一种更通用的顺序学习框架，这就引出了 Robbins–Monro (RM) 算法。考虑一对随机变量 θ 和 z，其联合分布为 $p(z,\theta)$。给定 θ 时 z 的条件期望定义了一个确定性函数 $f(\theta)$：

$$f(\theta) \equiv \mathbb{E}[z \mid \theta] = \int z\, p(z \mid \theta)\, \mathrm{d}z, \tag{2.127}$$

如图 2.10 所示。这样定义的函数称为回归函数。

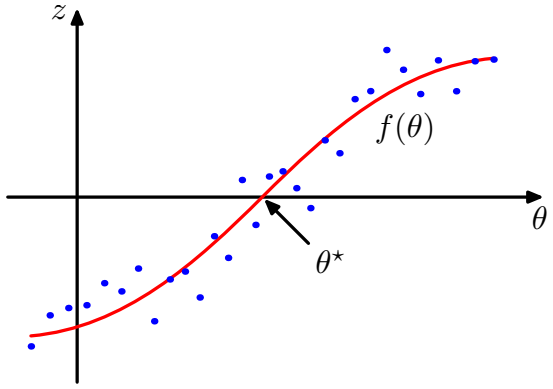

图 2.10 两个相关随机变量 z 和 θ 的示意图，同时给出了回归函数 $f(\theta)$，该函数由条件期望 $\mathrm{E}[z \mid \theta]$ 定义。Robbins–Monro 算法提供了一种通用的序列化方法，用于求解此类函数的根 θ^*。

我们的目标是寻找根 $\theta^\star$，使得 $f(\theta^\star) = 0$。如果拥有大量 z 和 θ 的观测数据，我们可以直接对回归函数建模，再求其根。然而，实际中我们往往是逐个观测到 z 的值，希望找到对应的顺序估计方案来逼近 $\theta^\star$。Robbins 和 Monro（1951）给出了求解这类问题的通用顺序算法。我们假定 z 的条件方差有限，即

$$\mathbb{E}\left[(z - f)^2 \mid \theta\right] < \infty, \tag{2.128}$$

并且不失一般性地考虑图 2.10 所示的情形：当 $\theta > \theta^\star$ 时 $f(\theta) > 0$，当 $\theta < \theta^\star$ 时 $f(\theta) < 0$。Robbins–Monro 算法定义根 $\theta^\star$ 的逐次估计序列为

$$\theta^{(N)} = \theta^{(N-1)} + a_{N-1} z\left(\theta^{(N-1)}\right), \tag{2.129}$$

其中 $z(\theta^{(N)})$ 是 θ 取值为 $\theta^{(N)}$ 时 z 的一个观测值。系数序列 $\{a_N\}$ 为正数，且满足

$$\lim_{N\to\infty} a_N = 0, \tag{2.130}$$

$$\sum_{N=1}^{\infty} a_N = \infty, \tag{2.131}$$

$$\sum_{N=1}^{\infty} a_N^2 < \infty. \tag{2.132}$$

可以证明（Robbins and Monro, 1951; Fukunaga, 1990），由（2.129）给出的估计序列几乎必然收敛到根 $\theta^\star$。其中，条件（2.130）保证逐次修正幅度逐渐减小，使算法能够收

敛到极限值；条件（2.131）保证算法不会在到达根之前停滞；条件（2.132）保证累积噪声具有有限方差，从而不破坏收敛性。

现在我们来看如何用 Robbins–Monro 算法顺序求解一般的最大似然问题。按定义，最大似然解 θ_{ML} 是对数似然函数的驻点，因此满足

$$\left.\frac{\partial}{\partial\theta}\left\{\frac{1}{N}\sum_{n=1}^{N}\ln p(\mathbf{x}_n\mid\theta)\right\}\right|_{\theta_{\mathrm{ML}}}=0. \tag{2.133}$$

交换求导与求和顺序并取极限 $N\to\infty$，得到

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}\ln p(x_n\mid\theta)=\mathbb{E}_x\left[\frac{\partial}{\partial\theta}\ln p(x\mid\theta)\right], \tag{2.134}$$

可见求最大似然解等价于求一个回归函数的根。因此我们可以直接应用 Robbins–Monro 算法，此时更新公式为

$$\theta^{(N)}=\theta^{(N-1)}+a_{N-1}\frac{\partial}{\partial\theta^{(N-1)}}\ln p\left(x_N\mid\theta^{(N-1)}\right). \tag{2.135}$$

作为一个具体例子，我们再次考虑高斯分布均值的顺序估计。此时参数 $\theta^{(N)}$ 即为高斯均值的估计 $\boldsymbol{\mu}_{\mathrm{ML}}^{(N)}$，随机变量 z 为

$$z=\frac{\partial}{\partial\boldsymbol{\mu}_{\mathrm{ML}}}\ln p\left(x\mid\boldsymbol{\mu}_{\mathrm{ML}},\sigma^2\right)=\frac{1}{\sigma^2}\left(x-\boldsymbol{\mu}_{\mathrm{ML}}\right). \tag{2.136}$$

因此 z 的分布是以 $\boldsymbol{\mu}-\boldsymbol{\mu}_{\mathrm{ML}}$ 为均值的高斯分布，如图 2.11 所示。将（2.136）代入（2.135），只要选择系数为 $a_N=\sigma^2/N$，即可得到（2.126）的单变量形式。需要指出，虽然我们讨论的是单变量情形，但同样的技术连同对系数 a_N 的相同限制（2.130）–（2.132）同样适用于多元情形（Blum, 1965）。

2.3.6 高斯分布的贝叶斯推断

最大似然框架为参数 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 提供了点估计。现在我们通过引入这些参数的先验分布来发展贝叶斯处理方法。先从一个简单例子开始：考虑单个高斯随机变量 x。假设方差 σ^2 已知，我们的任务是根据 N 个观测 $\mathbf{X}=\{x_1,\ldots,x_N\}$ 推断均值 $\boldsymbol{\mu}$。似然函数（即给定 $\boldsymbol{\mu}$ 时观测数据的概率，作为 $\boldsymbol{\mu}$ 的函数）为

$$p(\mathbf{X}\mid\boldsymbol{\mu})=\prod_{n=1}^{N}p(x_n\mid\boldsymbol{\mu})=\frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\boldsymbol{\mu})^2\right\}. \tag{2.137}$$

再次强调，似然函数 $p(\mathbf{X}\mid\boldsymbol{\mu})$ 并不是关于 $\boldsymbol{\mu}$ 的概率分布，也不归一化。

我们看到，似然函数具有指数形式的二次型。因此，如果选择高斯先验 $p(\mu)$，它将是该似然函数的共轭先验，因为相应的后验分布将是两个关于 μ 的二次函数指数的乘积，因而也是高斯分布。于是我们取先验分布为

$$p(\mu)=\mathcal{N}\left(\mu\mid\mu_0,\sigma_0^2\right), \tag{2.138}$$

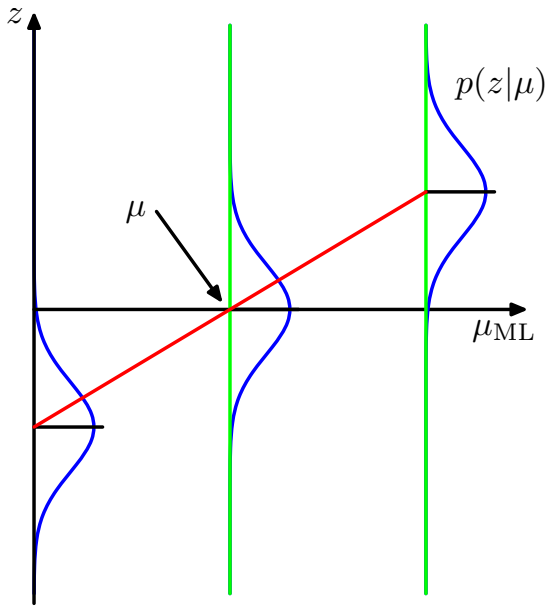

图 2.11 在高斯分布的情形下，令 θ 对应均值 μ，则图 2.10 中所示的回归函数呈现为一条直线（红色所示）。此时随机变量 z 对应于对数似然函数的导数，表示为 $(x-\mu_{\mathrm{ML}})/\sigma^2$，其期望（即定义回归函数的量）为 $(\mu-\mu_{\mathrm{ML}})/\sigma^2$，同样是一条直线。该回归函数的根对应最大似然估计量 μ_{ML}。

后验分布为

$$p(\mu \mid \mathbf{X}) \propto p(\mathbf{X} \mid \mu)p(\mu). \tag{2.139}$$

通过在指数中完成平方，简单推导可得后验分布为

$$p(\mu \mid \mathbf{X}) = \mathcal{N}\left(\mu \mid \mu_N, \sigma_N^2\right), \tag{2.140}$$

其中

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}}, \tag{2.141}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}, \tag{2.142}$$

这里 μ_{ML} 是样本均值给出的最大似然解

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n. \tag{2.143}$$

值得花一点时间研究后验均值和方差的形式。首先，后验均值（2.141）是先验均值 μ_0 与最大似然解 μ_{ML} 的折中。如果观测数据点数 $N=0$，则（2.141）退化为先验均值，这符合预期。当 $N \to \infty$ 时，后验均值变为最大似然解。

同样，考察后验方差的表达式（2.142）。我们发现用倒数方差（即精度）表示最为自然。而且精度是可加的：后验精度等于先验精度加上每个观测数据贡献的数据精度。随着观测数据点增多，精度持续增加，对应的后验分布方差持续减小。没有观测数据时，只有先验方差；当 $N \to \infty$ 时，方差 $\sigma_N^2 \to 0$，后验分布在最大似然解处变得无限尖锐。因此我们看到，当观测数据无限多时，贝叶斯方法精确恢复了最大似然给出的点估计结果（2.143）。

另外注意，对于有限 N，如果取极限 $\sigma_0^2 \to \infty$（先验方差无穷大，即无信息先验），则后验均值（2.141）退化为最大似然结果，而由（2.142）得后验方差为 $\sigma_N^2 = \sigma^2/N$。

我们用图 2.12 来说明对高斯均值的贝叶斯推断分析。将此结果推广到已知协方差、未知均值的 D 维高斯随机变量 $\mathbf{x}$ 是直观的。

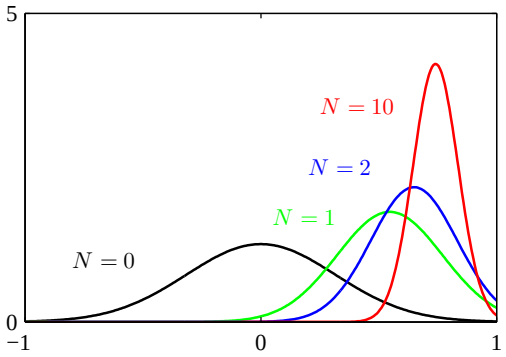

图 2.12 高斯分布均值 μ 的贝叶斯推断示意图，其中方差假定已知。图中曲线展示了对 μ 的先验分布（标记为 $N = 0$ 的曲线），该先验本身为高斯分布；同时还展示了随着数据点数量 N 增加，由式 (2.140) 给出的后验分布。数据点来自均值为 0.8、方差为 0.1 的高斯分布，先验均值取为 0。在先验和似然函数中，方差均设为真实值。

我们已经看到，高斯均值的最大似然表达式可以重写为顺序更新公式，其中观测 N 个数据点后的均值用观测 $N-1$ 个数据点后的均值加上第 N 个数据点 $\mathbf{x}_N$ 的贡献表示。事实上，贝叶斯范式天然导致对推断问题的顺序视角。在高斯均值推断的背景下，我们把最后一个数据点 $\mathbf{x}_N$ 的贡献分离出来，写出后验分布

$$p(\boldsymbol{\mu} \mid \mathbf{X}) \propto \left[p(\boldsymbol{\mu}) \prod_{n=1}^{N-1} p(\mathbf{x}_n \mid \boldsymbol{\mu}) \right] p(\mathbf{x}_N \mid \boldsymbol{\mu}). \tag{2.144}$$

方括号中的项（除归一化常数外）正是观测 $N-1$ 个数据点后的后验分布。我们看到，它可以被视为先验分布，通过贝叶斯定理与数据点 $\mathbf{x}_N$ 的似然函数结合，得到观测 N 个数据点后的后验分布。这种贝叶斯推断的顺序视角非常普适，适用于所有观测数据被假设为独立同分布的情形。

到目前为止，我们一直假设数据的高斯分布方差已知，目标是推断均值。现在假设均值已知，我们希望推断方差。同样，如果选择共轭先验形式，计算将大大简化。最方便的是对精度 $\lambda \equiv 1/\sigma^2$ 进行操作。精度 λ 的似然函数形式为

$$p(\mathbf{X} \mid \lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu, \lambda^{-1}) \propto \lambda^{N/2} \exp\left\{ -\frac{\lambda}{2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\}. \tag{2.145}$$

因此，共轭先验应与 λ 的幂次乘积以及 λ 的线性函数指数成正比。这对应于伽马分布（gamma distribution），其定义为

$$\mathrm{Gam}(\lambda \mid a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda). \tag{2.146}$$

这里 $\Gamma(a)$ 是伽马函数，由（1.141）定义，它保证（2.146）被正确归一化。当 $a > 0$ 时伽马分布积分有限，当 $a \geqslant 1$ 时分布本身有限。图 2.13 画出了不同 a 和 b 取值下的伽马分布。伽马分布的均值和方差为

$$\mathbb{E}[\lambda] = \frac{a}{b}, \tag{2.147}$$

$$\text{var}[\lambda] = \frac{a}{b^2}. \tag{2.148}$$

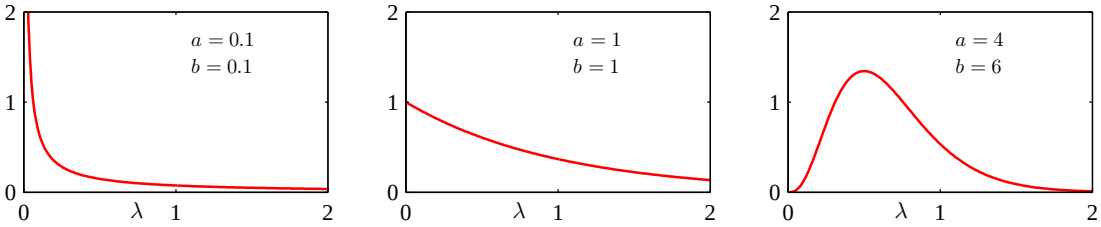

图 2.13 根据式（2.146）定义的伽马分布 $\text{Gam}(\lambda \mid a, b)$ 在不同参数 a 和 b 取值下的绘图。

考虑先验分布 $\text{Gam}(\lambda \mid a_0, b_0)$。将其与似然函数（2.145）相乘，得到后验分布

$$p(\lambda \mid \mathbf{X}) \propto \lambda^{a_0 - 1} \lambda^{N/2} \exp\left\{ -b_0 \lambda - \frac{\lambda}{2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\}, \tag{2.149}$$

这显然是形如 $\text{Gam}(\lambda \mid a_N, b_N)$ 的伽马分布，其中

$$a_N = a_0 + \frac{N}{2}, \tag{2.150}$$

$$b_N = b_0 + \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^2 = b_0 + \frac{N}{2} \sigma^2_{\text{ML}}, \tag{2.151}$$

这里 σ^2_{ML} 是方差的最大似然估计量。注意，在（2.149）中无需跟踪先验和似然函数中的归一化常数，因为必要时可最后利用伽马分布的归一化形式（2.146）补上正确的系数。

从（2.150）可见，观测 N 个数据点使参数 a 增加了 $N/2$。因此我们可以将先验中的 a_0 解释为对应于 $2a_0$ 个“有效”先验观测。类似地，由（2.151）可见，N 个数据点向参数 b 贡献了 $N\sigma^2_{\text{ML}}/2$，其中 σ^2_{ML} 是方差，因此先验中的 b_0 可解释为来自这 $2a_0$ 个“有效”先验观测，其方差为 $2b_0/(2a_0) = b_0/a_0$。回想我们在 Dirichlet 先验中也做过类似的解释。这些分布都属于指数族，我们将看到，对指数族分布而言，将共轭先验解释为有效虚构数据点是一种通用的做法。

我们也可以直接对方差本身操作，此时共轭先验称为逆伽马分布，但我们不再进一步讨论，因为对精度操作更方便。

现在假设均值和精度均未知。为找到共轭先验，我们考察似然函数对 μ 和 λ 的依

赖关系

$$\begin{aligned} p(\mathbf{X} \mid \mu, \lambda) &= \prod_{n=1}^{N} \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left\{-\frac{\lambda}{2}(x_n - \mu)^2\right\} \\ &\propto \left[\lambda^{1/2} \exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu \sum_{n=1}^{N} x_n - \frac{\lambda}{2} \sum_{n=1}^{N} x_n^2\right\}. \end{aligned} \tag{2.152}$$

我们希望找到一个先验 $p(\mu, \lambda)$，其对 μ 和 λ 的函数依赖形式与似然相同，因此应具有形式

$$\begin{aligned} p(\mu, \lambda) &\propto \left[\lambda^{1/2} \exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^{\beta} \exp\{c\lambda\mu - d\lambda\} \\ &= \exp\left\{-\frac{\beta\lambda}{2}\left(\mu - \frac{c}{\beta}\right)^2\right\} \lambda^{\beta/2} \exp\left\{-\left(d - \frac{c^2}{2\beta}\right)\lambda\right\}, \end{aligned} \tag{2.153}$$

其中 c、d 和 β 为常数。由于总可写成 $p(\mu, \lambda) = p(\mu \mid \lambda)p(\lambda)$，我们可直接看出：$p(\mu \mid \lambda)$ 是精度与 λ 成正比的高斯分布，$p(\lambda)$ 是伽马分布，因此归一化的先验为

$$p(\mu, \lambda) = \mathcal{N}\left(\mu \mid \mu_0, (\beta\lambda)^{-1}\right) \mathrm{Gam}(\lambda \mid a, b), \tag{2.154}$$

其中新常数定义为 $\mu_0 = c/\beta$，$a = 1 + \beta/2$，$b = d - c^2/(2\beta)$。分布（2.154）称为正态-伽马分布（normal-gamma）或高斯-伽马分布，如图 2.14 所示。注意这不是 μ 的高斯先验与 λ 的伽马先验的独立乘积，因为 μ 的精度与 λ 成正比。即使我们选择 μ 和 λ 独立的先验，后验分布中 μ 的精度与 λ 的取值仍会耦合。

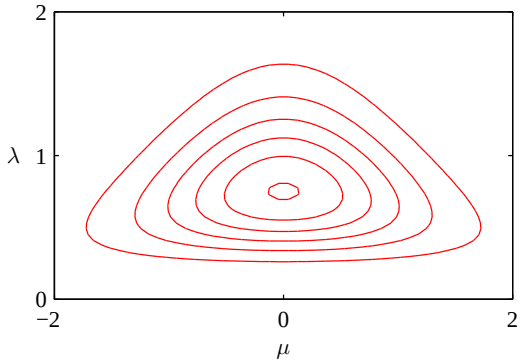

图 2.14 对参数取值 $\mu_0 = 0$, $\beta = 2$, $a = 5$, $b = 6$ 的正态-伽马分布 (??) 的等高线图。

对于 D 维变量 $\mathbf{x}$ 的多元高斯分布 $\mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right)$，当精度矩阵 $\boldsymbol{\Lambda}$ 已知时，均值 $\boldsymbol{\mu}$ 的共轭先验仍是高斯分布。当均值已知而精度矩阵 $\boldsymbol{\Lambda}$ 未知时，共轭先验为 *Wishart* 分布：

$$\mathcal{W}(\boldsymbol{\Lambda} \mid \mathbf{W}, \nu) = B|\boldsymbol{\Lambda}|^{(\nu - D - 1)/2} \exp\left(-\frac{1}{2}\mathrm{Tr}\left(\mathbf{W}^{-1}\boldsymbol{\Lambda}\right)\right), \tag{2.155}$$

其中 ν 称为自由度（degrees of freedom），$\mathbf{W}$ 是 $D \times D$ 的尺度矩阵（scale matrix），$\mathrm{Tr}(\cdot)$ 表示矩阵的迹。归一化常数 B 为

$$B(\mathbf{W}, \nu) = |\mathbf{W}|^{-\nu/2} \left(2^{\nu D/2} \pi^{D(D-1)/4} \prod_{i=1}^{D} \Gamma\left(\frac{\nu + 1 - i}{2}\right)\right)^{-1}. \tag{2.156}$$

同样，也可以对协方差矩阵本身定义共轭先验，得到逆 Wishart 分布，但此处不再详述。当均值 $\boldsymbol{\mu}$ 和精度矩阵 $\boldsymbol{\Lambda}$ 均未知时，仿照单变量情形的推导，共轭先验为

$$p\left(\boldsymbol{\mu}, \boldsymbol{\Lambda} \mid \boldsymbol{\mu}_0, \beta, \mathbf{W}, \nu\right)=\mathcal{N}\left(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0,(\beta \boldsymbol{\Lambda})^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda} \mid \mathbf{W}, \nu), \tag{2.157}$$

称为正态-Wishart 分布或高斯-Wishart 分布。

2.3.7 学生氏分布

我们已经看到，高斯分布精度的共轭先验是伽马分布。若有一个单变量高斯分布 $\mathcal{N}(x \mid \mu, \tau^{-1})$，并对精度 τ 赋予伽马先验 $\mathrm{Gam}(\tau \mid a, b)$，将精度 τ 边缘化（积分消掉）后，得到 x 的边缘分布为

$$\begin{aligned} p(x \mid \mu, a, b) &= \int_0^{\infty} \mathcal{N}(x \mid \mu, \tau^{-1}) \,\mathrm{Gam}(\tau \mid a, b) \,\mathrm{d}\tau \\ &= \int_0^{\infty} \frac{b^a e^{-b\tau} \tau^{a-1}}{\Gamma(a)} \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{\tau}{2}(x-\mu)^2\right\} \mathrm{d}\tau \\ &= \frac{b^a}{\Gamma(a)} \left(\frac{1}{2\pi}\right)^{1/2} \left[b + \frac{(x-\mu)^2}{2}\right]^{-a-1/2} \Gamma\left(a + \frac{1}{2}\right) \end{aligned} \tag{2.158}$$

其中我们做了变量替换 $z = \tau\left[b + \frac{(x-\mu)^2}{2}\right]$。按惯例引入新参数 $\nu = 2a$ 和 $\lambda = a/b$，于是 $p(x \mid \mu, a, b)$ 变为

$$\mathrm{St}(x \mid \mu, \lambda, \nu) = \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left(\frac{\lambda}{\pi\nu}\right)^{1/2} \left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]^{-\nu/2-1/2} \tag{2.159}$$

这就是 Student's t 分布。参数 λ 有时被称为 t 分布的“精度”，尽管它通常并不等于方差的倒数。参数 ν 称为自由度，其影响见图 2.15。当 $\nu = 1$ 时，t 分布退化为 Cauchy 分布；当 $\nu \to \infty$ 时，$\mathrm{St}(x \mid \mu, \lambda, \nu) \to \mathcal{N}(x \mid \mu, \lambda^{-1})$，即均值为 μ、精度为 λ 的高斯分布。

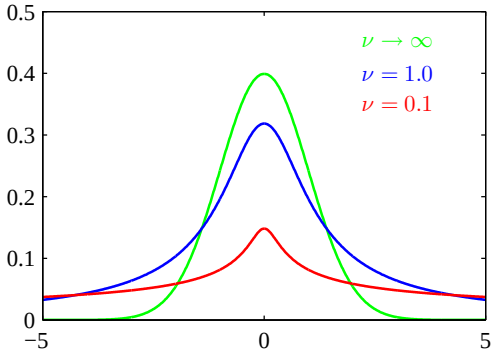

图 2.15 对不同的 ν 值绘制 Student's t 分布 (2.159)，其中 $\mu = 0$ 且 $\lambda = 1$。当 $\nu \to \infty$ 时，该分布趋近于均值为 μ、精度为 λ 的高斯分布。

从（2.158）可见，Student's t 分布相当于对无数个均值相同但精度不同的高斯分布求和。这可以看作高斯分布的无限混合（高斯混合将在 2.3.9 节详细讨论）。结果是一个比高斯分布具有更长“尾部”的分布，如图 2.15 所示。这种重尾赋予了 t 分布一个重要

性质——鲁棒性，即对少数离群点远不如高斯分布敏感。t 分布的鲁棒性在图 2.16 中与高斯分布的最大似然解进行了对比（t 分布的最大似然解可用 EM 算法求得）。可见少数离群点对 t 分布的影响远小于高斯分布。实际应用中离群点可能来自：数据生成过程本身具有重尾；数据标注错误。鲁棒性在回归问题中同样重要。最小二乘回归（对应条件高斯分布的最大似然）不具备鲁棒性。若将回归模型的噪声换成重尾的 t 分布，则得到更鲁棒的回归模型。

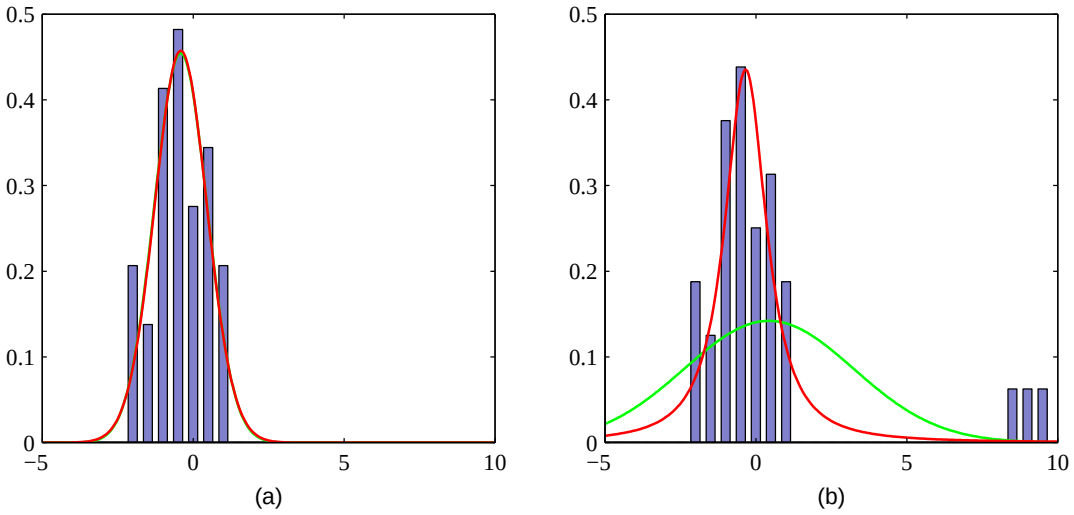

图 2.16 示例说明了与高斯分布相比，Student’s t 分布的鲁棒性。(a) 对从高斯分布抽取的 30 个数据点绘制直方图，并分别给出基于 t 分布（红色曲线）和高斯分布（绿色曲线，几乎被红色曲线遮住）的最大似然拟合。由于 t 分布将高斯分布作为特殊情形，因此其结果与高斯分布几乎相同。(b) 使用同一数据集，但额外加入三个离群点，可以看到高斯分布（绿色曲线）受离群点影响严重，而 t 分布（红色曲线）则相对不受影响。

回到（2.158），若代入参数 $\nu = 2a$、$\lambda = a/b$ 并令 $\eta = \tau b/a$，则 t 分布可重写为

$$\mathrm{St}(x \mid \mu, \lambda, \nu) = \int_0^\infty \mathcal{N}\left(x \mid \mu, (\eta\lambda)^{-1}\right) \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) \mathrm{d}\eta \tag{2.160}$$

将其推广到多元情形 $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda})$，即得多元 Student’s t 分布

$$\mathrm{St}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \int_0^\infty \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) \mathrm{d}\eta \tag{2.161}$$

采用与单变量情形相同的技巧，对该积分求值可得

$$\mathrm{St}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \frac{\Gamma\left(\frac{D}{2} + \frac{\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{|\boldsymbol{\Lambda}|^{1/2}}{(\pi\nu)^{D/2}} \left[1 + \frac{\Delta^2}{\nu}\right]^{-\frac{D}{2}-\frac{\nu}{2}} \tag{2.162}$$

其中 D 是 $\mathbf{x}$ 的维数，Δ^2 为平方马氏距离，定义为

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu}). \tag{2.163}$$

这就是多元 Student's t 分布，它具有以下性质

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} \quad (\text{当 } \nu > 1 \text{ 时存在}) \tag{2.164}$$

$$\text{cov}[\mathbf{x}] = \frac{\nu}{\nu - 2}\boldsymbol{\Lambda}^{-1} \quad (\text{当 } \nu > 2\,) \tag{2.165}$$

$$\text{mode}[\mathbf{x}] = \boldsymbol{\mu} \tag{2.166}$$

与单变量情况对应的结果类似。

2.3.8 周期变量

虽然高斯分布具有重要的实际意义，既可独立使用，也可作为构建更复杂概率模型的基本模块，但在某些情况下，它们并不适合作为连续变量的密度模型。一个在实际应用中重要的情形就是周期变量。

周期变量的一个例子是在特定地理位置的风向。例如，我们可能在多天测量风向值，并希望用参数分布来总结这些数据。另一个例子是日历时间，我们可能对那些被认为在 24 小时或年度周期内呈现周期性的量进行建模。这类量可以很方便地用角度（极）坐标 $0 \leqslant \theta < 2\pi$ 表示。

我们可能会倾向于通过选择某个方向作为原点，然后应用常规分布（如高斯分布）来处理周期变量。然而，这种方法会导致结果强烈依赖于原点的任意选择。例如，假设我们有两个观测值 $\theta_1 = 1^\circ$ 和 $\theta_2 = 359^\circ$，并使用标准的单变量高斯分布建模。如果选择原点在 0°，则该数据集的样本均值为 180°，标准差为 179°，而如果选择原点在 180°，则均值为 0°，标准差为 1°。显然，我们需要开发一种特殊的方法来处理周期变量。

让我们考虑对一组周期变量观测值 $\mathcal{D} = \{\theta_1, \ldots, \theta_N\}$ 求均值的问题。从现在起，我们假设 θ 以弧度为单位。我们已经看到，简单平均 $(\theta_1 + \cdots + \theta_N)/N$ 将强烈依赖于坐标选择。为了找到一个不变的均值度量，我们注意到这些观测值可以看作单位圆上的点，因此可以用二维单位向量 $\mathbf{x}_1, \ldots, \mathbf{x}_N$ 来描述，其中 $\|\mathbf{x}_n\| = 1$（$n = 1, \ldots, N$），如图 2.17 所示。我们可以对向量 $\{\mathbf{x}_n\}$ 取平均，得到

$$\overline{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \tag{2.167}$$

然后求出这个平均向量的对应角度 $\bar{\theta}$。显然，这种定义将确保均值的位置与角度坐标的原点选择无关。注意，$\overline{\mathbf{x}}$ 通常位于单位圆内部。观测值的笛卡尔坐标为 $\mathbf{x}_n = (\cos\theta_n, \sin\theta_n)$，我们可以将样本均值的笛卡尔坐标写成 $\overline{\mathbf{x}} = (\bar{r}\cos\bar{\theta}, \bar{r}\sin\bar{\theta})$。代入 (2.167) 并分别对 x_1 和 x_2 分量取等，得到

$$\bar{r}\cos\bar{\theta} = \frac{1}{N}\sum_{n=1}^{N}\cos\theta_n, \quad \bar{r}\sin\bar{\theta} = \frac{1}{N}\sum_{n=1}^{N}\sin\theta_n \tag{2.168}$$

取比值，并利用恒等式 $\tan\theta = \sin\theta/\cos\theta$，我们可以解出 $\bar{\theta}$，得到

$$\bar{\theta} = \tan^{-1}\left\{\frac{\sum_n \sin\theta_n}{\sum_n \cos\theta_n}\right\} \tag{2.169}$$

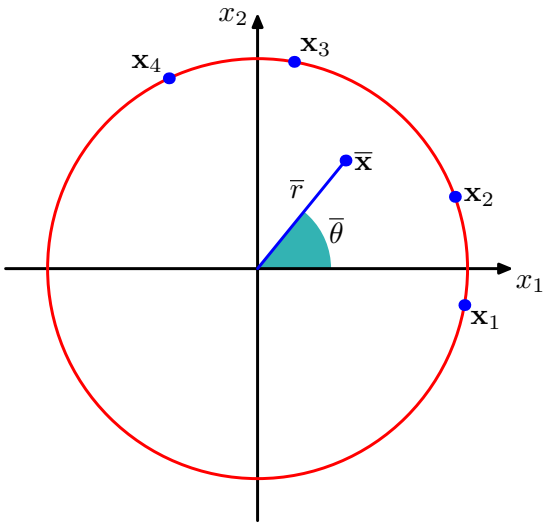

图 2.17 将周期变量的取值 θ_n 表示为位于单位圆上的二维向量 $\mathbf{x}_n$ 的示意图，同时给出了这些向量的平均值 $\overline{\mathbf{x}}$。

稍后，我们将看到这一结果自然地作为适当定义的周期变量分布的最大似然估计量出现。

现在我们考虑高斯分布的周期推广形式——冯 • 米塞斯。这里我们仅关注单变量分布，尽管周期分布也可以在任意维度的超球面上定义。有关周期分布的详细讨论，请参见 Mardia 和 Jupp (2000)。

按照惯例，我们考虑周期为 2π 的分布 $p(\theta)$。任何定义在 θ 上的概率密度 $p(\theta)$ 不仅必须非负且积分为 1，还必须具有周期性。因此 $p(\theta)$ 必须满足三个条件：

$$p(\theta) \geqslant 0 \tag{2.170}$$

$$\int_0^{2\pi} p(\theta)\,\mathrm{d}\theta = 1 \tag{2.171}$$

$$p(\theta + 2\pi) = p(\theta) \tag{2.172}$$

由 (2.172) 可知，对于任意整数 M，有 $p(\theta + M2\pi) = p(\theta)$。

我们可以很容易地得到一个满足这三个性质的类高斯分布，方法如下。考虑二维变量 $\mathbf{x} = (x_1, x_2)$ 上的高斯分布，其均值为 $\boldsymbol{\mu} = (\mu_1, \mu_2)$，协方差矩阵为 $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$，其中 $\mathbf{I}$ 是 2×2 单位矩阵，于是

$$p\left(x_1, x_2\right) = \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{\left(x_1 - \mu_1\right)^2 + \left(x_2 - \mu_2\right)^2}{2\sigma^2}\right\} \tag{2.173}$$

常数 $p(\mathbf{x})$ 的等高线是圆，如图 2.18 所示。现在假设我们考虑该分布在固定半径圆上的取值。由于构造方式，这种分布将是周期的，尽管尚未归一化。我们可以通过从笛卡尔坐标 (x_1, x_2) 变换到极坐标 (r, θ) 来确定该分布的形式，即

$$x_1 = r\cos\theta, \quad x_2 = r\sin\theta \tag{2.174}$$

我们还将均值 $\boldsymbol{\mu}$ 映射到极坐标，写作

$$\mu_1 = r_0\cos\theta_0, \quad \mu_2 = r_0\sin\theta_0 \tag{2.175}$$

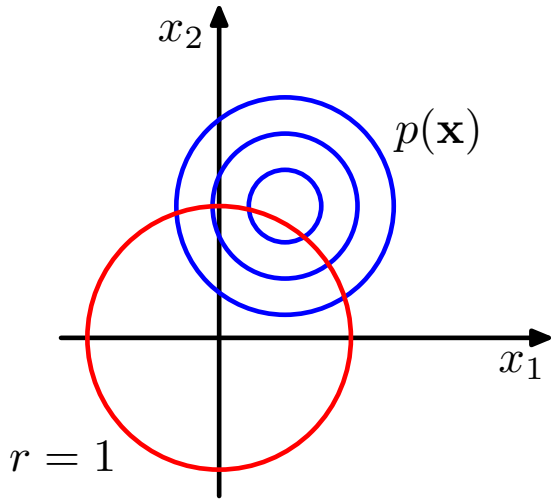

图 2.18 von Mises 分布可以通过考虑形如 (2.173) 的二维高斯分布推导而来，其密度等高线如蓝色所示，并对位于红色单位圆上的点进行条件化。

接下来将这些变换代入二维高斯分布 (2.173)，然后在单位圆 $r = 1$ 上条件化，注意我们只关心对 θ 的依赖。聚焦于高斯分布的指数部分，我们有

$$\begin{aligned} &-\frac{1}{2\sigma^2}\left\{(r\cos\theta - r_0\cos\theta_0)^2 + (r\sin\theta - r_0\sin\theta_0)^2\right\} \\ &= -\frac{1}{2\sigma^2}\left\{1 + r_0^2 - 2r_0\cos\theta\cos\theta_0 - 2r_0\sin\theta\sin\theta_0\right\} \\ &= \frac{r_0}{\sigma^2}\cos(\theta - \theta_0) + \text{const} \end{aligned} \tag{2.176}$$

其中“const”表示不依赖于 θ 的项，我们利用了以下三角恒等式

$$\cos^2 A + \sin^2 A = 1 \tag{2.177}$$

$$\cos A\cos B + \sin A\sin B = \cos(A - B) \tag{2.178}$$

如果现在定义 $m = r_0/\sigma^2$，我们就得到沿单位圆 $r = 1$ 的分布 $p(\theta)$ 的最终表达式

$$p(\theta \mid \theta_0, m) = \frac{1}{2\pi I_0(m)}\exp\{m\cos(\theta - \theta_0)\} \tag{2.179}$$

这被称为冯 • 米塞斯分布，或称圆形正态分布。这里参数 θ_0 对应分布的均值，而 m 被称为集中参数，它类似于高斯分布的逆方差（精度）。(2.179) 中的归一化系数用 $I_0(m)$ 表示，这是第一类零阶贝塞尔函数（Abramowitz and Stegun, 1965），定义为

$$I_0(m) = \frac{1}{2\pi}\int_0^{2\pi}\exp\{m\cos\theta\}\,\mathrm{d}\theta \tag{2.180}$$

当 m 很大时，该分布近似成为高斯分布。冯 • 米塞斯分布如图 2.19 所示，函数 $I_0(m)$ 如图 2.20 所示。

现在考虑冯 • 米塞斯分布参数 θ_0 和 m 的最大似然估计量。对数似然函数为

$$\ln p(\mathcal{D} \mid \theta_0, m) = -N\ln(2\pi) - N\ln I_0(m) + m\sum_{n=1}^{N}\cos(\theta_n - \theta_0) \tag{2.181}$$

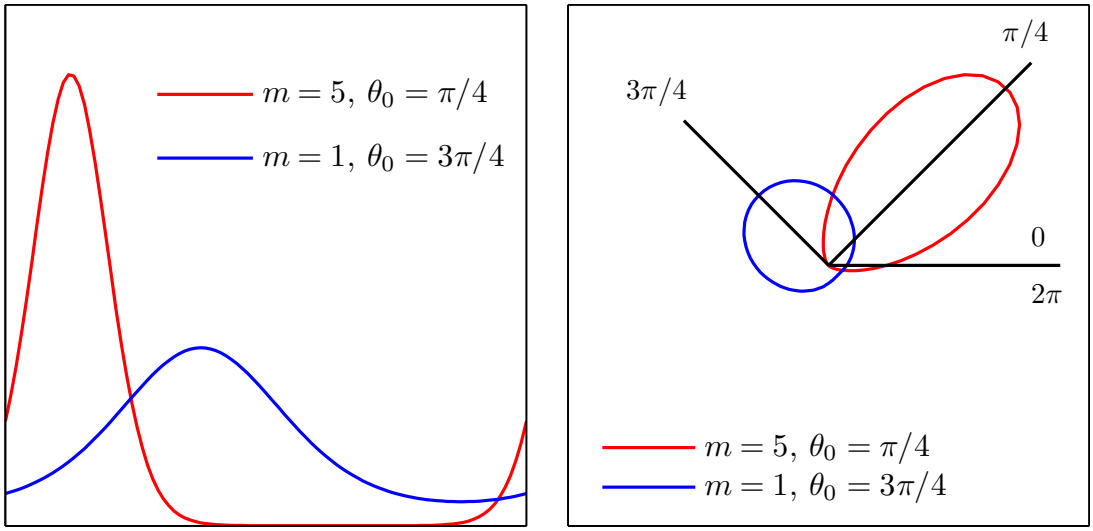

图 2.19 对两组不同参数取值绘制的 von Mises 分布，左图为笛卡尔坐标表示，右图为对应的极坐标表示。

将对 θ_0 的导数设为零，得到

$$\sum_{n=1}^{N} \sin\left(\theta_n - \theta_0\right) = 0 \tag{2.182}$$

为了求解 θ_0，我们利用三角恒等式

$$\sin(A - B) = \cos B \sin A - \cos A \sin B \tag{2.183}$$

由此得到

$$\theta_0^{\mathrm{ML}} = \tan^{-1}\left\{\frac{\sum_n \sin\theta_n}{\sum_n \cos\theta_n}\right\} \tag{2.184}$$

这正是我们先前在二维笛卡尔空间中观察均值时得到的 (2.169)。

类似地，对 (2.181) 关于 m 求最大值，并利用 $I_0'(m) = I_1(m)$（Abramowitz and Stegun, 1965），得到

$$A(m) = \frac{1}{N}\sum_{n=1}^{N} \cos\left(\theta_n - \theta_0^{\mathrm{ML}}\right) \tag{2.185}$$

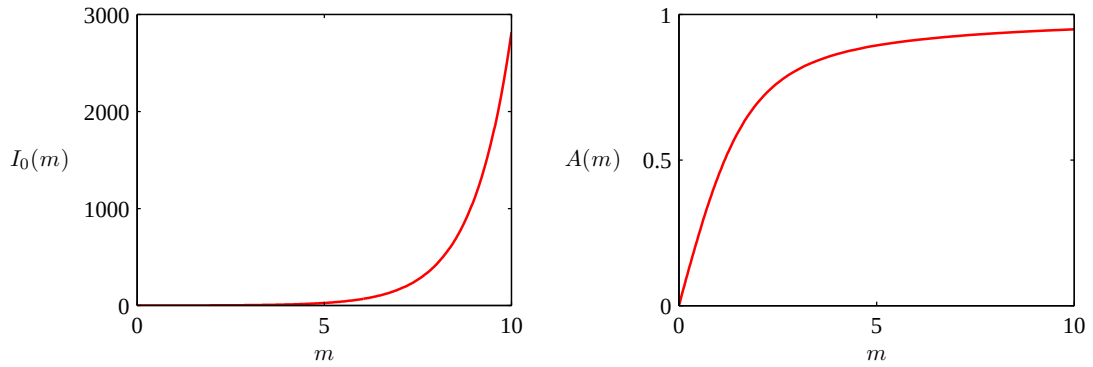

图 2.20 绘制由式 (2.180) 定义的贝塞尔函数 $I_0(m)$，以及由式 (2.186) 定义的函数 $A(m)$。

其中已代入 θ_0^{ML} 的最大似然解（记住我们是对 θ_0 和 m 进行联合优化），并定义

$$A(m) = \frac{I_1(m)}{I_0(m)} \tag{2.186}$$

函数 $A(m)$ 如图 2.20 所示。利用三角恒等式 (2.178)，我们可以将 (2.185) 写成

$$A\left(m_{\mathrm{ML}}\right) = \left(\frac{1}{N}\sum_{n=1}^{N}\cos\theta_n\right)\cos\theta_0^{\mathrm{ML}} - \left(\frac{1}{N}\sum_{n=1}^{N}\sin\theta_n\right)\sin\theta_0^{\mathrm{ML}} \tag{2.187}$$

(2.187) 的右端很容易计算，而函数 A(m) 可以数值求反。

为了完整性，我们简要提及几种构造周期分布的替代方法。最简单的方法是使用观测值的直方图，将角度坐标划分为固定区间。这种方法简单灵活，但也存在显著局限性，我们将在 2.5 节详细讨论直方图方法时看到。另一种方法与冯 • 米塞斯分布类似，从欧氏空间上的高斯分布出发，但现在是对单位圆进行边缘化而非条件化（Mardia and Jupp, 2000）。然而，这会导致更复杂的形式，我们不再进一步讨论。最后，任何在实轴上的合法分布（如高斯分布）都可以通过将宽度为 2π 的连续区间映射到周期变量 (0,2π) 上而转为周期分布，这相当于将实轴“缠绕”在单位圆上。同样，所得分布比冯 • 米塞斯分布更难处理。

冯 • 米塞斯分布的一个局限性是它是单峰的。通过构造冯 • 米塞斯分布的混合，我们得到一个灵活的周期变量建模框架，能够处理多峰情况。关于使用冯 • 米塞斯分布的机器学习应用示例，见 Lawrence et al. (2002)；关于回归问题中条件密度建模的扩展，见 Bishop 和 Nabney (1996)。

2.3.9 高斯混合

虽然高斯分布具有重要的解析性质，但在建模真实数据集时存在显著局限性。考虑图 2.21 所示的例子。这就是著名的“老忠实间歇泉”（Old Faithful）数据集，包含美国黄石国家公园老忠实间歇泉的 272 次喷发测量记录。每条记录包括喷发持续时间（分钟，横轴）和下一次喷发间隔时间（分钟，纵轴）。我们看到数据集形成了两个主要簇，单一高斯分布无法捕捉这种结构，而两个高斯的线性叠加则能更好地刻画数据集。

这种通过对更基本分布（如高斯）取线性组合形成的叠加，可以表述为称为混合分布的概率模型（McLachlan and Basford, 1988; McLachlan and Peel, 2000）。在图 2.22 中我们看到，高斯的线性组合可以产生非常复杂的密度。通过使用足够多的高斯，并调整它们的均值、协方差以及线性组合中的系数，几乎任意连续密度都可以以任意精度逼近。

因此，我们考虑 K 个高斯密度的叠加，形式为

$$p(\mathbf{x}) = \sum_{k=1}^{K}\pi_k\mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \tag{2.188}$$

这被称为高斯混合。每一个高斯密度 $\mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$ 都被称为混合的一个分量，它拥有自己的均值 $\boldsymbol{\mu}_k$ 和协方差 $\boldsymbol{\Sigma}_k$。具有 3 个分量的高斯混合的等高线图和表面图如图 2.23 所示。

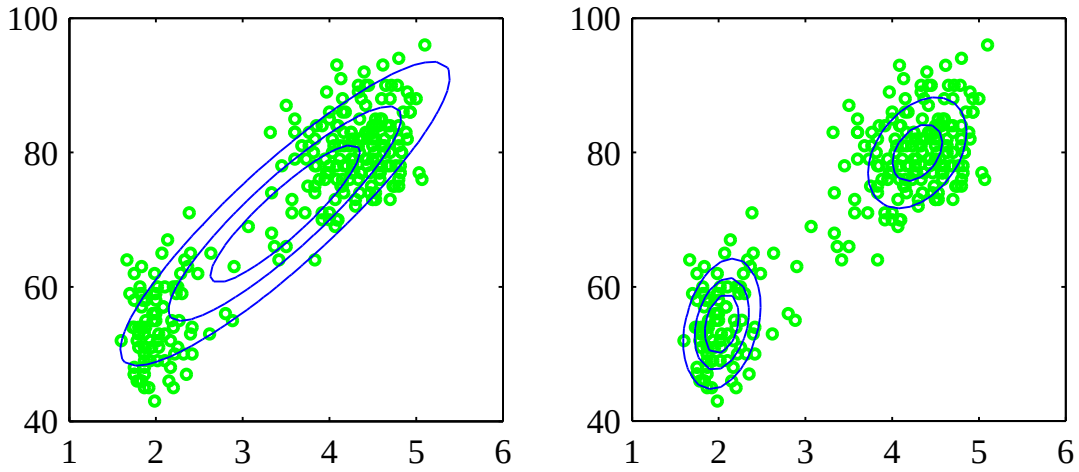

图 2.21 对“老忠实间歇泉”数据的绘图，其中蓝色曲线表示等概率密度轮廓。左图为使用最大似然拟合得到的单个高斯分布。注意该分布无法捕捉数据中的两个聚集区域，并且将大部分概率质量放在这两个聚集之间的中间区域，而该区域的数据相对稀疏。右图为由两个高斯的线性组合构成的分布，利用第 9 章讨论的方法通过最大似然进行拟合，能够更好地刻画数据结构。

在本节中，我们以高斯分量为例来说明混合模型的框架。更一般地，混合模型可以由其他分布的线性组合构成。例如，在 9.3.3 节我们将讨论伯努利分布的混合，作为离散变量混合模型的一个例子。

(2.188) 中的参数 π_k 被称为混合系数。如果我们对 (2.188) 两边关于 $\mathbf{x}$ 积分，并注意到 $p(\mathbf{x})$ 和各个高斯分量都是归一化的，便得到

$$\sum_{k=1}^{K} \pi_k = 1 \tag{2.189}$$

此外，$p(\mathbf{x}) \geqslant 0$ 的要求，连同 $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \geqslant 0$ 意味着 $\pi_k \geqslant 0$（对所有 k 成立）。结合条件 (2.189)，我们得到

$$0 \leqslant \pi_k \leqslant 1 \tag{2.190}$$

因此我们看到，混合系数满足作为概率的要求。

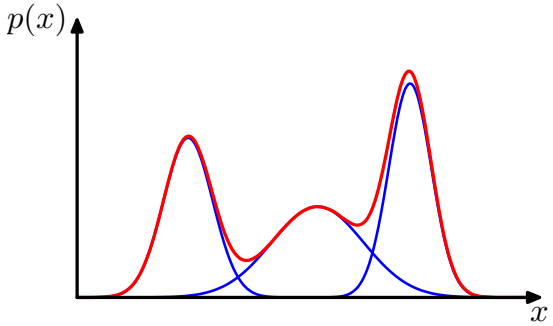

图 2.22 一维高斯混合分布示例，其中三个经过系数缩放的高斯分布以蓝色显示，它们的和以红色显示。

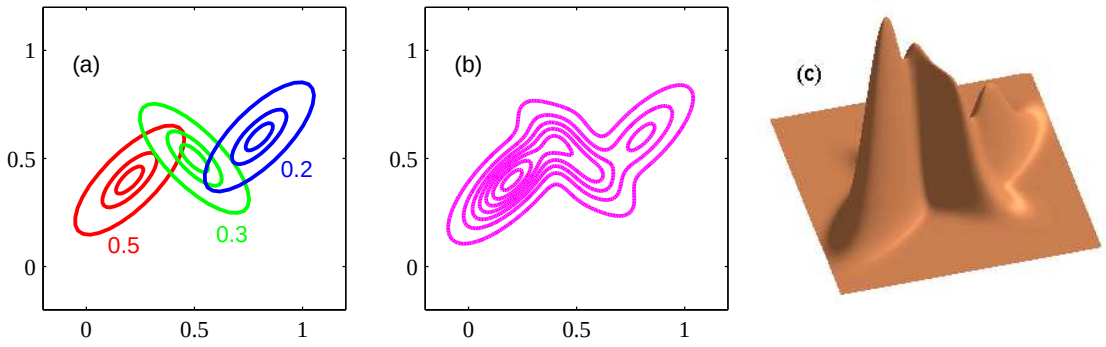

图 2.23 一维高斯混合分布示例，其中三个经过系数缩放的高斯分布以蓝色显示，它们的和以红色显示。

根据求和法则与乘积法则，边际密度为

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k)\, p(\mathbf{x} \mid k) \tag{2.191}$$

这等价于 (2.188)，其中我们可以将 $\pi_k = p(k)$ 视为选择第 k 个分量的先验概率，而 $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = p(\mathbf{x} \mid k)$ 视为给定 k 时 $\mathbf{x}$ 的条件概率。正如我们在后面章节将看到的，后验概率 $p(k \mid \mathbf{x})$（也称为 responsibilities）扮演着重要角色。根据贝叶斯定理，它们为

$$\begin{aligned} \gamma_k(\mathbf{x}) &\equiv p(k \mid \mathbf{x}) \\ &= \frac{p(k)\, p(\mathbf{x} \mid k)}{\sum_l p(l)\, p(\mathbf{x} \mid l)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_l \pi_l \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \end{aligned} \tag{2.192}$$

我们将在第 9 章更详细地讨论混合分布的概率解释。

高斯混合分布的形式由参数 $\boldsymbol{\pi}$、$\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 决定，这里采用记号 $\boldsymbol{\pi} \equiv \{\pi_1, \ldots, \pi_K\}$、$\boldsymbol{\mu} \equiv \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$ 以及 $\boldsymbol{\Sigma} \equiv \{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}$。确定这些参数值的一种方法是使用最大似然。从 (2.188) 可得对数似然函数为

$$\ln p(\mathbf{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \tag{2.193}$$

其中 $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$。我们立即看到，由于对数内部出现了对 k 的求和，现在的情况比单一高斯复杂得多。因此，参数的最大似然解不再具有闭合解析形式。最大化似然函数的一种方法是使用迭代数值优化技术（Fletcher, 1987; Nocedal and Wright, 1999; Bishop and Nabney, 2008）。另一种方法是采用一种称为期望最大的强大框架，我们将在第 9 章详细讨论。

2.4 指数家族

本章迄今研究的概率分布（高斯混合除外）都是一个广义分布类——指数家族的具体例子（Duda and Hart, 1973; Bernardo and Smith, 1994）。指数家族的成员具有许多共同的重要性质，用一定的通用性讨论这些性质是很有启发性的。

给定参数 $\boldsymbol{\eta}$，关于 $\mathbf{x}$ 的指数家族分布定义为如下形式的分布集合：

$$p(\mathbf{x} \mid \boldsymbol{\eta}) = h(\mathbf{x}) g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\right\} \tag{2.194}$$

其中 $\mathbf{x}$ 可以是标量或向量，可以是离散的或连续的。这里的 $\boldsymbol{\eta}$ 被称为分布的自然参数，$\mathbf{u}(\mathbf{x})$ 是 $\mathbf{x}$ 的某个函数。函数 $g(\boldsymbol{\eta})$ 可以解释为确保分布归一化的系数，因此满足

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\right\} \mathrm{d}\mathbf{x} = 1 \tag{2.195}$$

若 $\mathbf{x}$ 是离散变量，则积分改为求和。

我们先从本章前面介绍的一些分布例子入手，说明它们确实属于指数家族。首先考虑伯努利分布

$$p(x \mid \mu) = \mathrm{Bern}(x \mid \mu) = \mu^x (1-\mu)^{1-x} \tag{2.196}$$

将右端表示为对数的指数形式，得到

$$\begin{aligned} p(x \mid \mu) &= \exp\{x \ln \mu + (1-x) \ln(1-\mu)\} \\ &= (1-\mu) \exp\left\{\ln\left(\frac{\mu}{1-\mu}\right) x\right\} \end{aligned} \tag{2.197}$$

与 (2.194) 对比，可识别出

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right) \tag{2.198}$$

我们可以解出 μ 得到 $\mu = \sigma(\eta)$，其中

$$\sigma(\eta) = \frac{1}{1+\exp(-\eta)} \tag{2.199}$$

被称为 logistic sigmoid 函数。于是我们可以用标准形式 (2.194) 将伯努利分布写为

$$p(x \mid \eta) = \sigma(-\eta) \exp(\eta x) \tag{2.200}$$

这里利用了 $1 - \sigma(\eta) = \sigma(-\eta)$，这由 (2.199) 易证。与 (2.194) 对比可得

$$u(x) = x \tag{2.201}$$

$$h(x) = 1 \tag{2.202}$$

$$g(\eta) = \sigma(-\eta) \tag{2.203}$$

接下来考虑多项分布，对于单个观测 $\mathbf{x}$ 它具有形式

$$p(\mathbf{x} \mid \boldsymbol{\mu}) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} \tag{2.204}$$

其中 $\mathbf{x} = (x_1, \ldots, x_M)^{\mathrm{T}}$。同样，我们可以将它写成标准形式 (2.194)：

$$p(\mathbf{x} \mid \boldsymbol{\eta}) = \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \mathbf{x}\right) \tag{2.205}$$

其中 $\eta_k = \ln \mu_k$，并定义 $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_M)^{\mathrm{T}}$。再与 (2.194) 对比得到

$$\mathbf{u}(\mathbf{x}) = \mathbf{x} \tag{2.206}$$

$$h(\mathbf{x}) = 1 \tag{2.207}$$

$$g(\boldsymbol{\eta}) = 1 \tag{2.208}$$

需要注意的是，参数 η_k 并非相互独立，因为参数 μ_k 受约束

$$\sum_{k=1}^{M} \mu_k = 1 \tag{2.209}$$

因此，给定任意 $M-1$ 个 μ_k，剩余一个参数的值就被确定了。在某些情况下，为了消除这一约束，方便地用仅 $M-1$ 个参数表达分布是可取的。这可以通过关系式 (2.209) 将 μ_M 表示为其余 $\{\mu_k\}$（$k = 1, \ldots, M-1$）的函数来实现，从而只剩下 $M-1$ 个参数。注意这些剩余参数仍受约束

$$0 \leqslant \mu_k \leqslant 1, \quad \sum_{k=1}^{M-1} \mu_k \leqslant 1 \tag{2.210}$$

利用约束 (2.209)，多项分布在此表示下变为

$$\begin{aligned}
&\exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} \\
&= \exp\left\{\sum_{k=1}^{M-1} x_k \ln \mu_k + \left(1 - \sum_{k=1}^{M-1} x_k\right) \ln\left(1 - \sum_{k=1}^{M-1} \mu_k\right)\right\} \\
&= \exp\left\{\sum_{k=1}^{M-1} x_k \ln\left(\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j}\right) + \ln\left(1 - \sum_{k=1}^{M-1} \mu_k\right)\right\}
\end{aligned} \tag{2.211}$$

现在令

$$\ln\left(\frac{\mu_k}{1 - \sum_j \mu_j}\right) = \eta_k \tag{2.212}$$

对 k 求和后整理并回代，得到

$$\mu_k = \frac{\exp(\eta_k)}{1 + \sum_j \exp(\eta_j)} \tag{2.213}$$

这被称为 softmax 函数，或归一化指数函数。在此表示下，多项分布因此具有形式

$$p(\mathbf{x} \mid \boldsymbol{\eta}) = \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k)\right)^{-1} \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \mathbf{x}\right) \tag{2.214}$$

这是指数家族的标准形式，参数向量为 $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{M-1})^{\mathrm{T}}$，其中

$$\mathbf{u}(\mathbf{x}) = \mathbf{x} \tag{2.215}$$

$$h(\mathbf{x}) = 1 \tag{2.216}$$

$$g(\boldsymbol{\eta}) = \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k)\right)^{-1} \tag{2.217}$$

最后，我们考虑高斯分布。对于单变量高斯分布，有

$$p\left(x \mid \mu, \sigma^2\right) = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{2.218}$$

$$= \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\right\} \tag{2.219}$$

经过简单整理后，可将其改写为指数家族标准形式 (2.194)，其中

$$\boldsymbol{\eta} = \begin{pmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{pmatrix} \tag{2.220}$$

$$\mathbf{u}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \tag{2.221}$$

$$h(x) = (2\pi)^{-1/2} \tag{2.222}$$

$$g(\boldsymbol{\eta}) = (-2\eta_2)^{1/2} \exp\left(\frac{\eta_1^2}{4\eta_2}\right) \tag{2.223}$$

2.4.1 最大似然与充分统计量

现在我们考虑使用最大似然方法估计指数家族分布 (2.194) 中的参数向量 $\boldsymbol{\eta}$。对 (2.195) 两边关于 $\boldsymbol{\eta}$ 求梯度，得到

$$\begin{aligned} \nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathrm{d}\mathbf{x} \\ + g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x}) \mathrm{d}\mathbf{x} = 0 \end{aligned} \tag{2.224}$$

整理并再次利用 (2.195)，得到

$$-\frac{1}{g(\boldsymbol{\eta})}\nabla g(\boldsymbol{\eta}) = g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x}) \mathrm{d}\mathbf{x} = \mathbb{E}[\mathbf{u}(\mathbf{x})] \tag{2.225}$$

这里用到了 (2.194)。因此得到结果

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})] \tag{2.226}$$

需要指出，$\mathbf{u}(\mathbf{x})$ 的协方差可以用 $g(\boldsymbol{\eta})$ 的二阶导数表示，高阶矩同理。因此，只要能对指数家族分布进行归一化，我们总可以通过简单求导得到它的各阶矩。

现在考虑一组独立同分布的数据 $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$，其似然函数为

$$p(\mathbf{X} \mid \boldsymbol{\eta}) = \left(\prod_{n=1}^{N} h(\mathbf{x}_n)\right) g(\boldsymbol{\eta})^N \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)\right\} \tag{2.227}$$

将 $\ln p(\mathbf{X} \mid \boldsymbol{\eta})$ 关于 $\boldsymbol{\eta}$ 的梯度设为零，得到最大似然估计 $\boldsymbol{\eta}_{\mathrm{ML}}$ 必须满足的条件

$$-\nabla \ln g(\boldsymbol{\eta}_{\mathrm{ML}}) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n) \tag{2.228}$$

原则上可通过求解得到 $\boldsymbol{\eta}_{\mathrm{ML}}$。我们看到，最大似然估计的解仅通过 $\sum_n \mathbf{u}(\mathbf{x}_n)$ 依赖于数据，因此称其为分布 (2.194) 的充分统计量。我们无需存储整个数据集，只需保留充分统计量的值即可。例如，对于伯努利分布，$\mathbf{u}(x) = x$，因此只需记录数据点 $\{x_n\}$ 的和；而对于高斯分布，$\mathbf{u}(x) = (x, x^2)^{\mathrm{T}}$，因此需要保留 $\{x_n\}$ 的和以及 $\{x_n^2\}$ 的和。

如果考虑极限 $N \to \infty$，则 (2.228) 右端变为 $\mathbb{E}[\mathbf{u}(\mathbf{x})]$，与 (2.226) 对比可知，在此极限下 $\boldsymbol{\eta}_{\mathrm{ML}}$ 将等于真实值 $\boldsymbol{\eta}$。

事实上，这种充分性性质在贝叶斯推断中同样成立，不过我们将推迟到第 8 章讨论，那时我们已具备图模型工具，从而能更深入地理解这些重要概念。

2.4.2 共轭先验

我们已多次遇到共轭先验的概念，例如伯努利分布（其共轭先验为 beta 分布）或高斯分布（均值的共轭先验为高斯分布，精度的共轭先验为 Wishart 分布）。一般地，对于给定的概率分布 $p(\mathbf{x} \mid \boldsymbol{\eta})$，我们可以寻找一个与似然函数共轭的先验 $p(\boldsymbol{\eta})$，使得后验分布与先验具有相同的函数形式。对于指数家族 (2.194) 的任意成员，都存在可写成如下形式的共轭先验：

$$p(\boldsymbol{\eta} \mid \boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu)\, g(\boldsymbol{\eta})^{\nu} \exp\left\{\nu \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi}\right\} \tag{2.229}$$

其中 $f(\boldsymbol{\chi}, \nu)$ 为归一化系数，$g(\boldsymbol{\eta})$ 与 (2.194) 中出现的函数相同。为了验证其共轭性，将先验 (2.229) 与似然函数 (2.227) 相乘，得到后验分布（除归一化系数外）形式为

$$p(\boldsymbol{\eta} \mid \mathbf{X}, \boldsymbol{\chi}, \nu) \propto g(\boldsymbol{\eta})^{\nu+N} \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \left(\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n) + \nu \boldsymbol{\chi}\right)\right\} \tag{2.230}$$

这再次具有与先验 (2.229) 相同的函数形式，证实了共轭性。此外，我们看到参数 ν 可解释为先验中有效伪观测的数量，每个伪观测的充分统计量 $\mathbf{u}(\mathbf{x})$ 值为 $\boldsymbol{\chi}$。

2.4.3 无信息先验

在概率推断的某些应用中，我们可能拥有可通过先验分布方便表达的先验知识。例如，若先验对变量的某个值赋予零概率，则无论后续观测到什么数据，后验分布也必然对该值赋予零概率。然而在许多情况下，我们对分布形式几乎一无所知。这时我们会寻找一种称为无信息先验的先验分布，旨在对后验分布的影响尽可能小（Jeffries, 1946; Box and Tao, 1973; Bernardo and Smith, 1994）。这有时被称为“让数据自己发声”。

如果分布 $p(x \mid \lambda)$ 由参数 λ 控制，我们可能会倾向于提出先验分布 $p(\lambda) = \text{const}$ 作为合适的先验。如果 λ 是具有 K 个状态的离散变量，这只相当于将每个状态的先验概率设为 $1/K$。然而对于连续参数，这种做法存在两个潜在困难。

第一个困难是，如果 λ 的定义域无界，这个先验分布就无法正确归一化，因为对 λ 的积分发散。这类先验被称为非正常先验。在实践中，只要相应的后验分布是正常的（即可正确归一化），通常仍可使用非正常先验。例如，若对高斯分布的均值施加均匀先验，只要观测到至少一个数据点，均值的后验分布就是正常的。

第二个困难源于概率密度在非线性变量变换下的变换行为，即 (1.27)。如果函数 $h(\lambda)$ 为常数，而我们换元为 $\lambda = \eta^2$，则 $\widehat{h}(\eta) = h(\eta^2)$ 也为常数。然而，若我们选择密度 $p_\lambda(\lambda)$ 为常数，则根据 (1.27)，η 的密度为

$$p_\eta(\eta) = p_\lambda(\lambda) \left| \frac{\mathrm{d}\lambda}{\mathrm{d}\eta} \right| = p_\lambda(\eta^2) \cdot 2\eta \propto \eta \tag{2.231}$$

因此 η 上的密度不再是常数。使用最大似然时不会出现这个问题，因为似然函数 $p(x \mid \lambda)$ 是 λ 的简单函数，我们可以自由选用任何方便的参数化。然而，如果要选择一个常数先验，就必须谨慎选用参数的适当表示。

这里我们考虑两个无信息先验的简单例子（Berger, 1985）。首先，如果密度具有形式

$$p(x \mid \mu) = f(x - \mu) \tag{2.232}$$

则参数 μ 被称为位置参数（location parameter）。这一密度族具有平移不变性，因为若将 x 平移常数得到 $\widehat{x} = x + c$，则

$$p(\widehat{x} \mid \widehat{\mu}) = f(\widehat{x} - \widehat{\mu}) \tag{2.233}$$

其中定义了 $\widehat{\mu} = \mu + c$。于是密度在新变量下具有与原变量相同的形式，因此密度与原点的选择无关。我们希望选择一个能反映这种平移不变性的先验分布，即对区间 $A \leqslant \mu \leqslant B$ 赋予与平移后的区间 $A - c \leqslant \mu \leqslant B - c$ 相同的概率质量。这意味着

$$\int_A^B p(\mu)\,\mathrm{d}\mu = \int_{A-c}^{B-c} p(\mu)\,\mathrm{d}\mu = \int_A^B p(\mu - c)\,\mathrm{d}\mu \tag{2.234}$$

由于这对任意 A 和 B 都必须成立，故有

$$p(\mu - c) = p(\mu) \tag{2.235}$$

这意味着 $p(\mu)$ 为常数。高斯分布均值 μ 就是一个位置参数的例子。如前所述，此情况下 μ 的共轭先验为高斯分布 $p(\mu \mid \mu_0, \sigma_0^2) = \mathcal{N}(\mu \mid \mu_0, \sigma_0^2)$，取极限 $\sigma_0^2 \to \infty$ 即可得到无信息先验。事实上，由 (2.141) 和 (2.142) 可知，这将使后验分布中先验的贡献消失。

第二个例子，考虑形如

$$p(x \mid \sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right) \tag{2.236}$$

的密度，其中 $\sigma > 0$。只要 $f(x)$ 正确归一化，这就是一个归一化密度。参数 σ 称为尺度参数，该密度具有尺度不变性，因为若将 x 缩放常数得到 $\widehat{x} = cx$，则

$$p(\widehat{x} \mid \widehat{\sigma}) = \frac{1}{\widehat{\sigma}} f\left(\frac{\widehat{x}}{\widehat{\sigma}}\right) \tag{2.237}$$

其中定义了 $\widehat{\sigma} = c\sigma$。这种变换对应尺度改变，例如从米变为千米（若 x 是长度），我们希望选择能反映这种尺度不变性的先验分布。若考虑区间 $A \leqslant \sigma \leqslant B$ 与缩放后的区间 $A/c \leqslant \sigma \leqslant B/c$，先验应对这两个区间赋予相等的概率质量。于是

$$\int_A^B p(\sigma)\,\mathrm{d}\sigma = \int_{A/c}^{B/c} p(\sigma)\,\mathrm{d}\sigma = \int_A^B p\left(\frac{\sigma}{c}\right) \frac{1}{c}\,\mathrm{d}\sigma \tag{2.238}$$

由于这对任意 A 和 B 都必须成立，故有

$$p(\sigma) = p\left(\frac{\sigma}{c}\right)\frac{1}{c} \tag{2.239}$$

从而 $p(\sigma) \propto 1/\sigma$。这又是一个非正常先验，因为在 $0 \leqslant \sigma \leqslant \infty$ 上积分发散。有时用参数对数的密度来考虑尺度参数的先验分布也很方便。利用密度变换规则 (1.27)，得到 $p(\ln\sigma) = \text{const}$。因此，对于该先验，区间 $1 \leqslant \sigma \leqslant 10$、$10 \leqslant \sigma \leqslant 100$ 以及 $100 \leqslant \sigma \leqslant 1000$ 具有相同的概率质量。

高斯分布的标准差 σ 就是一个尺度参数的例子（在已考虑位置参数 μ 之后），因为

$$\mathcal{N}(x \mid \mu, \sigma^2) \propto \sigma^{-1}\exp\left\{-(\tilde{x}/\sigma)^2\right\} \tag{2.240}$$

其中 $\tilde{x} = x - \mu$。如前所述，通常更方便用精度 $\lambda = 1/\sigma^2$ 而非 σ 本身工作。利用密度变换规则，可见 $p(\sigma) \propto 1/\sigma$ 对应于精度 λ 上的分布 $p(\lambda) \propto 1/\lambda$。我们已知 λ 的共轭先验是 gamma 分布 $\text{Gam}(\lambda \mid a_0, b_0)$，见 (2.146)。无信息先验对应特例 $a_0 = b_0 = 0$。同样，若考察后验分布 (2.150) 与 (2.151)，当 $a_0 = b_0 = 0$ 时，后验仅依赖于数据项，而与先验无关。

2.5 非参数方法

在本章中，我们一直关注具有特定函数形式、仅由少量参数控制的概率分布，这些参数的值需从数据集中确定。这被称为密度建模的参数方法。该方法的一个重要局限在于，所选密度可能无法很好地刻画生成数据的真实分布，从而导致较差的预测性能。例如，若生成数据的过程是多峰的，则高斯分布（必然单峰）永远无法捕捉这一特性。

在本节最后部分，我们将讨论几种非参数密度估计方法，它们对分布形式几乎不作假设。这里主要聚焦于简单的频数主义方法。但需注意，非参数贝叶斯方法正日益受到关注（Walker et al., 1999; Neal, 2000; Müller and Quintana, 2004; Teh et al., 2006）。

我们从直方图密度估计方法开始讨论，这在我们先前讨论边缘分布与条件分布（图 1.11）以及中心极限定理（图 2.6）时已遇到。这里将更详细地探讨直方图密度模型的性质，聚焦于单个连续变量 x 的情形。

标准直方图将 x 简单划分为若干互不重叠的区间，每个区间宽度为 Δ_i，然后统计落入第 i 个区间的观测数 n_i。要将其转化为归一化概率密度，只需将计数除以总观测数 N 再除以区间宽度 Δ_i，得到每个区间的概率值为

$$p_i = \frac{n_i}{N\Delta_i} \tag{2.241}$$

易见 $\int p(x)\,\mathrm{d}x = 1$。这给出了一个在每个区间内恒定的密度模型 $p(x)$，通常各区间宽度相等，即 $\Delta_i = \Delta$。

在图 2.24 中，我们展示了一个直方图密度估计的例子。这里的数据是从对应于绿色曲线的分布中抽取的，该分布由两个高斯分布的混合构成。图中还展示了三种直方图密度估计的结果，对应于三种不同的区间宽度 Δ。我们看到，当 Δ 非常小时（最上方的图），得到的密度模型非常尖锐，出现了许多生成数据集的底层分布中并不存在的结

构。相反，如果 Δ 过大（最下方的图），则得到一个过于平滑的模型，从而无法捕捉绿色曲线的双峰特性。最佳结果是在某个中间值的 Δ 处获得的（中间的图）。原则上，直方图密度模型还依赖于区间边界的选择，不过这通常远不如 Δ 的值重要。

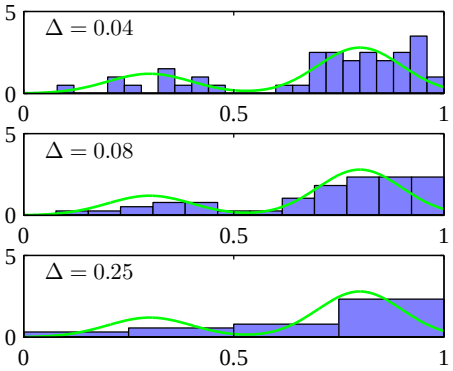

图 2.24 密度估计中直方图方法的示意图，其中 50 个数据点由绿色曲线所示的分布生成。基于式 (2.241) 并采用统一分箱宽度 Δ 的直方图密度估计在不同 Δ 取值下给出。

需要注意的是，直方图方法具有这样一个性质（与接下来将要讨论的方法不同）：一旦直方图计算完成，数据集本身就可以被丢弃，这在数据集很大时非常有利。此外，如果数据点是依次到达的，直方图方法也非常容易应用。

在实践中，直方图技术可以用于在一维或二维中快速可视化数据，但不适合大多数密度估计应用。一个明显的问题是估计得到的密度在区间边界处存在不连续性，而这些不连续性来源于区间的划分，而不是生成数据的底层分布的任何特性。直方图方法的另一个主要局限在于其随维数增加的扩展性。如果我们在 D 维空间中将每个变量划分为 M 个区间，那么总的区间数目将是 M^D。这种随 D 指数级增长的现象就是维数灾难的一个例子。在高维空间中，要提供有意义的局部概率密度估计所需的数据量将非常庞大。

然而，直方图密度估计方法确实给我们带来了两个重要的经验教训。首先，要估计某个特定位置的概率密度，我们应该考虑落在该点某个局部邻域内的数据点。注意，局部的概念要求我们假设某种距离度量，这里我们一直假设的是欧氏距离。对于直方图，这个邻域性质由区间定义，并且存在一个自然的“平滑”参数来描述局部区域的空间范围，在本例中就是区间宽度。其次，平滑参数的值既不能太大也不能太小，才能得到好的结果。这让人联想到第一章中讨论的多项式曲线拟合中的模型复杂度选择，其中多项式的次数 M，或者正则化参数 α 的值，在某个既不太大也不太小的中间值时达到最优。带着这些洞察，我们现在转向两种广泛使用的非参数密度估计技术——核估计器和最近邻方法，这两种方法在维数增加时的扩展性都优于简单的直方图模型。

2.5.1 核密度估计器

假设在某个 D 维欧氏空间中，观测样本是从某个未知概率密度 $p(\mathbf{x})$ 中独立抽取的，我们希望估计 $p(\mathbf{x})$ 的值。根据之前的局部性讨论，考虑包含 $\mathbf{x}$ 的某个小区间 $\mathcal{R}$。该区

域包含的概率质量为

$$P = \int_{\mathcal{R}} p(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{2.242}$$

现在假设我们已经收集了一个包含 N 个样本的数据集，这些样本来自 $p(\mathbf{x})$。由于每个数据点落入 $\mathcal{R}$ 的概率为 P，因此落在 $\mathcal{R}$ 内部的点数 K 服从二项分布

$$\mathrm{Bin}(K \mid N, P) = \frac{N!}{K!(N-K)!} P^K (1-P)^{N-K} \tag{2.243}$$

根据（2.11），落入该区域的点的平均比例为 $\mathbb{E}[K/N] = P$；同样根据（2.12），该比例围绕均值的方差为 $\mathrm{var}[K/N] = P(1-P)/N$。当 N 很大时，该分布将在均值附近急剧集中，因此

$$K \simeq NP \tag{2.244}$$

另一方面，如果我们还假设区域 $\mathcal{R}$ 足够小，使得概率密度 $p(\mathbf{x})$ 在该区域上近似为常数，则有

$$P \simeq p(\mathbf{x})V \tag{2.245}$$

其中 V 是 $\mathcal{R}$ 的体积。将（2.244）和（2.245）合并，可得到密度估计的形式

$$p(\mathbf{x}) = \frac{K}{NV} \tag{2.246}$$

需要注意的是，（2.246）的有效性依赖于两个相互矛盾的假设：一方面区域 $\mathcal{R}$ 要足够小，以保证密度在区域内近似恒定；另一方面区域又要足够大（相对于密度的大小），以使得落入其中的点数 K 足够多，从而二项分布能够急剧集中在均值附近。

我们可以以两种不同的方式利用结果 (2.246)。要么固定 K 并从数据中确定 V 的值，这就产生了即将讨论的 K 近邻技术；要么固定 V 并从数据中确定 K，这就产生了核方法。可以证明，只要 V 随着 N 适当地缩小，而 K 随着 N 增大，则 K 近邻密度估计器和核密度估计器在 $N \to \infty$ 极限下都会收敛到真实的概率密度（Duda and Hart, 1973）。

我们首先详细讨论核方法，一开始我们取区域 $\mathcal{R}$ 为以希望确定概率密度的点 $\mathbf{x}$ 为中心的一个小超立方体。为了统计落入该区域内的点数 K，定义如下函数很方便

$$k(\mathbf{u}) = \begin{cases} 1, & |u_i| \leqslant 1/2, \quad i = 1, \ldots, D \\ 0, & \text{otherwise} \end{cases} \tag{2.247}$$

这个函数表示一个以原点为中心、边长为 1 的单位立方体。函数 $k(\mathbf{u})$ 是一种核函数，在此背景下也称为 Parzen 窗。根据 (2.247)，若数据点 $\mathbf{x}_n$ 位于以 $\mathbf{x}$ 为中心、边长为 h 的立方体内部，则 $k\left((\mathbf{x} - \mathbf{x}_n)/h\right)$ 取值为 1，否则为 0。因此，位于该立方体内部的数据点总数为

$$K = \sum_{n=1}^{N} k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right). \tag{2.248}$$

将此表达式代入 (2.246)，即可得到点 $\mathbf{x}$ 处的估计密度

$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{h^D} k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right) \tag{2.249}$$

其中我们用了 D 维空间中边长为 h 的超立方体的体积 $V = h^D$。利用函数 $k(\mathbf{u})$ 的对称性，我们现在可以将这个方程重新解释为不是以 $\mathbf{x}$ 为中心的一个立方体，而是对以 N 个数据点 $\mathbf{x}_n$ 为中心的 N 个立方体求和。

就目前的形式而言，核密度估计器 (2.249) 会遇到与直方图方法相同的缺陷之一，即在立方体边界处存在人为的不连续性。如果我们选择更平滑的核函数，就可以得到更平滑的密度模型，一个常见的选择是高斯核，从而得到如下核密度模型

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{(2\pi h^2)^{1/2}}\exp\left\{-\frac{\|\mathbf{x}-\mathbf{x}_n\|^2}{2h^2}\right\} \tag{2.250}$$

其中 h 表示高斯分量的标准差。因此，我们的密度模型是通过在每个数据点上放置一个高斯分布，然后对整个数据集的贡献求和，再除以 N 以确保密度正确归一化。在图 2.25 中，我们将模型 (2.250) 应用于之前用来展示直方图技术的数据集。可以看出，正如预期的那样，参数 h 起到了平滑参数的作用：在 h 较小时对噪声过于敏感，而在 h 较大时又会出现过度平滑的现象。因此，h 的优化是一个模型复杂度的权衡问题，类似于直方图密度估计中 bin 宽度的选择，或曲线拟合中多项式阶数的选择。

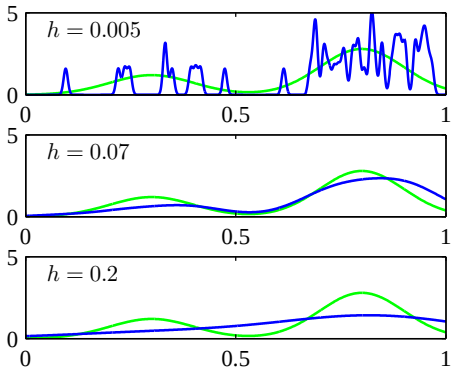

图 2.25 将核密度模型 (2.250) 应用于与图 2.24 中直方图方法相同的数据集。可以看到，h 起平滑参数作用：若 h 取值过小（上图），得到的密度模型噪声很大；若 h 取值过大（下图），则数据生成的真实分布（绿色曲线所示）本来的双峰结构被抹平。某个中间取值的 h（中图）则给出了最佳密度模型。

我们也可以在 (2.249) 中选择任意其他核函数 $k(\mathbf{u})$，只要它满足以下条件：

$$k(\mathbf{u}) \geqslant 0, \tag{2.251}$$

$$\int k(\mathbf{u})\mathrm{d}\mathbf{u} = 1 \tag{2.252}$$

这些条件保证了得到的概率分布处处非负且积分为 1。由 (2.249) 给出的这一类密度模型被称为核密度估计器（kernel density estimator），或 Parzen 估计器。它有一个很大的优点：在“训练”阶段完全不需要计算，因为这仅仅需要存储训练集。然而，这也是它的一个重大缺陷，因为密度估计的计算代价随着数据集大小线性增长。

2.5.2 最近邻方法

核方法在密度估计中的一个困难在于，控制核宽度的参数 h 对所有核都是固定的。在数据密度高的区域，较大的 h 可能会导致过度平滑，从而抹平原本可以从数据中提取的结构。然而，减小 h 又可能在数据空间中密度较低的其他区域导致噪声估计。因此，h 的最优选择可能依赖于数据空间中的位置。最近邻密度估计方法正是为了解决这一问题而提出的。

我们因此回到局部密度估计的通用结果 (2.246)，不再固定体积 V 并从数据中确定 K 的值，而是固定 K 的值，并利用数据找到合适的 V。具体做法是：在希望估计密度 $p(\mathbf{x})$ 的点 $\mathbf{x}$ 处放置一个小的球体，让球体的半径逐渐增大，直到正好包含 K 个数据点。此时密度 $p(\mathbf{x})$ 的估计由 (2.246) 给出，其中 V 取为该球体的体积。这种技术被称为 K 最近邻（K nearest neighbours），如图 2.26 所示，我们对不同的 K 值使用了与图 2.24 和图 2.25 相同的数据集。可以看出，K 的取值现在控制了平滑程度，同样存在一个最优的 K 值，既不能太大也不能太小。需要注意的是，K 最近邻方法产生的模型并不是真正的概率密度模型，因为其在整个空间上的积分是发散的。

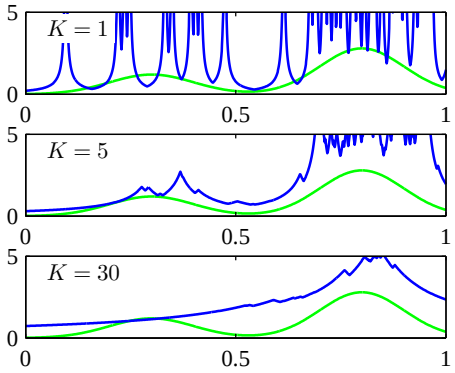

图 2.26 使用与图 2.25 和图 2.24 相同的数据集进行 K 近邻密度估计的示意图。可以看到，参数 K 决定平滑程度：K 较小时（上图），密度模型噪声很大；K 较大时（下图），数据生成的真实分布（绿色曲线所示）的双峰结构被过度平滑。

我们通过展示如何将 K 最近邻密度估计技术扩展到分类问题来结束本章。为此，我们分别对每个类别单独应用 K 最近邻密度估计，然后利用贝叶斯定理。假设我们有一个数据集，其中类别 $\mathcal{C}_k$ 包含 N_k 个点，总共 N 个点，因此 $\sum_k N_k = N$。如果我们要对一个新点 $\mathbf{x}$ 进行分类，就以 $\mathbf{x}$ 为中心画一个球体，使其恰好包含 K 个点（不论类别）。假设该球体的体积为 V，其中来自类别 $\mathcal{C}_k$ 的点有 K_k 个。那么 (2.246) 给出了各个类别的密度估计

$$p\left(\mathbf{x} \mid \mathcal{C}_k\right) = \frac{K_k}{N_k V} \tag{2.253}$$

同样，无条件密度为

$$p(\mathbf{x}) = \frac{K}{NV} \tag{2.254}$$

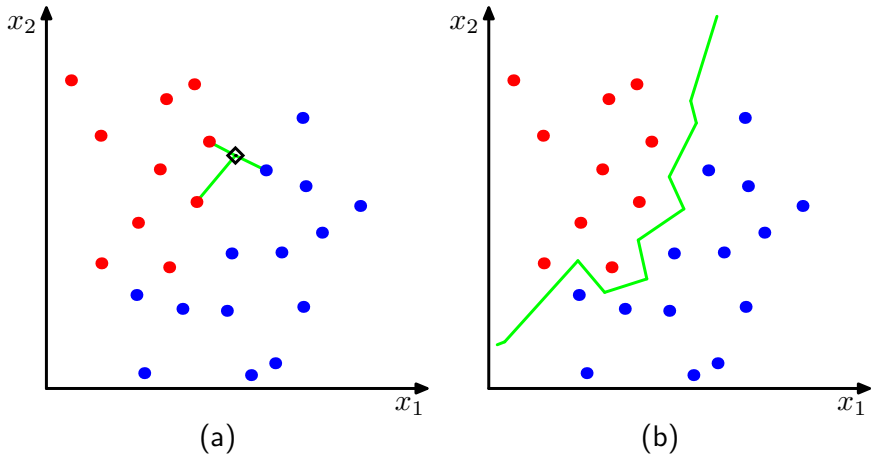

图 2.27 (a) 在 K 近邻分类器中，新数据点（黑色菱形所示）的类别由其最近的 K 个训练样本中占多数的类别决定，此处 $K = 3$。(b) 在最近邻（$K = 1$）分类方法中，得到的决策边界由一系列超平面组成，这些超平面是来自不同类别的成对样本之间的垂直平分面。

而类别先验概率为

$$p(\mathcal{C}_k) = \frac{N_k}{N} \tag{2.255}$$

现在我们利用贝叶斯定理合并 (2.253)、(2.254) 和 (2.255)，得到类别后验概率

$$p(\mathcal{C}_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathcal{C}_k)\, p(\mathcal{C}_k)}{p(\mathbf{x})} = \frac{K_k}{K} \tag{2.256}$$

如果我们希望最小化误分类概率，只需将测试点 $\mathbf{x}$ 分配给后验概率最大的类别，也就是 K_k/K 最大的那个类别。因此，对一个新点进行分类时，我们从训练集中找出 K 个最近的点，然后将新点分配给这 K 个点中代表数量最多的类别。如果出现平局，可随机打破平局。$K = 1$ 的特殊情况称为最近邻规则，此时测试点直接被分配给训练集中最近点的同一类别。这些概念在图 2.27 中得到了说明。

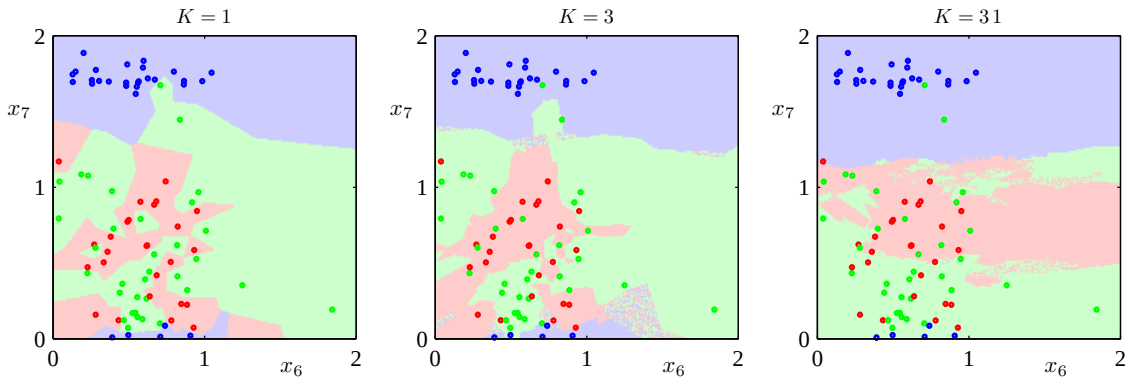

图 2.28 从油流数据集中抽取的 200 个数据点，其特征 x_6 与 x_7 的散点图如上所示。其中红色、绿色与蓝色点分别对应“层流（laminar）”、“环状流（annular）”和“均匀流（homogeneous）”三类。图中还展示了 K 近邻算法在不同 K 取值下对输入空间的分类结果。

在图 2.28 中，我们对第 1 章介绍的油流数据应用了 K 最近邻算法，展示了不同 K 值的结果。正如预期，K 控制了平滑程度：K 较小时会产生许多小的类别区域，而 K

较大时则导致更少但更大的区域。

最近邻（K=1）分类器的一个有趣性质是，当 $N \to \infty$ 时，其错误率永远不会超过最优分类器（即使用真实类别分布的分类器）最小可达错误率的两倍（Cover and Hart, 1967）。

到目前为止讨论的 K 最近邻方法和核密度估计器都需要存储整个训练数据集，当数据集很大时会导致计算代价高昂。通过构造基于树的搜索结构，可以在一次性额外计算的代价下，高效地找到（近似的）最近邻，而无需对数据集进行穷尽搜索，从而在一定程度上缓解这一问题。尽管如此，这些非参数方法仍然受到严重限制。另一方面，我们已经看到，简单的参数模型在所能表示的分布形式上非常受限。因此，我们需要找到非常灵活的密度模型，同时模型的复杂度能够独立于训练集大小进行控制，在后续章节中我们将看到如何实现这一点。

3 线性回归模型

本书至今的重点一直是非监督学习，包括密度估计和数据聚类等主题。现在我们转向监督学习的讨论，从回归开始。回归的目标是：在给定 D 维输入向量 $\mathbf{x}$ 的值时下，预测一个或多个连续目标变量 t 的值。我们在第 1 章讨论多项式曲线拟合时，已经接触过一个回归问题的例子。多变量就是一大类称为线性回归模型的函数中的一个具体例子，这类模型的共同特点是它们是可调参数的线性函数，本章将以此为重点。最简单的线性回归模型同时也是输入变量的线性函数。然而，通过对一组固定的输入变量非线性函数（称为基函数）取线性组合，我们可以得到一类更加有用的函数。这类模型关于参数是线性的，因而具有简单的解析性质，同时关于输入变量却可以是非线性的。

给定一个包含 N 个观测的训练数据集 $\{\mathbf{x}_n\}$（其中 $n = 1, \ldots, N$），以及对应的目标值 $\{t_n\}$，我们的目标是预测当输入为新的 $\mathbf{x}$ 时 t 的值。最简单的方法是直接构造一个合适的函数 $y(\mathbf{x})$，使其在新输入 x 上的取值构成对相应 t 值的预测。更一般地，从概率的观点来看，我们的目标是建模预测分布 $p(t \mid \mathbf{x})$，因为它表达了对于每一个 $\mathbf{x}$，我们对 t 取值的不确定性。利用这个条件分布，我们可以通过最小化适当选择的损失函数的期望值，来对任意新 $\mathbf{x}$ 下的 t 进行预测。正如 1.5.5 节所讨论的，对于实值变量，常用的损失函数是平方损失，其最优解由 t 的条件期望给出。

尽管线性模型作为模式识别的实用技术存在显著局限性，尤其是在高维输入空间问题中，但它们具有良好的解析性质，并且构成了后续章节将要讨论的更复杂模型的基础。

3.1 线性基函数模型

最简单的回归线性模型是对输入变量进行线性组合的形式

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D \tag{3.1}$$

其中 $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$。这通常被简称为线性回归。该模型的关键性质是它关于参数 $w_0, \ldots, w_D$ 是线性的。然而，它同时也关于输入变量 x_i 是线性的，这给模型带来了显著的局限性。因此，我们通过考虑输入变量的固定非线性函数的线性组合来扩展模型类，得到如下形式

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) \tag{3.2}$$

其中 $\phi_j(\mathbf{x})$ 被称为基函数。通过令索引 j 的最大值为 $M-1$，该模型中的参数总数为 M。

参数 w_0 用来适应数据中的任何固定偏移，有时被称为偏置参数（不要与统计意义上的“偏差”混淆）。为了方便，通常定义一个额外的哑基函数 $\phi_0(\mathbf{x}) = 1$，从而

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}) \tag{3.3}$$

其中 $\mathbf{w} = (w_0, \ldots, w_{M-1})^{\mathrm{T}}$，$\phi = (\phi_0, \ldots, \phi_{M-1})^{\mathrm{T}}$。在许多模式识别的实际应用中，我们会对原始数据变量施加某种固定的预处理或特征提取。如果原始变量组成向量 $\mathbf{x}$，那么

特征就可以用基函数 $\{\phi_j(\mathbf{x})\}$ 来表示。

通过使用非线性基函数，我们允许函数 $y(\mathbf{x}, \mathbf{w})$ 成为输入向量 $\mathbf{x}$ 的非线性函数。然而，形如 (3.2) 的函数仍被称为线性模型，因为该函数关于参数 $\mathbf{w}$ 是线性的。正是这种参数的线性性质极大地简化了这类模型的分析。不过，这也带来了一些重要的局限性，我们将在 3.6 节中讨论。

第 1 章中考虑的多项式回归是这类模型的一个特例，其中只有一个输入变量 x，基函数采用 x 的幂形式，即 $\phi_j(x) = x^j$。多项式基函数的一个局限性在于它们是输入变量的全局函数，因此输入空间某一区域的变化会影响所有其他区域。这一问题可以通过将输入空间划分成多个区域，并在每个区域内拟合不同的多项式来解决，从而得到样条函数（Hastie et al., 2001）。

基函数还有许多其他选择，例如

$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\} \tag{3.4}$$

其中 μ_j 控制基函数在输入空间中的位置，参数 s 控制其空间尺度。这些通常被称为“高斯”基函数，但需要注意的是，它们并不需要具有概率意义，特别是归一化系数并不重要，因为这些基函数最终会乘以可调参数 w_j。

另一种可能是采用如下形式的 sigmoid 基函数

$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right) \tag{3.5}$$

其中 $\sigma(a)$ 是 logistic sigmoid 函数，定义为

$$\sigma(a) = \frac{1}{1+\exp(-a)} \tag{3.6}$$

等价地，我们也可以使用‘tanh’函数，因为它与 logistic sigmoid 的关系为 $\tanh(a) = 2\sigma(a) - 1$，因此 logistic sigmoid 函数的一般线性组合等价于‘tanh’函数的一般线性组合。这些不同的基函数选择在图 3.1 中得到了展示。

另一种可能的基函数选择是傅里叶基，它会导致正弦函数的展开。每个基函数代表一个特定的频率并具有无限的空间范围。相比之下，局限于输入空间有限区域的基函数必然包含不同空间频率的谱。在许多信号处理应用中，人们感兴趣的是同时在空间和频率上都局部化的基函数，这导致了一类称为小波的函数。这些小波还被定义为相互正交，以简化其应用。小波最适用于输入值位于规则网格上的情况，例如时间序列中的连续时间点或图像中的像素。关于小波的有用参考文献包括 Ogden (1997)、Mallat (1999) 和 Vidakovic (1999)。

本章的大部分讨论并不依赖于基函数集的具体选择，因此在大多数讨论中我们不会指定基函数的具体形式，除非为了数值说明。事实上，我们的许多讨论同样适用于基函数向量 $\boldsymbol{\phi}(\mathbf{x})$ 就是恒等的情形，即 $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}$。此外，为了保持记号简洁，我们将重点关注单个目标变量 t 的情况。不过，在 3.1.5 节中，我们将简要讨论处理多个目标变量所需的修改。

3.1.1 最大似然与最小二乘

在第 1 章中，我们通过最小化平方和误差函数将多项式函数拟合到数据集。我们还证明了该误差函数可以在假设高斯噪声模型下被解释为最大似然解。现在让我们回到这一讨论，更详细地考察最小二乘方法及其与最大似然的关系。

与之前一样，我们假设目标变量 t 由一个确定性函数 $y(\mathbf{x}, \mathbf{w})$ 加上加性高斯噪声给出，即

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon \tag{3.7}$$

其中 ϵ 是均值为零、精度（方差的倒数）为 β 的高斯随机变量。因此我们可以写出

$$p(t \mid \mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}\left(t \mid y(\mathbf{x}, \mathbf{w}), \beta^{-1}\right) \tag{3.8}$$

回忆一下，如果我们采用平方损失函数，那么对于新的 $\mathbf{x}$，最优预测将由目标变量的条件均值给出。对于形如 (3.8) 的高斯条件分布，条件均值简化为

$$\mathbb{E}[t \mid \mathbf{x}] = \int t p(t \mid \mathbf{x}) \mathrm{d}t = y(\mathbf{x}, \mathbf{w}) \tag{3.9}$$

需要注意的是，高斯噪声假设意味着给定 $\mathbf{x}$ 时 t 的条件分布是单峰的，这在某些应用中可能不合适。将条件分布扩展为高斯混合的形式（允许条件分布为多峰）将在 14.5.1 节讨论。

现在考虑输入集合 $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ 以及对应的目标值 $t_1, \ldots, t_N$。我们将目标变量 $\{t_n\}$ 组合成一个列向量，用 $\mathbf{t}$ 表示（采用不同字体以区别于多变量目标的单个观测 $\mathbf{t}$）。假设这些数据点独立地从分布 (3.8) 中抽取，可得似然函数关于可调参数 $\mathbf{w}$ 和 β 的表达式为

$$p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \mathbf{w}^{\mathrm{T}} \phi\left(\mathbf{x}_n\right), \beta^{-1}\right) \tag{3.10}$$

这里使用了 (3.3)。注意，在回归（以及分类）等监督学习问题中，我们并不试图对输入变量的分布建模。因此 $\mathbf{x}$ 总是出现在条件变量集合中，从现在起我们将省略诸如 $p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta)$ 中的显式 $\mathbf{x}$，以保持记号简洁。对似然函数取对数，并利用单变量高斯的标准形式 (1.46)，得到

$$\begin{aligned} \ln p(\mathbf{t} \mid \mathbf{w}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}\left(t_n \mid \mathbf{w}^{\mathrm{T}} \phi\left(\mathbf{x}_n\right), \beta^{-1}\right) \\ &= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w}) \end{aligned} \tag{3.11}$$

其中平方和误差函数定义为

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left\{t_n - \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}\left(\mathbf{x}_n\right)\right\}^2 \tag{3.12}$$

写出似然函数后，我们即可使用最大似然法确定 $\mathbf{w}$ 和 β。首先考虑关于 $\mathbf{w}$ 的最大化。正如 1.2.5 节已指出的，在线性模型的条件高斯噪声分布下，最大化似然函数等价

于最小化平方和误差函数 $E_D(\mathbf{w})$。对数似然函数 (3.11) 的梯度形式为

$$\nabla \ln p(\mathbf{t} \mid \mathbf{w}, \beta) = \sum_{n=1}^{N} \left\{ t_n - \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \right\} \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} \tag{3.13}$$

将该梯度设为零得到

$$0 = \sum_{n=1}^{N} t_n \phi(\mathbf{x}_n)^{\mathrm{T}} - \mathbf{w}^{\mathrm{T}} \left(\sum_{n=1}^{N} \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^{\mathrm{T}} \right) \tag{3.14}$$

求解 $\mathbf{w}$ 可得

$$\mathbf{w}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t} \tag{3.15}$$

这就是最小二乘问题的正规方程（normal equations）。其中 $\boldsymbol{\Phi}$ 是 $N \times M$ 矩阵，称为设计矩阵（design matrix），其元素为 $\Phi_{nj} = \phi_j(\mathbf{x}_n)$，因此

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix} \tag{3.16}$$

其中

$$\boldsymbol{\Phi}^{\dagger} \equiv \left(\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \tag{3.17}$$

称为矩阵 $\boldsymbol{\Phi}$ 的 Moore–Penrose 伪逆（Rao and Mitra, 1971; Golub and Van Loan, 1996）。它可视为矩阵逆概念对非方阵的推广。事实上，如果 $\boldsymbol{\Phi}$ 是可逆方阵，利用 $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ 可知 $\boldsymbol{\Phi}^{\dagger} = \boldsymbol{\Phi}^{-1}$。

此时我们可以深入理解偏置参数 w_0 的作用。如果显式写出偏置参数，误差函数 (3.12) 变为

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left\{ t_n - w_0 - \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}_n) \right\}^2 \tag{3.18}$$

对 w_0 求导并令其为零，解得

$$w_0 = \bar{t} - \sum_{j=1}^{M-1} w_j \overline{\phi_j} \tag{3.19}$$

其中定义了

$$\bar{t} = \frac{1}{N} \sum_{n=1}^{N} t_n, \quad \overline{\phi_j} = \frac{1}{N} \sum_{n=1}^{N} \phi_j(\mathbf{x}_n) \tag{3.20}$$

可见，偏置 w_0 用来补偿目标值平均值与基函数值加权平均值之和之间的差值。

我们还可以对噪声精度参数 β 最大化对数似然函数 (3.11)，得到

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{n=1}^{N} \left\{ t_n - \mathbf{w}_{\mathrm{ML}}^{\mathrm{T}} \phi(\mathbf{x}_n) \right\}^2 \tag{3.21}$$

因此，噪声精度的倒数正是目标值围绕回归函数的残差方差。

3.1.2 最小二乘的几何解释

此时考察最小二乘解的几何意义是很有启发的。为此，我们考虑一个 N 维空间，其坐标轴由 t_n 给出，从而 $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$ 是该空间中的一个向量。每个基函数 $\phi_j(\mathbf{x}_n)$ 在 N 个数据点上的取值也可以表示为同一空间中的向量，记作 $\boldsymbol{\varphi}_j$，如图 3.2 所示。注意 $\boldsymbol{\varphi}_j$ 对应于 $\boldsymbol{\Phi}$ 的第 j 列，而 $\phi(\mathbf{x}_n)$ 对应于 $\boldsymbol{\Phi}$ 的第 n 行。如果基函数数量 M 小于数据点数量 N，则 M 个向量 $\boldsymbol{\varphi}_j$ 将张成一个维度为 M 的线性子空间 $\mathcal{S}$。我们定义 $\mathbf{y}$ 为一个 N 维向量，其第 n 个元素为 $y(\mathbf{x}_n, \mathbf{w})$，其中 $n = 1, \ldots, N$。由于 $\mathbf{y}$ 是向量 $\boldsymbol{\varphi}_j$ 的任意线性组合，它可以位于 M 维子空间中的任意位置。平方和误差 (3.12) 则（除以因子 $1/2$ 外）等于 $\mathbf{y}$ 与 $\mathbf{t}$ 之间的平方欧氏距离。因此，$\mathbf{w}$ 的最小二乘解对应于子空间 $\mathcal{S}$ 中离 $\mathbf{t}$ 最近的那个 $\mathbf{y}$。从图 3.2 直观来看，我们预期该解对应于 $\mathbf{t}$ 在子空间 $\mathcal{S}$ 上的正交投影。事实也的确如此，这一点很容易验证：只需注意到 $\mathbf{y}$ 的解为 $\boldsymbol{\Phi}\mathbf{w}_{\mathrm{ML}}$，然后确认它正是正交投影的形式。

在实际中，直接求解正规方程可能在 $\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$ 接近奇异时导致数值困难。特别是当两个或多个基向量 $\boldsymbol{\varphi}_j$ 共线或几乎共线时，得到的参数值可能具有很大的幅度。在处理真实数据集时，这种近似退化并不罕见。由此产生的数值困难可以使用奇异值分解（SVD）技术来解决（Press et al., 1992; Bishop and Nabney, 2008）。需要注意的是，加入正则化项可以保证矩阵非奇异，即使存在退化情况。

3.1.3 序贯学习

批处理技术（如最大似然解 (3.15)）需要一次性处理整个训练集，对于大数据集来说计算代价高昂。正如第 1 章所讨论的，如果数据集足够大，使用序贯算法（也称为在线算法）可能是值得的：在这种算法中，数据点逐个考虑，每次呈现一个数据点后就更新模型参数。序贯学习也适用于实时应用场景，此时数据观测以连续流的形式到来，并且必须在看到所有数据点之前就做出预测。

我们可以通过应用随机梯度下降技术（也称为序贯梯度下降）得到一种序贯学习算法，具体如下。如果误差函数是关于数据点的求和 $E = \sum_n E_n$，则在呈现第 n 个模式后，随机梯度下降算法使用以下方式更新参数向量 $\mathbf{w}$

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_n \tag{3.22}$$

其中 τ 表示迭代步数，η 是学习率参数。我们稍后将讨论 η 的取值选择。$\mathbf{w}$ 的值初始化为某个起始向量 $\mathbf{w}^{(0)}$。对于平方和误差函数 (3.12)，这给出了

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \eta \left(t_n - \mathbf{w}^{(\tau)\mathrm{T}} \phi_n\right) \phi_n \tag{3.23}$$

其中 $\phi_n = \phi(\mathbf{x}_n)$。这被称为最小均方算法（least-mean-squares，简称 LMS）。η 的取值需要小心选择，以确保算法收敛（Bishop and Nabney, 2008）。

3.1.4 正则化最小二乘

在 1.1 节中，我们介绍了向误差函数中添加正则化项以控制过拟合的思想，从而要求最小化的总误差函数形式为

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) \tag{3.24}$$

其中 λ 是正则化系数，用于控制数据相关误差 $E_D(\mathbf{w})$ 与正则化项 $E_W(\mathbf{w})$ 的相对重要性。最简单的正则化器之一是权重向量元素的平方和

$$E_W(\mathbf{w}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \tag{3.25}$$

若同时考虑平方和误差函数

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\right\}^2 \tag{3.26}$$

则总误差函数变为

$$\frac{1}{2}\sum_{n=1}^{N}\left\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \tag{3.27}$$

这种特定的正则化器在机器学习文献中被称为权重衰减（weight decay），因为在序贯学习算法中，它会促使权重值向零衰减，除非得到数据的支持。在统计学中，它是参数收缩（parameter shrinkage）方法的一个例子，因为它使参数值向零收缩。其优点在于误差函数关于 $\mathbf{w}$ 仍保持为二次函数，因此可以以闭合形式精确求得其极小值。具体地，将 (3.27) 关于 $\mathbf{w}$ 的梯度设为零并求解，得到

$$\mathbf{w} = \left(\lambda\mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} \tag{3.28}$$

这是在最小二乘解 (3.15) 基础上的一个简单扩展。

有时会使用更一般的正则化器，此时正则化误差采取如下形式

$$\frac{1}{2}\sum_{n=1}^{N}\left\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\right\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q \tag{3.29}$$

其中 $q = 2$ 对应于二次正则化器 (3.27)。图 3.3 显示了不同 q 值下正则化项的等高线。

$q = 1$ 的情形在统计学文献中被称为 lasso（Tibshirani, 1996）。它具有如下性质：当 λ 足够大时，某些系数 w_j 会被压缩到零，从而得到一个稀疏模型，其中对应的基函数不起任何作用。为了看出这一点，我们首先注意到，最小化 (3.29) 等价于在如下约束下最小化未经正则化的平方和误差 (3.12)

$$\sum_{j=1}^{M}|w_j|^q \leqslant \eta \tag{3.30}$$

其中 η 是适当的参数值，这两种方法可以通过拉格朗日乘子相互关联。稀疏性的来源可见于图 3.4，该图展示了在约束 (3.30) 下误差函数的最小值。随着 λ 的增大，越来越多的参数被压缩到零。

正则化允许在有限规模的数据集上训练复杂模型，而不会出现严重的过拟合，其本质是通过限制有效模型复杂度来实现。然而，确定最优模型复杂度的问题也随之从寻找适当的基函数数量转移到确定正则化系数 λ 的合适取值上。我们将在本章后面重新讨论模型复杂度问题。

在本章余下部分，我们将专注于二次正则化器 (3.27)，既因为它在实践中重要，也因为它在数学上易于处理。

3.1.5 多输出

到目前为止，我们一直考虑单个目标变量 t 的情况。在某些应用中，我们可能希望同时预测 $K > 1$ 个目标变量，统称为目标向量 $\mathbf{t}$。一种做法是为 $\mathbf{t}$ 的每个分量引入不同的基函数集合，从而得到多个独立的回归问题。然而，更有趣也更常见的方法是使用同一组基函数来对目标向量的所有分量进行建模，即

$$\mathbf{y}(\mathbf{x}, \mathbf{w}) = \mathbf{W}^{\mathrm{T}}\phi(\mathbf{x}) \tag{3.31}$$

其中 $\mathbf{y}$ 是 K 维列向量，$\mathbf{W}$ 是 $M \times K$ 参数矩阵，$\phi(\mathbf{x})$ 是 M 维列向量，其元素为 $\phi_j(\mathbf{x})$，且仍令 $\phi_0(\mathbf{x}) = 1$。假设目标向量的条件分布为各向同性高斯分布

$$p(\mathbf{t} \mid \mathbf{x}, \mathbf{W}, \beta) = \mathcal{N}\left(\mathbf{t} \mid \mathbf{W}^{\mathrm{T}}\phi(\mathbf{x}), \beta^{-1}\mathbf{I}\right) \tag{3.32}$$

若有观测 $\mathbf{t}_1, \ldots, \mathbf{t}_N$，我们可将其组装成 $N \times K$ 矩阵 $\mathbf{T}$，其第 n 行为 $\mathbf{t}_n^{\mathrm{T}}$。类似地，将输入向量 $\mathbf{x}_1, \ldots, \mathbf{x}_N$ 组装成矩阵 $\mathbf{X}$。此时对数似然函数为

$$\begin{aligned} \ln p(\mathbf{T} \mid \mathbf{X}, \mathbf{W}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}\left(\mathbf{t}_n \mid \mathbf{W}^{\mathrm{T}}\phi(\mathbf{x}_n), \beta^{-1}\mathbf{I}\right) \\ &= \frac{NK}{2}\ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}\sum_{n=1}^{N}\left\|\mathbf{t}_n - \mathbf{W}^{\mathrm{T}}\phi(\mathbf{x}_n)\right\|^2. \end{aligned} \tag{3.33}$$

与之前一样，我们对 $\mathbf{W}$ 最大化该函数，得到

$$\mathbf{W}_{\mathrm{ML}} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{T}. \tag{3.34}$$

如果我们对每个目标变量 t_k 单独考察，就有

$$\mathbf{w}_k = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}_k = \mathbf{\Phi}^{\dagger}\mathbf{t}_k \tag{3.35}$$

其中 $\mathbf{t}_k$ 是 N 维列向量，其分量为 t_{nk}，$n = 1, \ldots, N$。因此回归问题的解在不同目标变量之间是解耦的，我们只需计算一次伪逆矩阵 $\mathbf{\Phi}^{\dagger}$，它被所有 $\mathbf{w}_k$ 共享。

扩展到具有任意协方差矩阵的一般高斯噪声分布也是直观的，同样会导致 K 个独立的回归问题。这一结果并不令人意外，因为参数 $\mathbf{W}$ 只定义了高斯噪声分布的均值，而我们从 2.3.4 节知道，多变量高斯分布均值的最大似然解与协方差无关。从现在起，为简便起见，我们将只考虑单个目标变量 t。

3.2 偏差-方差分解

到目前为止，在讨论回归的线性模型时，我们一直假设基函数的形式和数量都是固定的。正如第 1 章所见，如果用数据规模有限的数据集训练复杂模型，最大似然（或等价的最小二乘）会导致严重的过拟合。然而，为了避免过拟合而限制基函数数量的副作用是限制了模型捕捉数据中重要趋势的灵活性。虽然引入正则化项可以对参数较多的模型控制过拟合，但这又引出了如何确定正则化系数 λ 合适值的问题。显然，同时对权重向量 $\mathbf{w}$ 和正则化系数 λ 最小化正则化误差函数并不是正确做法，因为这会导致无正则化的解，即 $\lambda = 0$。

正如前几章所见，过拟合本质上是最大似然的一个缺陷，而在贝叶斯框架下对参数边缘化时就不会出现这种现象。在本章中，我们将深入探讨模型复杂度的贝叶斯观点。不过在此之前，先从频度学派的视角考察模型复杂度问题——即偏差-方差权衡（bias-variance trade-off）——是很有启发的。虽然我们将在简单易于说明的线性基函数模型背景下引入这一概念，但相关讨论具有更广泛的适用性。

在 1.5.5 节讨论回归问题的决策理论时，我们考虑了多种损失函数，每一种在给定条件分布 $p(t \mid \mathbf{x})$ 后都会导向相应的最优预测。常用的平方损失函数对应的最优预测是条件期望，我们记作 $h(\mathbf{x})$，定义为

$$h(\mathbf{x}) = \mathbb{E}[t \mid \mathbf{x}] = \int t\, p(t \mid \mathbf{x})\, \mathrm{d}t \tag{3.36}$$

此时有必要区分决策理论中产生的平方损失函数与最大似然参数估计中出现的平方和误差函数。我们可能会采用比最小二乘更复杂的技术（例如正则化或完全贝叶斯方法）来确定条件分布 $p(t \mid \mathbf{x})$。这些方法都可以与平方损失函数结合，用于做出预测。

我们在 1.5.5 节证明，期望平方损失可以写成如下形式

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x}) \mathrm{d}\mathbf{x} + \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) \mathrm{d}\mathbf{x} \mathrm{d}t \tag{3.37}$$

回忆一下，第二项与 $y(\mathbf{x})$ 无关，它源于数据的固有噪声，代表期望损失可达到的最小值。第一项依赖于我们对函数 $y(\mathbf{x})$ 的选择，我们将寻求使该项最小的 $y(\mathbf{x})$。由于它非负，我们希望的最好结果是使其为零。如果拥有无限数据（以及无限计算资源），原则上我们可以以任意精度求得回归函数 $h(\mathbf{x})$，这将是 $y(\mathbf{x})$ 的最优选择。然而实际中我们只有包含有限 N 个数据点的有限数据集 $\mathcal{D}$，因此无法精确知道回归函数 $h(\mathbf{x})$。

如果我们使用参数函数 $y(\mathbf{x}, \mathbf{w})$ 来建模 $h(\mathbf{x})$，其中由参数向量 $\mathbf{w}$ 控制，那么从贝叶斯视角，模型的不确定性通过 $\mathbf{w}$ 的后验分布表达。而频度学派处理方式则是基于数据集 $\mathcal{D}$ 对 $\mathbf{w}$ 做一个点估计，并通过如下思想实验来解释该估计的不确定性：假设我们有大量大小均为 N 的数据集，每个数据集都独立地从分布 $p(t, \mathbf{x})$ 中抽取。对于任意给定的数据集 $\mathcal{D}$，我们运行学习算法得到预测函数 $y(\mathbf{x}; \mathcal{D})$。来自集合的不同数据集会给出不同的函数，从而产生不同的平方损失值。特定学习算法的性能通过对这一数据集集合取平均来评估。

考虑 (3.37) 中第一项的被积函数，对于特定数据集 $\mathcal{D}$ 其形式为

$$\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2 \tag{3.38}$$

由于该量依赖于具体数据集 $\mathcal{D}$，我们对其在数据集集合上取平均。如果在大括号内加减 $\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]$，然后展开，可得

$$\begin{aligned}
&\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 \\
&= \{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 \\
&\quad + 2\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}
\end{aligned} \tag{3.39}$$

现在我们对 $\mathcal{D}$ 取该表达式的期望，注意到最后一项将消失，得到

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2\right] \\
&\quad = \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]}_{\text{variance}}
\end{aligned} \tag{3.40}$$

可见，$y(\mathbf{x};\mathcal{D})$ 与回归函数 $h(\mathbf{x})$ 之间期望平方差可以表示为两个项之和。第一项称为平方偏差（squared bias），它衡量所有数据集上的平均预测与理想回归函数的差异程度。第二项称为方差（variance），它衡量各个数据集的解围绕其平均值的波动程度，从而反映函数 $y(\mathbf{x};\mathcal{D})$ 对具体数据集选择的敏感程度。我们稍后通过一个简单例子来直观说明这些定义。

到目前为止，我们只考虑单个输入值 $\mathbf{x}$。若将此展开式代回 (3.37)，就得到期望平方损失的如下分解

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise} \tag{3.41}$$

其中

$$(\text{bias})^2 = \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{3.42}$$

$$\text{variance} = \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{3.43}$$

$$\text{noise} = \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \tag{3.44}$$

此时偏差和方差项指代的是积分量。

我们的目标是最小化期望损失，它被分解为（平方）偏差、方差以及一个常数噪声项之和。正如我们将看到的，偏差与方差之间存在权衡：非常灵活的模型偏差低但方差高，相对刚性的模型偏差高但方差低。预测能力最优的模型是在偏差与方差之间取得最佳平衡的模型。这可以通过第 1 章中的正弦数据集来说明。这里我们独立地从正弦曲线 $h(x) = \sin(2\pi x)$ 生成 100 个数据集，每个数据集包含 $N = 25$ 个数据点。数据集编号为 $l = 1,\ldots,L$，其中 $L = 100$。对每个数据集 $\mathcal{D}^{(l)}$，我们通过最小化正则化误差函数 (3.27) 拟合 24 个高斯基函数，得到预测函数 $y^{(l)}(x)$，如图 3.5 所示。

第一行对应于正则化系数 λ 很大的情况：方差低（左图中红色曲线看起来很相似），但偏差高（右图中平均拟合与真实正弦函数相差很大）。最后一行对应于 λ 很小的情形：

方差大（左图中红色曲线之间变化剧烈），但偏差低（右图中平均模型拟合与原始正弦函数吻合良好）。值得注意的是，对复杂模型（$M = 25$）的多个解取平均后，能极好地拟合回归函数，这提示我们平均可能是一种有益的操作。实际上，对多个解进行加权平均正是贝叶斯方法的核心，只不过平均是对参数后验分布进行的，而不是对多个数据集进行的。

我们也可以对该例子进行偏差-方差权衡的定量分析。平均预测通过下式估计

$$\bar{y}(x) = \frac{1}{L}\sum_{l=1}^{L} y^{(l)}(x) \tag{3.45}$$

积分平方偏差和积分方差则由下式给出

$$(\text{bias})^2 = \frac{1}{N}\sum_{n=1}^{N}\{\bar{y}(x_n) - h(x_n)\}^2 \tag{3.46}$$

$$\text{variance} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{l=1}^{L}\left\{y^{(l)}(x_n) - \bar{y}(x_n)\right\}^2 \tag{3.47}$$

其中关于 x 加权分布 $p(x)$ 的积分通过从该分布抽取的数据点的有限和来近似。这些量连同它们的和作为 $\ln\lambda$ 的函数绘于图 3.6 中。我们看到，λ 的取值很小时，模型会精细地拟合每个单独数据集上的噪声，导致方差很大。相反，λ 取值很大时会把权重参数拉向零，导致偏差很大。

尽管偏差-方差分解从频度学派视角为模型复杂度问题提供了一些有趣的洞察，但其实际价值有限，因为偏差-方差分解基于对数据集集合的平均，而实际中我们只有一个观测到的数据集。如果我们有大量给定大小的独立训练集，最好将它们合并成一个大的训练集，这当然会降低给定模型复杂度下的过拟合程度。

鉴于这些局限性，我们将在下一节转向对线性基函数模型的贝叶斯处理，这不仅为过拟合问题提供了深刻的洞察，还导向了解决模型复杂度问题的实用技术。

3.3 贝叶斯线性回归

在讨论线性回归模型参数的最大似然设定时，我们已经看到，有效模型复杂度（由基函数数量决定）必须根据数据集大小进行控制。在对数似然函数中加入正则化项后，有效模型复杂度就可以通过正则化系数的值来控制，不过基函数的数量和形式选择当然仍然对模型整体行为至关重要。

这就留下了一个问题：如何为具体问题确定合适的模型复杂度？单纯最大化似然函数显然不行，因为这总是导致过于复杂的模型和过拟合。如 1.3 节所述，可以使用独立的验证集来确定模型复杂度，但这样做既计算昂贵，又浪费宝贵的数据。因此，我们转向线性回归的贝叶斯处理，它将避免最大似然带来的过拟合问题，并仅使用训练数据就能自动确定模型复杂度。为简单起见，我们仍重点讨论单个目标变量 t 的情况。扩展到多个目标变量是直观的，只需遵循 3.1.5 节的讨论即可。

3.3.1 参数分布

我们从对模型参数 $\mathbf{w}$ 引入先验概率分布开始讨论线性回归的贝叶斯处理。暂时将噪声精度参数 β 视为已知常数。首先注意，由 (3.10) 定义的似然函数 $p(\mathbf{t} \mid \mathbf{w})$ 是 $\mathbf{w}$ 的二次函数的指数形式。因此对应的共轭先验是高斯分布

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_0, \mathbf{S}_0) \tag{3.48}$$

其中均值为 $\mathbf{m}_0$，协方差为 $\mathbf{S}_0$。

接下来我们计算后验分布，它正比于似然函数与先验的乘积。由于选择了共轭高斯先验，后验分布也将是高斯的。我们可以通过在指数中完成平方并利用标准高斯归一化结果来求得该分布。不过，我们已经在推导通用结果 (2.116) 时完成了必要工作，因此可以直接写出后验分布的形式

$$p(\mathbf{w} \mid \mathbf{t}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_N, \mathbf{S}_N) \tag{3.49}$$

其中

$$\mathbf{m}_N = \mathbf{S}_N \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\right) \tag{3.50}$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} \tag{3.51}$$

需要注意的是，由于后验分布是高斯的，其众数与均值重合。因此最大后验权重向量就是 $\mathbf{w}_{\mathrm{MAP}} = \mathbf{m}_N$。如果考虑无限宽的先验 $\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$ 并令 $\alpha \to 0$，则后验均值 $\mathbf{m}_N$ 将退化为最大似然值 $\mathbf{w}_{\mathrm{ML}}$（由 (3.15) 给出）。类似地，若 $N = 0$，后验分布退回到先验。此外，如果数据点逐个到达，则任意阶段的后验分布可作为下一个数据点的先验，新后验分布仍由 (3.49) 给出。

在本章余下部分，我们将考虑一种特殊的高斯先验以简化处理。具体地，我们采用零均值、各向同性的高斯先验，仅由单一精度参数 α 控制，即

$$p(\mathbf{w} \mid \alpha) = \mathcal{N}\left(\mathbf{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I}\right) \tag{3.52}$$

此时关于 $\mathbf{w}$ 的后验分布仍由 (3.49) 给出，但参数变为

$$\mathbf{m}_N = \beta\mathbf{S}_N\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} \tag{3.53}$$

$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} \tag{3.54}$$

后验分布的对数是似然函数对数与先验对数之和，作为 $\mathbf{w}$ 的函数具有如下形式

$$\ln p(\mathbf{w} \mid \mathbf{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}\left\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\}^2 - \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \text{const.} \tag{3.55}$$

因此，对该后验分布关于 $\mathbf{w}$ 的最大化等价于最小化平方和误差函数加上二次正则化项，即 (3.27) 中 $\lambda = \alpha/\beta$ 的情形。

我们可以通过一个简单的直线拟合例子来说明线性基函数模型中的贝叶斯学习以及后验分布的序贯更新。考虑单个输入变量 x、单个目标变量 t 和形如 $y(x, \mathbf{w}) = w_0 + w_1 x$

的线性模型。由于只有两个可调参数，我们可以直接在参数空间中绘制先验和后验分布。我们可以通过一个简单的直线拟合示例，在线性基函数模型中说明贝叶斯学习以及后验分布的序列更新。考虑单一输入变量 x、单一目标变量 t，以及形如 $y(x, \mathbf{w}) = w_0 + w_1 x$ 的线性模型。由于该模型只有两个可调参数，我们可以直接在参数空间中绘制先验和后验分布。我们从函数 $f(x, \mathbf{a}) = a_0 + a_1 x$ 生成合成数据，其中参数取值为 $a_0 = -0.3$ 和 $a_1 = 0.5$。具体方法是，首先从均匀分布 $U(x \mid -1, 1)$ 中选取 x_n 的值，然后计算 $f(x_n, \mathbf{a})$，最后添加标准差为 0.2 的高斯噪声以得到目标值 t_n。我们的目标是根据这些数据恢复 a_0 和 a_1 的取值，并将探讨数据集大小的影响。这里假设噪声方差已知，因此将精度参数设为其真实值 $\beta = (1/0.2)^2 = 25$ 类似地，将参数 $\alpha = 2.0$ 固定下来。稍后我们将讨论从训练数据中确定 α 和 β 的策略。图 3.7 展示了在该模型中，随着数据集大小增加的贝叶斯学习结果，并演示了贝叶斯学习的序列特性：当观察到新数据点时，当前后验分布将作为新的先验。值得花时间详细分析该图，因为它说明了贝叶斯推断中的多个重要方面。

图的第一行对应未观察到任何数据点的情况，显示了 w 空间中的先验分布，以及从先验中采样得到的六条函数 $y(x, \mathbf{w})$ 的曲线。第二行展示了观察到一个数据点后的情况。右侧列用蓝色圆点表示数据点 (x, t)。左侧列显示了该数据点对应的似然函数 $p(t \mid x, \mathbf{w})$ 作为 **w** 的函数。注意，似然函数提供的是一种“柔性”约束，即直线需要经过数据点附近，而“附近”的程度由噪声精度 β 决定。为了对比，生成数据集时使用的真实参数 $a_0 = -0.3$、$a_1 = 0.5$ 在左侧的图中以白色叉号标出。将该似然函数与第一行的先验相乘并归一化，即得到第二行中间的后验分布。从该后验分布中对 **w** 进行采样，得到的回归函数样本显示在右侧的图中。可以看到，这些样本曲线都经过该数据点附近。第三行展示了观察第二个数据点的效果。同样在右侧列以蓝色圆点表示。左侧图显示仅由第二个数据点得到的似然函数。当我们将该似然函数与第二行的后验分布相乘时，得到第三行中间的后验分布。注意，这与直接将原始先验与两个数据点的似然函数相结合所得结果完全相同。此时后验已受到两个数据点的影响，而两个点足以确定一条直线，因此后验已经相对集中。该后验的样本生成的函数以红色显示，可以看到它们同时经过两个数据点附近。第四行展示了观察到总共 20 个数据点的情况。左侧图为第 20 个数据点单独的似然函数，中间图为结合所有 20 个观测后的后验分布。可以注意到此时后验比第三行显著更尖锐。在数据点数趋于无穷时，后验分布将收缩为一个以真实参数值（白色叉号所示）为中心的狄拉克 delta 分布。

可以考虑参数上的其他形式的先验。例如，我们可以将高斯先验推广为

$$p(\mathbf{w} \mid \alpha) = \left[\frac{q}{2} \left(\frac{\alpha}{2} \right)^{1/q} \frac{1}{\Gamma(1/q)} \right]^M \exp\left(-\frac{\alpha}{2} \sum_{j=1}^{M} |w_j|^q \right) \tag{3.56}$$

其中 $q = 2$ 时对应高斯分布，并且只有在这种情况下先验与似然函数 (3.10) 是共轭的。在 w 上对后验分布取最大值等价于最小化正则化误差函数 (3.29)。对于高斯先验，后验分布的众数等于其均值，但如果 $q \neq 2$，这一性质将不再成立。

3.3.2 预测分布

在实际中，我们通常并不直接关心 **w** 的取值，而是希望对新的 **x** 预测 t。这需要计算定义为

$$p(t \mid \mathbf{t}, \alpha, \beta) = \int p(t \mid \mathbf{w}, \beta) p(\mathbf{w} \mid \mathbf{t}, \alpha, \beta) \mathrm{d}\mathbf{w} \tag{3.57}$$

的预测分布，其中 **t** 是训练集中的目标值向量，并且为了简化符号，我们在条件语句中省略了对应的输入向量。目标变量的条件分布 $p(t \mid \mathbf{x}, \mathbf{w}, \beta)$ 由 (3.8) 给出，而权重的后验分布由 (3.49) 给出。注意 (3.57) 涉及两个高斯分布的卷积，因此利用第 8.1.4 节中的结果 (2.115)，可得预测分布为

$$p(t \mid \mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}\left(t \mid \mathbf{m}_N^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x})\right) \tag{3.58}$$

其中预测分布方差 $\sigma_N^2(\mathbf{x})$ 为

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}) \tag{3.59}$$

(3.59) 中的第一项表示数据噪声，而第二项反映参数 w 的不确定性。由于噪声和 **w** 的分布均为高斯分布，它们的方差可以相加。注意，随着观察到更多的数据点，后验分布会逐渐收缩。因此可以证明 (Qazaz et al., 1997)，有 $\sigma_{N+1}^2(\mathbf{x}) \leqslant \sigma_N^2(\mathbf{x})$。当 $N \to \infty$ 时，(3.59) 中的第二项趋于 0，此时预测分布的方差仅由噪声参数 β 决定。

为了说明贝叶斯线性回归模型的预测分布，我们回到第 1.1 节中的合成正弦数据集。在图 3.8 中，我们将由高斯基函数线性组合构成的模型拟合到不同规模的数据集上，并观察相应的后验分布。图中绿色曲线表示生成数据点（加高斯噪声）的函数 $\sin(2\pi x)$。四幅图中蓝色圆点分别表示规模为 $N = 1$、$N = 2$、$N = 4$ 和 $N = 25$ 的数据集。对于每幅图，红色曲线表示相应高斯预测分布的均值，而红色阴影区域表示均值左右一个标准差的范围。可以看到，预测不确定性依赖于 x，并且在数据点附近最小。同时注意到，随着观测数据点的数量增加，不确定性逐渐降低。

图 3.8 中的图仅展示了预测方差随 x 变化的点对点形式。为了进一步理解在不同 x 取值下预测之间的协方差关系，我们可以从 **w** 的后验分布中进行采样，然后绘制对应的函数 $y(x, \mathbf{w})$，如图 3.9 所示。

如果使用局部化的基函数（例如高斯函数），那么在远离基函数中心的区域中，(3.59) 中预测方差的第二项贡献将趋于零，只剩下噪声项 β^{-1}。因此，在超出基函数覆盖范围进行外推时，模型会对自己的预测表现得非常自信，而这通常是一种不理想的行为。这个问题可以通过采用另一种用于回归的贝叶斯方法来避免，该方法称为高斯过程。

注意，如果将 **w** 和 β 都视为未知量，那么我们可以为 $p(\mathbf{w}, \beta)$ 引入一个共轭先验分布，根据第 2.3.6 节的讨论，该先验将是高斯-伽马分布（Denison et al., 2002）。在这种情况下，预测分布将是 Student’s t 分布。

3.3.3 等价核

线性基函数模型的后验均值解（3.53）有一个有趣的解释，这将为包括高斯过程在内的核方法奠定基础。如果我们将（3.53）代入表达式（3.3），可以看到预测均值可以写成如下形式

$$y(\mathbf{x}, \mathbf{m}_N) = \mathbf{m}_N^{\mathrm{T}}\phi(\mathbf{x}) = \beta\phi(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} = \sum_{n=1}^{N}\beta\phi(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\phi(\mathbf{x}_n)t_n \tag{3.60}$$

其中 $\mathbf{S}_N$ 由（3.51）定义。因此，在点 $\mathbf{x}$ 处预测分布的均值可以表示为训练集目标变量 t_n 的线性组合，从而我们可以写为

$$y(\mathbf{x}, \mathbf{m}_N) = \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n)t_n \tag{3.61}$$

其中函数

$$k(\mathbf{x}, \mathbf{x}') = \beta\phi(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\phi(\mathbf{x}') \tag{3.62}$$

被称为平滑矩阵或等效核。像这样的回归函数通过对训练集目标值进行线性组合来进行预测，被称为线性平滑器。注意，等效核依赖于数据集中输入值 $\mathbf{x}_n$，因为这些值出现在 $\mathbf{S}_N$ 的定义中。图 3.10 展示了高斯基函数情况下的等效核，其中核函数 $k(x, x')$ 随着 x' 绘制，并对应三个不同的 x 值。可见这些函数都集中在 x 附近，因此在 x 处预测分布的均值 $y(x, \mathbf{m}_N)$ 是通过加权组合目标值得到的，其中靠近 x 的数据点具有更高权重，而离 x 较远的数据点权重较低。直观地说，更强地权衡局部证据而非远处证据是合理的。注意，这种局部化性质不仅出现在局部的高斯基函数中，也出现在非局部的多项式和 S 形（sigmoidal）基函数中，如图 3.11 所示。

对等价核作用的进一步理解可以通过考虑 $y(\mathbf{x})$ 与 $y(\mathbf{x}')$ 之间的协方差来获得，其形式为

$$\begin{aligned}\mathrm{cov}[y(\mathbf{x}), y(\mathbf{x}')] &= \mathrm{cov}[\phi(\mathbf{x})^{\mathrm{T}}\mathbf{w}, \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}')] \\ &= \phi(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\phi(\mathbf{x}') = \beta^{-1}k(\mathbf{x}, \mathbf{x}')\end{aligned} \tag{3.63}$$

这里使用了 (3.49) 和 (3.62)。由等价核的形式可以看出，相邻点处的预测均值具有较强的相关性，而较远点之间的相关性则较弱。

图 3.8 所示的预测分布允许我们通过 (3.59) 观察预测在每一个点上的不确定性。然而，通过从 $\mathbf{w}$ 的后验分布中抽取样本并在图 3.9 中绘制相应的模型函数 $y(\mathbf{x}, \mathbf{w})$，我们实际上是在可视化由等价核决定的，在两个（或更多）不同 x 值处的 y 值之间的联合不确定性。

将线性回归表示为核函数形式提示了另一种回归方法。与其引入一组基函数（从而隐式决定等价核），我们可以直接定义一个局部化的核函数，并利用它在给定训练集的情况下对新的输入向量 $\mathbf{x}$ 进行预测。这引出了一个用于回归（和分类）的实用框架，称为高斯过程，我们将在第 6.4 节中详细讨论。

我们已经看到，有效核决定了在对新的 $\mathbf{x}$ 进行预测时，训练集目标值被组合的权重，并且可以证明这些权重之和为 1，即

$$\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) = 1 \tag{3.64}$$

对所有 $\mathbf{x}$ 都成立。这个直观上令人满意的结果可以通过一个非正式推断得到：考虑一个训练集，其中所有目标值 $t_n = 1$，则其预测均值记为 $\widehat{y}(\mathbf{x})$。只要基函数线性无关、数据点数量多于基函数数量，并且包含一个常数基函数（对应偏置参数），那么我们可以精确拟合训练数据，从而预测均值必为 $\widehat{y}(x) = 1$，由此得到 (3.64)。需要注意的是，核函数的值可以为负也可以为正，因此尽管满足加和约束，预测并不一定是训练集目标值的凸组合。

最后，我们注意到等价核 (3.62) 具有核函数的一项重要通性，即它可以表示为关于一组非线性函数向量 $\psi(x)$ 的内积形式，因此有

$$k(\mathbf{x}, \mathbf{z}) = \psi(\mathbf{x})^{\mathrm{T}} \psi(\mathbf{z}) \tag{3.65}$$

其中 $\psi(\mathbf{x}) = \beta^{1/2} \mathbf{S}_N^{1/2} \phi(\mathbf{x})$。

3.4 贝叶斯模型比较

在第 1 章中，我们强调了过拟合的问题，以及使用交叉验证来设定正则化参数取值或在备选模型之间进行选择的方法。在这里，我们从贝叶斯的角度来讨论模型选择的问题。本节将给出一个非常一般性的讨论，并在第 3.5 节中看到如何将这些思想用于线性回归中正则化参数的确定。

正如我们将看到的，通过对模型参数进行边缘化（求和或积分），而不是对其进行点估计，可以避免最大似然引起的过拟合。这样，模型可以直接使用训练数据进行比较，而不需要验证集。这使得可以利用全部可用数据进行训练，也避免了交叉验证中每个模型都需要多次训练的问题。同时，这还允许在训练过程中同时确定多个复杂度参数。例如，在第 7 章中我们将介绍相关向量机，它是一个贝叶斯模型，并为每一个训练数据点设置一个复杂度参数。

贝叶斯模型比较的观点非常简单，即使用概率来表示对模型选择的不确定性，并一致地使用概率的和法则和积法则。假设我们希望比较一组模型 $\mathcal{M}_i$，其中 $i = 1, \ldots, L$。这里模型指的是定义在观测数据 $\mathcal{D}$ 上的概率分布。在多项式曲线拟合问题中，该分布定义在目标值向量 $\mathbf{t}$ 上，并假设输入值集合 $\mathbf{X}$ 已知。其他类型的模型会对 $\mathbf{X}$ 和 $\mathbf{t}$ 的联合分布进行建模。我们假设数据由这些模型之一生成，但我们并不确定是哪一个。这种不确定性用先验分布 $p(\mathcal{M}_i)$ 来表达。在给定训练集 $\mathcal{D}$ 的情况下，我们希望计算后验分布

$$p(\mathcal{M}_i \mid \mathcal{D}) \propto p(\mathcal{M}_i) p(\mathcal{D} \mid \mathcal{M}_i). \tag{3.66}$$

先验允许我们表达对不同模型的偏好。我们可以简单地假设所有模型具有相同的先验概率。更有趣的项是模型证据 $p(\mathcal{D} \mid \mathcal{M}_i)$，它表示数据对不同模型的偏好，我们将稍后

更详细地研究这一项。模型证据有时也称为边缘似然，因为它可以被看作模型空间上的似然函数，其中参数已经被边缘化掉。两个模型的模型证据之比 $p(\mathcal{D} \mid \mathcal{M}_i)/p(\mathcal{D} \mid \mathcal{M}_j)$ 被称为贝叶斯因子（Kass and Raftery, 1995）。

一旦我们知道模型的后验分布，根据和法则与积法则，预测分布可写为

$$p(t \mid \mathbf{x}, \mathcal{D}) = \sum_{i=1}^{L} p(t \mid \mathbf{x}, \mathcal{M}_i, \mathcal{D}), p(\mathcal{M}_i \mid \mathcal{D}) \tag{3.67}$$

这是一个混合分布的例子,其中整体预测分布由各个模型的预测分布 $p(t \mid \mathbf{x}, \mathcal{M}_i, \mathcal{D})$ 按照其后验概率 $p(\mathcal{M}_i \mid \mathcal{D})$ 加权求平均得到。例如，如果存在两个后验上同等可能的模型，一个给出在 $t = a$ 附近的窄分布，另一个给出在 $t = b$ 附近的窄分布，那么整体预测分布将是一个在 $t = a$ 和 $t = b$ 处具有两个峰值的双峰分布，而不是在 $t = (a + b)/2$ 处产生一个单峰分布。

对模型平均的一种简单近似是仅使用后验概率最大的模型来进行预测，这称为模型选择。

对于由参数向量 $\mathbf{w}$ 所决定的模型，根据概率的和法则与积法则，模型证据为

$$p(\mathcal{D} \mid \mathcal{M}_i) = \int p(\mathcal{D} \mid \mathbf{w}, \mathcal{M}_i), p(\mathbf{w} \mid \mathcal{M}_i) \mathrm{d}\mathbf{w} \tag{3.68}$$

从采样的角度看，边缘似然可以理解为：从先验中随机采样参数后由模型生成数据集 $\mathcal{D}$ 的概率。还值得注意的是，该模型证据正是贝叶斯公式中用于归一化参数后验分布的分母项，因为

$$p(\mathbf{w} \mid \mathcal{D}, \mathcal{M}_i) = \frac{p(\mathcal{D} \mid \mathbf{w}, \mathcal{M}_i), p(\mathbf{w} \mid \mathcal{M}_i)}{p(\mathcal{D} \mid \mathcal{M}_i)}. \tag{3.69}$$

我们可以通过对参数积分做一个简单近似来理解模型证据。首先考虑只有一个参数 w 的模型。参数的后验分布与 $p(\mathcal{D} \mid w)p(w)$ 成正比，这里为了简化符号省略模型 $\mathcal{M}_i$ 的依赖。若假设后验在其最可能值 w_{MAP} 附近非常尖锐，宽度为 $\Delta w_{\text{posterior}}$，则可将积分近似为取被积函数在峰值处的函数值乘以该峰的宽度。进一步地，若先验是宽度为 Δw_{prior} 的平坦先验，即 $p(w) = 1/\Delta w_{\text{prior}}$，则有

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid w)p(w)\mathrm{d}w \simeq p(\mathcal{D} \mid w_{\text{MAP}}) \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \tag{3.70}$$

取对数后得到

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D} \mid w_{\text{MAP}}) + \ln \left(\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right) \tag{3.71}$$

这一近似在图 3.12 中进行了说明。第一项表示由最可能参数值所给出的对数据的拟合，对于平坦先验来说，这一项就对应对数似然。第二项根据模型的复杂度对模型进行惩罚。由于 $\Delta w_{\text{posterior}} < \Delta w_{\text{prior}}$，因此该项为负，并且随着 $\Delta w_{\text{posterior}}/\Delta w_{\text{prior}}$ 的比例变小而其绝对值增大。于是，如果参数在后验分布中被精细地调整以拟合数据，那么惩罚项将会很大。

对于具有 M 个参数的模型，可以对每个参数做类似的近似。假设所有参数具有相同的比值 $\Delta w_{\text{posterior}}/\Delta w_{\text{prior}}$，则有

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D} \mid \mathbf{w}_{\text{MAP}}) + M \ln \left(\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right) \tag{3.72}$$

因此，在这一非常简单的近似中，复杂度惩罚的大小随模型中自适应参数数量 M 成线性增长。当增加模型复杂度时，第一项通常会减少，因为更复杂的模型能够更好地拟合数据；而第二项由于依赖于 M 而增加。由最大证据确定的最优模型复杂度正是这两个相互竞争项之间的折中。稍后我们将基于后验分布的高斯近似给出一个更精细的版本。

我们可以通过考察图 3.13 来进一步理解贝叶斯模型比较，并理解为何边缘似然会偏好中等复杂度的模型。这里的横轴是一维地表示所有可能数据集的空间，因此横轴上的每一点对应一个具体的数据集。我们考虑三个复杂度逐渐增加的模型 $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$。设想以生成模式运行这些模型以产生数据集，然后观察其生成的数据集的分布。由于参数由先验分布 $p(\mathbf{w})$ 控制，并且在目标变量中可能会有噪声，因此任何模型都可以生成多种不同的数据集。要从一个模型生成一个数据集，我们先从先验中选取参数 $\mathbf{w}$，再根据 $p(\mathcal{D} \mid \mathbf{w})$ 生成数据。一个简单模型（例如一阶多项式）具有较小的变化能力，因此会生成彼此相似的数据集，其分布 $p(\mathcal{D})$ 集中在横轴上的一个较小区域。相比之下，一个复杂模型（例如九阶多项式）能够生成种类繁多的数据集，因此其 $p(\mathcal{D})$ 分布覆盖横轴的更大范围。由于分布 $p(\mathcal{D} \mid \mathcal{M}_i)$ 都归一化，因此对于某个特定数据集 $\mathcal{D}_0$，中等复杂度的模型会给出最大的模型证据。直观上来说，简单模型无法很好地拟合数据，而复杂模型则将其预测概率分散到过于广阔的数据集范围中，从而对任何具体数据集都赋予较小的概率。

在贝叶斯模型比较框架中隐含着这样一个假设：生成数据的真实分布包含在所考虑的模型集合之中。在此条件下，可以证明贝叶斯模型比较平均而言会偏好正确模型。为说明这一点，考虑两个模型 $\mathcal{M}_1$ 和 $\mathcal{M}_2$，其中真实模型为 $\mathcal{M}_1$。对于给定的有限数据集，贝叶斯因子有可能会偏向错误的模型。然而，如果对数据集的分布取平均，则得到期望的贝叶斯因子，其形式为

$$\int p(\mathcal{D} \mid \mathcal{M}_1) \ln \frac{p(\mathcal{D} \mid \mathcal{M}_1)}{p(\mathcal{D} \mid \mathcal{M}_2)} \mathrm{d}\mathcal{D} \tag{3.73}$$

其中平均是相对于数据的真实分布进行的。该量是 Kullback-Leibler 散度的一个例子，具有总为正的性质，只有当两个分布相等时该式才为零。因此，平均而言，贝叶斯因子总是偏好正确模型。

我们已经看到，贝叶斯框架避免了过拟合问题，并允许仅基于训练数据来比较模型。然而，与任何模式识别方法一样，贝叶斯方法也需要对模型形式作出假设，如果这些假设不正确，结果仍可能产生误导。特别地，从图 3.12 中可以看到，模型证据对先验的多个方面（例如尾部行为）可能非常敏感。实际上，如果先验是不适定的，那么证据将无法定义，因为不适定先验具有任意缩放因子（即归一化系数未定义，因为分布无

法归一化）。如果我们考虑一个适定先验，并通过取极限使其成为一个不适定先验（例如令高斯先验的方差趋于无穷大），则证据将趋于零，这可以从 (3.70) 和图 3.12 中看出。然而，可以先计算两个模型之间的证据比值，然后再取极限，以得到有意义的结果。

因此，在实际应用中，明智的做法是保留一份独立的测试数据集，用于评估最终系统的整体性能。

3.5 证据近似

在对线性基函数模型进行完全贝叶斯处理时，我们会为超参数 α 和 β 引入先验分布，并通过对这些超参数以及参数 **w** 进行边缘化来进行预测。然而，尽管我们可以对 **w** 或对超参数中的任一个进行解析积分，但对所有这些变量的完整边缘化在解析上是不可行的。这里我们讨论一种近似方法，在该方法中，我们将超参数设定为特定值，这些值通过最大化先对参数 **w** 积分后得到的边缘似然函数来确定。该框架在统计文献中称为经验贝叶斯（Bernardo and Smith, 1994; Gelman et al., 2004），或类型 2 最大似然（Berger, 1985），或广义最大似然（Wahba, 1975），而在机器学习文献中也称为证据近似（Gull, 1989; MacKay, 1992a）。

如果我们为 α 和 β 引入超先验，则预测分布通过对 $\mathbf{w}, \alpha, \beta$ 进行边缘化得到，因此

$$p(t \mid \mathbf{t}) = \iiint p(t \mid \mathbf{w}, \beta), p(\mathbf{w} \mid \mathbf{t}, \alpha, \beta), p(\alpha, \beta \mid \mathbf{t}), \mathrm{d}\mathbf{w}, \mathrm{d}\alpha, \mathrm{d}\beta \tag{3.74}$$

其中 $p(t \mid \mathbf{w}, \beta)$ 由 (3.8) 给出，而 $p(\mathbf{w} \mid \mathbf{t}, \alpha, \beta)$ 由 (3.49) 给出，其中 $\mathbf{m} * N$ 和 $\mathbf{S} * N$ 分别由 (3.53) 和 (3.54) 定义。这里我们为简化符号省略了对输入变量 **x** 的依赖。

如果后验分布 $p(\alpha, \beta \mid \mathbf{t})$ 在某些值 $\widehat{\alpha}$ 和 $\widehat{\beta}$ 附近高度尖锐，则预测分布可以通过仅对 **w** 进行边缘化并将 α 和 β 固定为 $\widehat{\alpha}$ 和 $\widehat{\beta}$ 来获得，因此

$$p(t \mid \mathbf{t}) \simeq p(t \mid \mathbf{t}, \widehat{\alpha}, \widehat{\beta}) = \int p(t \mid \mathbf{w}, \widehat{\beta}), p(\mathbf{w} \mid \mathbf{t}, \widehat{\alpha}, \widehat{\beta}), \mathrm{d}\mathbf{w} \tag{3.75}$$

根据贝叶斯定理，α 和 β 的后验分布为

$$p(\alpha, \beta \mid \mathbf{t}) \propto p(\mathbf{t} \mid \alpha, \beta), p(\alpha, \beta). \tag{3.76}$$

如果先验相对平坦，则在证据框架中，$\widehat{\alpha}$ 和 $\widehat{\beta}$ 的取值通过最大化边缘似然函数 $p(\mathbf{t} \mid \alpha, \beta)$ 来获得。我们将通过对线性基函数模型的边缘似然进行求值并找到其最大值，从而仅利用训练数据来确定这些超参数，而无需进行交叉验证。回忆一下，比值 α/β 起到了类似正则化参数的作用。

顺便指出，如果我们为 α 和 β 定义共轭先验（Gamma 分布），则在式 (3.74) 中对这些超参数的边缘化可以解析完成，从而在 **w** 上得到一个 Student-t 分布（见 2.3.7 节）。尽管此时对 **w** 的积分不再可解析，但可能会想到用拉普拉斯近似（见 4.4 节），即在后验分布的众数附近用局部高斯近似来逼近该积分（Buntine and Weigend, 1991）。然而，积分作为 **w** 的函数往往具有强烈偏斜的峰形，拉普拉斯近似无法捕获大部分概率质量，从而导致结果反而不如最大化证据的方法（MacKay, 1999）。

回到证据框架，我们可以采取两种方法来最大化对数证据。其一是对证据函数进行解析求导，将其对 α 和 β 的偏导设为零，从而得到重新估计方程，这将在 3.5.2 节中讨论。另一种方法是使用期望最大化（EM）算法，将在 9.3.4 节中讨论，并会展示两种方法收敛到相同的解。

3.5.1 证据函数的求值

边缘似然函数 $p(\mathbf{t} \mid \alpha, \beta)$ 通过对权重参数 $\mathbf{w}$ 积分得到：

$$p(\mathbf{t} \mid \alpha, \beta) = \int p(\mathbf{t} \mid \mathbf{w}, \beta), p(\mathbf{w} \mid \alpha), \mathrm{d}\mathbf{w}. \tag{3.77}$$

一种求解该积分的方法是再次利用线性-高斯模型条件分布的结果 (2.115)。这里我们将改为通过配方（完成平方）的方法来整理指数，并使用高斯分布标准归一化系数的形式来求解。

根据 (3.11)、(3.12) 和 (3.52)，我们可以将证据函数写为

$$p(\mathbf{t} \mid \alpha, \beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \left(\frac{\alpha}{2\pi}\right)^{M/2} \int \exp -E(\mathbf{w}), \mathrm{d}\mathbf{w} \tag{3.78}$$

其中 M 是 $\mathbf{w}$ 的维度，并定义

$$E(\mathbf{w}) = \beta E_D(\mathbf{w}) + \alpha E_W(\mathbf{w}) \quad = \frac{\beta}{2}|\mathbf{t} - \mathbf{\Phi}\mathbf{w}|^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}. \tag{3.79}$$

我们注意到 (3.79) 与正则化平方和误差函数 (3.27) 等价（差一个比例常数）。现在对 $\mathbf{w}$ 完成平方，得到

$$E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^{\mathrm{T}}\mathbf{A}(\mathbf{w} - \mathbf{m}_N) \tag{3.80}$$

这里我们引入了

$$\mathbf{A} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi} \tag{3.81}$$

以及

$$E(\mathbf{m}_N) = \frac{\beta}{2}|\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N|^2 + \frac{\alpha}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N. \tag{3.82}$$

注意，$\mathbf{A}$ 对应误差函数的二阶导数矩阵

$$\mathbf{A} = \nabla\nabla E(\mathbf{w}) \tag{3.83}$$

称为 Hessian 矩阵。这里我们还定义了 $\mathbf{m}_N$，其为

$$\mathbf{m}_N = \beta\mathbf{A}^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}. \tag{3.84}$$

利用 (3.54)，我们有 $\mathbf{A} = \mathbf{S}_N^{-1}$，因此 (3.84) 与先前定义的 (3.53) 等价，从而表示后验分布的均值。

现在对 $\mathbf{w}$ 的积分可以直接利用多元高斯归一化系数的标准结果来计算，得到

$$\begin{aligned}&\int \exp -E(\mathbf{w})\mathrm{d}\mathbf{w}\\&= \exp -E(\mathbf{m}_N)\int \exp\left\{-\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{A}(\mathbf{w}-\mathbf{m}_N)\right\}\mathrm{d}\mathbf{w}\\&= \exp -E(\mathbf{m}_N)(2\pi)^{M/2}|\mathbf{A}|^{-1/2}.\end{aligned} \tag{3.85}$$

利用 (3.78)，我们可以将边缘似然的对数写成

$$\ln p(\mathbf{t}\mid\alpha,\beta)=\frac{M}{2}\ln\alpha+\frac{N}{2}\ln\beta-E(\mathbf{m}_N)-\frac{1}{2}\ln|\mathbf{A}|-\frac{N}{2}\ln(2\pi) \tag{3.86}$$

这就是所需的证据函数表达式。

回到多项式回归问题，我们可以将模型证据随多项式阶数的变化绘制出来，如图 3.14 所示。这里我们假设先验具有 (1.65) 的形式，并将参数 α 固定为 $\alpha=5\times10^{-3}$。这一图形的形状非常具有启发性。回顾图 1.4，可以看到 $M=0$ 的多项式对数据拟合非常差，因此其证据值相对较低。将模型提升到 $M=1$ 时，数据拟合能力大幅提高，因此证据显著增大。然而，从 $M=1$ 增加到 $M=2$ 时，数据拟合仅得到非常微小的改善，这是因为用于生成数据的底层正弦函数是奇函数，因此在多项式展开中没有偶次项。实际上，从图 1.5 可以看到，从 $M=1$ 到 $M=2$ 时残差误差仅略微降低。由于更复杂的模型会带来更高的复杂度惩罚，因此在从 $M=1$ 到 $M=2$ 的过程中证据反而下降。当我们增加到 $M=3$ 时，数据拟合又获得了显著改善，如图 1.4 所示，因此证据再次上升，并在多项式模型中达到最大值。继续增大 M 虽然可以带来微小的拟合改进，但复杂度惩罚增长更快，导致证据整体下降。再看图 1.5，可以发现 $M=3$ 到 $M=8$ 之间的泛化误差大致相当，仅凭该图很难在这些模型间做出选择。而证据值则清晰地偏好 $M=3$，因为这是最简单且能够很好解释观测数据的模型。

3.5.2 证据函数的最大化

首先考虑对 α 最大化 $p(\mathbf{t}\mid\alpha,\beta)$。为此，先定义如下特征向量方程

$$(\beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})\mathbf{u}_i=\lambda_i\mathbf{u}_i. \tag{3.87}$$

由 (3.81) 可知，$\mathbf{A}$ 的特征值为 $\alpha+\lambda_i$。现在考虑 (3.86) 中含有 $\ln|\mathbf{A}|$ 的那一项对 α 的导数。有

$$\frac{d}{d\alpha}\ln|\mathbf{A}|=\frac{d}{d\alpha}\ln\prod_i(\lambda_i+\alpha)=\frac{d}{d\alpha}\sum_i\ln(\lambda_i+\alpha)=\sum_i\frac{1}{\lambda_i+\alpha}. \tag{3.88}$$

因此，(3.86) 关于 α 的驻点满足

$$0=\frac{M}{2\alpha}-\frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N-\frac{1}{2}\sum_i\frac{1}{\lambda_i+\alpha}. \tag{3.89}$$

两边同乘以 2α 并整理，得到

$$\alpha\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N=M-\alpha\sum_i\frac{1}{\lambda_i+\alpha}=\gamma. \tag{3.90}$$

由于对 i 的求和有 M 项，量 γ 可写为

$$\gamma = \sum_i \frac{\lambda_i}{\alpha + \lambda_i}. \tag{3.91}$$

关于 γ 的解释将很快讨论。由 (3.90) 可见，使边缘似然达到最大值的 α 满足

$$\alpha = \frac{\gamma}{\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N}. \tag{3.92}$$

注意，这实际上是关于 α 的隐式解，不仅因为 γ 依赖于 α，而且因为后验分布的众数 $\mathbf{m}_N$ 本身也依赖于 α 的取值。因此，我们采用一种迭代过程：首先为 α 选择一个初始值，并利用 (3.53) 求得 $\mathbf{m}_N$，然后利用 (3.91) 计算 γ。接着利用 (3.92) 重新估计 α，并重复这一过程直到收敛。由于矩阵 $\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$ 是固定的，我们可以在开始时一次性计算出它的特征值，并通过乘以 β 来得到 λ_i。

需要强调的是，α 的取值完全由训练数据决定。与最大似然方法不同，这里不需要使用独立的数据集来优化模型复杂度。

类似地，我们也可以对 β 最大化对数边缘似然 (3.86)。为此，注意到由 (3.87) 定义的特征值 λ_i 与 β 成正比，因此 $d\lambda_i/d\beta = \lambda_i/\beta$，从而得到

$$\frac{d}{d\beta}\ln|\mathbf{A}| = \frac{d}{d\beta}\sum_i \ln(\lambda_i + \alpha) = \frac{1}{\beta}\sum_i \frac{\lambda_i}{\lambda_i + \alpha} = \frac{\gamma}{\beta}. \tag{3.93}$$

因此，边缘似然的驻点满足

$$0 = \frac{N}{2\beta} - \frac{1}{2}\sum_{n=1}^{N} t_n - \mathbf{m}_N^{\mathrm{T}}\phi(\mathbf{x}_n)^2 - \frac{\gamma}{2\beta}. \tag{3.94}$$

整理后得到

$$\frac{1}{\beta} = \frac{1}{N-\gamma}\sum_{n=1}^{N} t_n - \mathbf{m}_N^{\mathrm{T}}\phi(\mathbf{x}_n)^2. \tag{3.95}$$

同样地，这是关于 β 的隐式解，可以通过首先选择 β 的初值，然后利用其计算 $\mathbf{m}_N$ 和 γ，再通过 (3.95) 重新估计 β，重复迭代直到收敛。如果要同时从数据中确定 α 和 β，那么在每次更新 γ 后可以同时重新估计它们。

3.5.3 有效参数个数

结果 (3.92) 有一个优雅的解释 (MacKay, 1992a)，它为理解关于 α 的贝叶斯解提供了洞见。为此，考虑如图 3.15 所示的似然函数和先验的等高线。在这里，我们隐含地将参数空间转换到由 (3.87) 中定义的特征向量 $\mathbf{u}_i$ 对齐的旋转坐标系。在该坐标系中，似然函数的等高线是轴对齐的椭圆。特征值 λ_i 衡量似然函数的曲率，因此在图 3.15 中，特征值 λ_1 相对于 λ_2 较小（因为较小的曲率对应于似然等高线的较大拉伸）。由于 $\beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$ 是一个正定矩阵，它具有正特征值，因此比值 $\lambda_i/(\lambda_i + \alpha)$ 将位于 0 和 1 之间。于是，由 (3.91) 定义的数量 γ 将服从 $0 \leqslant \gamma \leqslant M$。对于满足 $\lambda_i \gg \alpha$ 的方向，相应的参数 w_i 将接

近其最大似然值，并且比值 $\lambda_i/(\lambda_i+\alpha)$ 将接近 1。这样的参数称为确定良好的参数，因为它们的值被数据严格约束。相反，对于满足 $\lambda_i \ll \alpha$ 的方向，相应的参数 w_i 将接近于零，且比值 $\lambda_i/(\lambda_i+\alpha)$ 也接近零。这些方向对应的是似然函数对参数值相对不敏感的情况，因此参数值被先验压缩到较小。由此可知，(3.91) 定义的 γ 衡量了有效的、被良好确定的参数个数之和。

我们通过将 (3.95) 与最大似然情况下的结果 (3.21) 进行比较，可以获得对重新估计 β 结果的进一步理解。这两个公式都将方差（或精度的倒数）表示为目标值与模型预测之间平方差的平均。然而，它们的区别在于，最大似然结果分母中的数据点个数 N 在贝叶斯结果中被 $N-\gamma$ 所取代。我们从 (1.56) 回忆，高斯分布下单变量 x 的最大似然方差估计为

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2 \tag{3.96}$$

并且该估计是有偏的，因为最大似然解 μ_{ML} 对数据中的噪声进行了拟合。本质上，这相当于在模型中使用了一个自由度。对应的无偏估计由 (1.59) 给出，其形式为

$$\sigma^2_{\mathrm{MAP}} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2. \tag{3.97}$$

我们将在第 10.1.3 节中看到，可以通过对未知均值进行边缘化的贝叶斯处理得到这一结果。贝叶斯结果分母中的 $N-1$ 因子考虑到在拟合均值时使用了一个自由度，从而消除了最大似然的偏差。现在考虑线性回归模型的对应结果。此时目标分布的均值由函数 $\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x})$ 给出，它包含 M 个参数。然而，并不是所有这些参数都被数据所调节。由数据决定的有效参数个数为 γ，剩余的 $M-\gamma$ 个参数则因为先验而被压缩到较小的值。这体现在贝叶斯的方差结果中，其分母包含 $N-\gamma$，从而校正了最大似然结果的偏差。

我们可以使用第 1.1 节中的正弦合成数据集，并采用包含 9 个高斯基函数的模型来说明用于设置超参数的证据框架，因此模型中的参数总数（包括偏置）为 $M=10$。在这里，为了简化说明，我们将 β 固定为其真实值 11.1，然后使用证据框架来确定 α，如图 3.16 所示。

我们还可以通过绘制各个参数 w_i 随有效参数个数 γ 的变化情况，来看出超参数 α 是如何控制参数大小的，如图 3.17 所示。

如果考虑极限 $N \gg M$，即数据点数量远大于参数数量，那么根据 (3.87)，所有参数都会被数据很好地确定，因为 $\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$ 中隐含了对数据点的求和，因此特征值 λ_i 会随着数据集的规模增长而变大。在这种情况下，有 $\gamma = M$，于是 α 和 β 的重新估计方程变为

$$\alpha = \frac{M}{2E_W(\mathbf{m}_N)} \tag{3.98}$$

$$\beta = \frac{N}{2E_D(\mathbf{m}_N)} \tag{3.99}$$

其中 E_W 和 E_D 分别由 (3.25) 和 (3.26) 定义。由于这一结果不需要计算 Hessian 的特征值谱，因此可以作为完整证据重估公式的易计算近似。

3.6 固定基函数的局限性

在本章中，我们一直关注由固定的非线性基函数的线性组合构成的模型。我们已经看到，对参数的线性假设带来了许多有用的性质，包括最小二乘问题的闭式解以及可处理的贝叶斯推断过程。此外，对于合适选择的基函数，我们可以对输入变量到目标之间的任意非线性映射进行建模。在下一章中，我们将研究用于分类的一类类似模型。

因此，看起来这类线性模型可以构成解决模式识别问题的通用框架。然而，不幸的是，线性模型存在一些显著的缺点，这将导致我们在后续章节中转向更复杂的模型，比如支持向量机和神经网络。

问题出在这样一个假设：基函数 $\phi_j(\mathbf{x})$ 在观察到训练数据之前就是固定的，而这实际上是第 1.4 节中讨论的维度灾难的一种表现。其结果是，基函数的数量需要随着输入空间维度 D 的增加而迅速增长，通常是指数级增长。

幸运的是，真实数据集具有两个可以利用来缓解该问题的性质。首先，数据向量 $\mathbf{x}_n$ 通常由于输入变量之间存在强相关性，而接近于一个低维的非线性流形，其内在维度小于输入空间的维度。当我们使用局部化基函数时，我们可以将它们仅分布在包含数据的输入空间区域中。这一策略被用于径向基函数网络（RBF networks）、支持向量机（SVM）以及相关向量机（RVM）中。神经网络模型使用具有 S 型非线性的自适应基函数，可以让基函数的作用区域与数据流形相匹配。第二个性质是目标变量可能仅依赖于数据流形中的少数几个方向。神经网络可以通过自动选择基函数响应的输入空间方向来利用这一性质。

4 线性分类模型

在上一章中，我们研究了一类具有特别简单的解析与计算性质的回归模型。现在我们讨论一类用于解决分类问题的类似模型。分类的目标是给定一个输入向量 $\mathbf{x}$，将其归入 K 个离散类别 $\mathcal{C}_k$ 中的某一个，其中 $k = 1, \ldots, K$。最常见的情形中，这些类别是互斥的，因此每个输入只能被分到唯一一个类别。这样输入空间被划分为若干决策区域，其边界称为决策边界或决策面。在本章中，我们考虑线性分类模型，即决策面是输入向量 $\mathbf{x}$ 的线性函数，因此在 D 维输入空间中由 $(D-1)$ 维超平面所定义。若一个数据集的类别可以被线性决策面完全分开，则称该数据集是线性可分的。

在回归问题中，目标变量 $\mathbf{t}$ 是我们希望预测的一组实数；而在分类问题中，有多种方式可以利用目标值来表示类别标签。对于概率模型，在二分类问题中最方便的方法是使用二元表示，即令单一目标变量 $t \in 0, 1$，其中 $t = 1$ 代表类别 $\mathcal{C}_1$，$t = 0$ 代表类别 $\mathcal{C}_2$。我们可以将 t 的取值解释为属于类别 $\mathcal{C}_1$ 的概率，只不过此时概率只能取 0 或 1 的极值。当 $K > 2$ 时，我们使用 1-of-K 编码，即令 $\mathbf{t}$ 为长度为 K 的向量，如果类别为 $\mathcal{C}_j$，则向量中除 $t_j = 1$ 外其余分量 t_k 全部为 0。例如，若 $K = 5$，且样本属于类别 2，则其目标向量为

$$\mathbf{t} = (0, 1, 0, 0, 0)^{\mathrm{T}} \tag{4.1}$$

我们同样可以将 t_k 理解为输入属于类别 $\mathcal{C}_k$ 的概率。而对于非概率模型，有时使用不同形式的类别标签表示会更加便利。

在第 1 章中，我们已经识别出三种不同的分类问题求解方法。最简单的一种是构造一个判别函数，它直接将每个向量 $\mathbf{x}$ 分配到一个特定的类别中。然而，更强大的方法是先在推断阶段对条件概率分布 $p(\mathcal{C}_k \mid \mathbf{x})$ 建模，然后再使用该分布进行最优决策。通过将推断和决策分离，我们可以获得许多好处，正如在 1.5.4 节中所讨论的。对条件概率 $p(\mathcal{C}_k \mid \mathbf{x})$ 有两种不同的处理方法。其中一种技术是直接对其建模，例如将它们表示为参数模型并使用训练集来优化这些参数。另一种方法是采用生成式方法，在该方法中我们对类条件密度 $p(\mathbf{x} \mid \mathcal{C}_k)$ 以及类的先验概率 $p(\mathcal{C}_k)$ 建模，然后使用贝叶斯定理计算所需的后验概率：

$$p(\mathcal{C}_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathcal{C}_k) p(\mathcal{C}_k)}{p(\mathbf{x})}. \tag{4.2}$$

在本章中，我们将讨论这三种方法的例子。

在线性回归模型中（第 3 章中讨论），模型预测 $y(\mathbf{x}, \mathbf{w})$ 由参数 $\mathbf{w}$ 的线性函数给出。在最简单的情况下，模型在输入变量上也是线性的，因此具有形式 $y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$，于是 y 是一个实数。然而，对于分类问题，我们希望预测的是离散的类别标签，或者更一般地说，是位于区间 $(0, 1)$ 的后验概率。为了实现这一点，我们将线性函数通过一个非线性函数 $f(\cdot)$ 进行变换，因此

$$y(\mathbf{x}) = f(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0). \tag{4.3}$$

在机器学习文献中，$f(\cdot)$ 被称为激活函数，而在统计文献中，其逆函数被称为链接函数。决策边界对应于 $y(\mathbf{x}) = \text{constant}$，因此 $\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 = \text{constant}$，于是即使函数 $f(\cdot)$

是非线性的，决策边界仍然是输入 $\mathbf{x}$ 的线性函数。因此，由式（4.3）描述的这一类模型被称为广义线性模型（McCullagh and Nelder, 1989）。然而需要注意的是，与用于回归的模型不同，由于存在非线性函数 $f(\cdot)$，这些模型在参数上不再是线性的。这将导致比线性回归模型更复杂的分析和计算性质。然而，与我们将在后续章节中研究的更一般的非线性模型相比，这些模型仍然相对简单。

本章讨论的算法同样适用于我们首先使用一组基函数 $\phi(\mathbf{x})$ 对输入变量进行固定的非线性变换的情况，就像我们在第 3 章回归模型中所做的那样。我们首先直接在原始输入空间 $\mathbf{x}$ 中考虑分类问题，而在第 4.3 节中，为了与后续章节保持一致，我们将发现使用涉及基函数的符号表示会更加方便。

4.1 判别函数

判别函数是一个将输入向量 $\mathbf{x}$ 分配到 K 个类别之一的函数，这些类别记为 $\mathcal{C}_k$。在本章中，我们将关注线性判别函数，也就是那些决策面为超平面的判别方法。为了简化讨论，我们首先考虑两类的情形，然后再研究扩展到 $K > 2$ 类的情况。

4.1.1 二分类

最简单的线性判别函数形式是对输入向量取线性函数：

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 \tag{4.4}$$

其中 $\mathbf{w}$ 称为权重向量，w_0 称为偏置（不要与统计学意义上的偏差混淆）。偏置的相反数有时也称为阈值。如果对点 $\mathbf{x}$ 有 $y(\mathbf{x}) \geq 0$，则将其判为类别 $\mathcal{C}_1$；否则判为 $\mathcal{C}_2$。因此，决策边界由 $y(\mathbf{x}) = 0$ 定义，对应于 D 维输入空间中的一个 $D-1$ 维超平面。设有两个点 $\mathbf{x}_A$ 和 $\mathbf{x}_B$ 都位于该决策边界上。由于 $y(\mathbf{x}_A) = y(\mathbf{x}_B) = 0$ 因此 $\mathbf{w}^{\mathrm{T}}(\mathbf{x}_A - \mathbf{x}_B) = 0$ 说明向量 $\mathbf{w}$ 与决策超平面内任何向量都正交，因此 $\mathbf{w}$ 决定了决策面的方向。同样地，如果 $\mathbf{x}$ 是决策面上的点，则 $y(\mathbf{x}) = 0$，于是决策面到原点的法向距离为：

$$\frac{\mathbf{w}^{\mathrm{T}}\mathbf{x}}{|\mathbf{w}|} = -\frac{w_0}{|\mathbf{w}|} \tag{4.5}$$

因此偏置 w_0 决定了决策面的平移位置。在 $D = 2$ 的情形中，这些性质如图 4.1 所示。

此外，$y(\mathbf{x})$ 的值还给出了点 $\mathbf{x}$ 到决策面的符号化垂直距离 r。为看到这一点，考虑任意一点 $\mathbf{x}$，并设其在决策面上的正交投影为 $\mathbf{x}_\perp$，则有：

$$\mathbf{x} = \mathbf{x}_\perp + r\frac{\mathbf{w}}{|\mathbf{w}|} \tag{4.6}$$

将该结果两边同时左乘 $\mathbf{w}^{\mathrm{T}}$ 并加上 w_0，并利用 $y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$ 以及 $y(\mathbf{x} * \perp) = \mathbf{w}^{\mathrm{T}}\mathbf{x} * \perp + w_0 = 0$，我们得到

$$r = \frac{y(\mathbf{x})}{|\mathbf{w}|} \tag{4.7}$$

该结果如图 4.1 所示。

如同第 3 章中的线性回归模型一样，有时使用一种更紧凑的记号会更为方便，即我们引入一个额外的虚拟“输入”值 $x_0 = 1$，然后定义 $\widetilde{\mathbf{w}} = (w_0, \mathbf{w})$ 和 $\widetilde{\mathbf{x}} = (x_0, \mathbf{x})$，从而

$$y(\mathbf{x}) = \widetilde{\mathbf{w}}^{\mathrm{T}}\widetilde{\mathbf{x}} \tag{4.8}$$

在这种情况下，决策面是在扩展输入空间中经过原点的 $D+1$ 维超平面，对应于原始 D 维空间中的超平面。

4.1.2 多类别情形

现在考虑将线性判别扩展到 $K > 2$ 个类别的情况。我们可能会倾向于通过组合若干个二分类判别函数来构建一个 K 类判别器。然而，这种做法会导致一些严重的问题（Duda 和 Hart, 1973），如下所示。

考虑使用 $K-1$ 个分类器，每个分类器解决一个将某个特定类别 $\mathcal{C}_k$ 与不属于该类别的点分开的二分类问题。这被称为一对其余（one-versus-the-rest）分类器。图 4.2 左侧的例子展示了一个包含三个类别的情况，在这种方法下输入空间中会出现分类结果不明确的区域。

另一种方法是为每一对可能的类别引入 $K(K-1)/2$ 个二元判别函数。这称为一对一（one-versus-one）分类器。然后根据这些判别函数的多数投票结果将每个点进行分类。然而，这种方法同样会出现分类模糊区域的问题，如图 4.2 右侧图所示。

我们可以通过考虑一个由 K 个线性函数组成的单一 K 类判别来避免这些困难，其形式为

$$y_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0} \tag{4.9}$$

然后对于点 $\mathbf{x}$，如果 $y_k(\mathbf{x}) > y_j(\mathbf{x})$ 对所有 $j \neq k$ 都成立，则将 $\mathbf{x}$ 判为类别 $\mathcal{C}_k$。因此，类别 $\mathcal{C}_k$ 和类别 $\mathcal{C}_j$ 之间的决策边界由 $y_k(\mathbf{x}) = y_j(\mathbf{x})$ 给出，从而对应如下定义的 $(D-1)$ 维超平面

$$-(\mathbf{w}_k - \mathbf{w}_j)^{\mathrm{T}}\mathbf{x} + (w_{k0} - w_{j0}) = 0 \tag{4.10}$$

这与第 4.1.1 节讨论的二分类情况下的决策边界形式相同，因此可以得到类似的几何性质。

这种判别器的决策区域总是单连通且凸的。为说明这一点，考虑两个点 $\mathbf{x}_{\mathrm{A}}$ 和 $\mathbf{x}_{\mathrm{B}}$，它们都位于决策区域 $\mathcal{R}k$ 内，如图 4.3 所示。连接 $\mathbf{x}_{\mathrm{A}}$ 与 $\mathbf{x}_{\mathrm{B}}$ 的直线上的任意点 $\widehat{\mathbf{x}}$ 都可以表示为

$$\widehat{\mathbf{x}} = \lambda\mathbf{x}_{\mathrm{A}} + (1-\lambda)\mathbf{x}_{\mathrm{B}} \tag{4.11}$$

其中 $0 \leqslant \lambda \leqslant 1$。由判别函数的线性性质可知，

$$y_k(\widehat{\mathbf{x}}) = \lambda y_k(\mathbf{x}\mathrm{A}) + (1-\lambda)yk(\mathbf{x}_{\mathrm{B}}) \tag{4.12}$$

由于 $\mathbf{x}_{\mathrm{A}}$ 和 $\mathbf{x}_{\mathrm{B}}$ 都位于 $\mathcal{R}k$ 内，因此对于所有 $j \neq k$，都有 $yk(\mathbf{x}_{\mathrm{A}}) > yj(\mathbf{x}_{\mathrm{A}})$，以及 $yk(\mathbf{x}_{\mathrm{B}}) > yj(\mathbf{x}_{\mathrm{B}})$，于是可得 $yk(\widehat{\mathbf{x}}) > y_j(\widehat{\mathbf{x}})$，从而 $\widehat{\mathbf{x}}$ 也位于 $\mathcal{R}_k$ 内。因此，$\mathcal{R}_k$ 是单连通且凸的。

注意，对于二分类问题，我们可以使用此处基于两个判别函数 $y_1(\mathbf{x})$ 和 $y_2(\mathbf{x})$ 的形式主义，或者使用第 4.1.1 节中基于单一判别函数 $y(\mathbf{x})$ 的更简单但等价的公式。

接下来我们将讨论三种用于学习线性判别函数参数的方法，分别基于最小二乘、Fisher 线性判别以及感知机算法。

4.1.3 用最小二乘法进行分类

在第 3 章中，我们讨论了参数的线性函数模型，并看到通过最小化平方和误差函数可以得到参数值的一个简单的闭式解。因此，我们可能会倾向于尝试将同样的形式主义应用于分类问题。考虑一个具有 K 个类别的一般分类问题，采用 1-of-K 的二元编码方案来表示目标向量 $\mathbf{t}$。在这种情形下使用最小二乘的一种理由是，它近似了给定输入向量后的目标值的条件期望 $\mathbb{E}[\mathbf{t} \mid \mathbf{x}]$。对于二元编码方案，这个条件期望由后验类别概率向量给出。然而，不幸的是，由于线性模型灵活性有限，这些概率通常被近似得很差，实际上这些近似甚至可能超出区间 $(0, 1)$，这一点我们很快就会看到。

每个类别 $\mathcal{C}_k$ 都由其自己的线性模型描述，因此

$$y_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0} \tag{4.13}$$

其中 $k = 1, \ldots, K$。我们可以使用向量表示将它们方便地组合在一起：

$$\mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{x}} \tag{4.14}$$

其中 $\widetilde{\mathbf{W}}$ 是一个矩阵，其第 k 列为 $D+1$ 维向量 $\widetilde{\mathbf{w}}_k = (w_{k0}, \mathbf{w}_k^{\mathrm{T}})^{\mathrm{T}}$，而 $\widetilde{\mathbf{x}}$ 是相应的增广输入向量 $(1, \mathbf{x}^{\mathrm{T}})^{\mathrm{T}}$，其中包含一个虚拟输入 $x_0 = 1$。这种表示法已在第 3.1 节中详细讨论。对新的输入 $\mathbf{x}$，我们将其分配给使输出 $y_k = \widetilde{\mathbf{w}}_k^{\mathrm{T}}\widetilde{\mathbf{x}}$ 最大的类别。

现在我们像第 3 章回归那样，通过最小化平方和误差函数来确定参数矩阵 $\widetilde{\mathbf{W}}$。考虑一个训练数据集 $\{\mathbf{x}_n, \mathbf{t}_n\}$，其中 $n = 1, \ldots, N$，并定义矩阵 $\mathbf{T}$，其第 n 行为向量 $\mathbf{t}_n^{\mathrm{T}}$，以及矩阵 $\widetilde{\mathbf{X}}$，其第 n 行为 $\widetilde{\mathbf{x}}_n^{\mathrm{T}}$。则平方和误差函数可写为

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\,\mathrm{Tr}\left\{(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})^{\mathrm{T}}(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})\right\}. \tag{4.15}$$

将对 $\widetilde{\mathbf{W}}$ 的导数设为零并整理后，可得 $\widetilde{\mathbf{W}}$ 的解为

$$\widetilde{\mathbf{W}} = \left(\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}^{\mathrm{T}}\mathbf{T} = \widetilde{\mathbf{X}}^{\dagger}\mathbf{T} \tag{4.16}$$

其中 $\widetilde{\mathbf{X}}^{\dagger}$ 为矩阵 $\widetilde{\mathbf{X}}$ 的伪逆，见第 3.1.1 节。于是可得判别函数为

$$\mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{x}} = \mathbf{T}^{\mathrm{T}}\left(\tilde{\mathbf{X}}^{\dagger}\right)^{\mathrm{T}}\tilde{\mathbf{x}} \tag{4.17}$$

关于具有多个目标变量的最小二乘解，有一个有趣的性质：如果训练集中每个目标向量都满足某个线性约束

$$\mathbf{a}^{\mathrm{T}}\mathbf{t}_n + b = 0 \tag{4.18}$$

其中 $\mathbf{a}$ 和 b 为常数，那么对于任意 $\mathbf{x}$，模型的预测同样满足该约束，即

$$\mathbf{a}^{\mathrm{T}}\mathbf{y}(\mathbf{x}) + b = 0. \tag{4.19}$$

因此，如果对 K 个类别使用 1-of-K 的编码方案，那么模型给出的预测将具有这样一个性质：对于任意的 $\mathbf{x}$，$\mathbf{y}(\mathbf{x})$ 的各个元素之和都等于 1。然而，仅有这个求和约束还不足以使模型输出能够被解释为概率，因为它们并没有被限制在区间 $(0, 1)$ 内。

最小二乘方法为判别函数参数提供了精确的闭式解。然而，即使将其仅作为判别函数来直接作出分类决策（而不去做概率解释），该方法仍然存在一些严重问题。我们已经看到，最小二乘解对离群点缺乏鲁棒性，这一点在分类应用中同样成立，如图 4.4 所示。从图中可以看到，在右图中额外的数据点使得决策边界发生了显著的变化，尽管这些点在左图中原有的决策边界下本可以被正确分类。平方和误差函数会惩罚“过于正确”的预测，即那些落在决策边界正确一侧且距离较远的点。在第 7.1.2 节中，我们将讨论几种用于分类的替代误差函数，并会看到它们不会出现这种问题。

然而，最小二乘法的问题甚至比缺乏鲁棒性更为严重，如图 4.5 所示。该图展示了一个三类的人工数据集，输入空间为二维 (x_1, x_2)，并且线性决策边界能够很好地分离各类别。实际上，本章后面介绍的逻辑回归方法就能得到很令人满意的结果，如右图所示。然而，最小二乘解却表现不佳，输入空间中只有很小的区域被判为绿色类别。

最小二乘法的失败并不令人意外，因为我们回忆到它对应于假设条件分布为高斯分布下的最大似然解，而二元目标向量显然具有与高斯分布相差甚远的分布特性。通过采用更合适的概率模型，我们将会得到比最小二乘法性质优良得多的分类技术。然而，目前我们将继续探讨用于为线性分类模型设定参数的其他非概率方法。

4.1.4 Fisher 线性判别

一种理解线性分类模型的方法是将其视为一种降维手段。首先考虑两类的情形，假设我们将 D 维输入向量 $\mathbf{x}$ 投影到一维空间：

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x}. \tag{4.20}$$

如果我们对 y 设置一个阈值，将 $y \geqslant -w_0$ 判为类别 $\mathcal{C}_1$，否则判为类别 $\mathcal{C}_2$，那么就得到了上一节讨论的标准线性分类器。一般而言，将数据投影到一维会导致信息的大量丢失，原本在 D 维空间中分离良好的类别在一维上可能会强烈重叠。然而，通过调整权向量 $\mathbf{w}$ 的分量，我们可以选择一个投影，使得类别间的分离程度最大。

首先考虑一个两类问题，其中类别 $\mathcal{C}_1$ 有 N_1 个样本，类别 $\mathcal{C}_2$ 有 N_2 个样本，则两个类别的均值向量为

$$\mathbf{m}_1 = \frac{1}{N_1}\sum_{n\in\mathcal{C}*1}\mathbf{x}_n, \quad \mathbf{m}_2 = \frac{1}{N_2}\sum_{n\in\mathcal{C}*2}\mathbf{x}_n. \tag{4.21}$$

在投影到 $\mathbf{w}$ 后，类别之间最简单的分离度量是投影后的类别均值之差。这提示我们可以选择 $\mathbf{w}$ 来最大化

$$m_2 - m_1 = \mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1), \tag{4.22}$$

其中

$$m_k = \mathbf{w}^{\mathrm{T}}\mathbf{m}_k \tag{4.23}$$

是类别 $\mathcal{C}_k$ 投影后数据的均值。然而，这个表达式可以通过增大 $\mathbf{w}$ 的长度而任意放大。为解决这个问题，我们可以约束 $\mathbf{w}$ 的长度为 1，即 $\sum_i w_i^2 = 1$。使用拉格朗日乘子法进行约束最大化，可得 $\mathbf{w} \propto (\mathbf{m}_2 - \mathbf{m}_1)$。

然而，这种方法仍然存在问题，如图 4.6 所示。图中两个类别在二维空间 (x_1, x_2) 中分离良好，但在投影到连接其均值的直线后出现了明显的重叠。这一问题来自类别分布具有强非对角协方差结构。Fisher 提出的思想是最大化一个函数，该函数不仅使投影后的类别均值分离较大，而且使每个类别内部的投影方差较小，从而减少类别间的重叠。

投影公式 (4.20) 将带标签的数据点集合 $\mathbf{x}$ 映射到一维空间 y 中的带标签数据集合。因此，类别 $\mathcal{C}_k$ 的投影数据的类内方差为

$$s_k^2 = \sum_{n \in \mathcal{C}k} (yn - m_k)^2 \tag{4.24}$$

其中 $y_n = \mathbf{w}^{\mathrm{T}}\mathbf{x}n$。我们可以将整个数据集的类内总方差简单地定义为 $s1^2 + s_2^2$。Fisher 判别准则定义为类间方差与类内方差之比，表示为

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}. \tag{4.25}$$

利用 (4.20)、(4.23) 和 (4.24)，可以将 Fisher 判别准则中对 $\mathbf{w}$ 的依赖显式改写为

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}} \tag{4.26}$$

其中 $\mathbf{S}_{\mathrm{B}}$ 是类间协方差矩阵，给出为

$$\mathbf{S}_{\mathrm{B}} = (\mathbf{m}2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^{\mathrm{T}} \tag{4.27}$$

而 $\mathbf{S}_{\mathrm{W}}$ 是总类内协方差矩阵，给出为

$$\mathbf{S}_{\mathrm{W}} = \sum n \in \mathcal{C}_1(\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathrm{T}} + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^{\mathrm{T}}. \tag{4.28}$$

对 (4.26) 式关于 $\mathbf{w}$ 求导，我们发现当满足以下条件时，$J(\mathbf{w})$ 达到最大值

$$\left(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}\right)\mathbf{S}_{\mathrm{W}}\mathbf{w} = \left(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}\right)\mathbf{S}_{\mathrm{B}}\mathbf{w} \tag{4.29}$$

从 (4.27) 式我们看到，$\mathbf{S}_{\mathrm{B}}\mathbf{w}$ 总是沿着 $(\mathbf{m}_2 - \mathbf{m}_1)$ 的方向。此外，我们不关心 $\mathbf{w}$ 的大小，只关心它的方向，因此我们可以去掉标量因子 $(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w})$ 和 $(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w})$。将 (4.29) 式的两边乘以 $\mathbf{S}_{\mathrm{W}}^{-1}$，我们得到

$$\mathbf{w} \propto \mathbf{S}_{\mathrm{W}}^{-1}(\mathbf{m}_2 - \mathbf{m}_1) \tag{4.30}$$

注意，如果类内协方差是各向同性的，即 $\mathbf{S}_{\mathbf{W}}$ 与单位矩阵成比例，我们发现 $\mathbf{w}$ 与类均值之差成比例，这与我们前面讨论的一致。

结果 (4.30) 被称为费舍尔线性判别 (Fisher's linear discriminant)，尽管严格来说它不是一个判别式，而是将数据投影到一维的特定方向的选择。然而，投影后的数据随后可以用于构建一个判别式，方法是选择一个阈值 y_0，使得如果 $y(\mathbf{x}) \geqslant y_0$，我们将一个新点分类为属于 $\mathcal{C}_1$，否则分类为属于 $\mathcal{C}_2$。例如，我们可以使用高斯分布来建模类条件概率密度 $p(y \mid \mathcal{C}_k)$，然后使用第 1.2.4 节的技术通过最大似然法来找到高斯分布的参数。在找到投影后的类的高斯近似后，第 1.5.1 节的形式主义随后给出了最优阈值的表达式。对于高斯假设的一些佐证来自于中心极限定理，因为 $y = \mathbf{w}^{\mathrm{T}}\mathbf{x}$ 是一组随机变量的和。

4.1.5 与最小二乘法的关系

确定线性判别函数的最小二乘法基于的目标是使模型预测尽可能接近一组目标值。相比之下，费舍尔准则的推导要求在输出空间中实现最大的类别可分离性。观察这两种方法之间的关系是很有趣的。特别是，我们将证明，对于两类问题，费舍尔准则可以作为最小二乘法的一个特例获得。

到目前为止，我们已经考虑了目标值的 1-of-K 编码。然而，如果我们采用略微不同的目标编码方案，那么权值的最小二乘解就等价于费舍尔解（Duda and Hart, 1973）。具体来说，我们将 $\mathcal{C}_1$ 类的目标值设为 N/N_1，其中 N_1 是 $\mathcal{C}_1$ 类中的模式数量，N 是模式总数。这个目标值近似于 $\mathcal{C}_1$ 类先验概率的倒数。对于 $\mathcal{C}_2$ 类，我们将目标值设为 $-N/N_2$，其中 N_2 是 $\mathcal{C}_2$ 类中的模式数量。

平方误差函数可以写成

$$E = \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0 - t_n\right)^2 \tag{4.31}$$

将 E 对 w_0 和 $\mathbf{w}$ 的导数设为零，我们分别得到

$$\sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0 - t_n\right) = 0 \tag{4.32}$$

$$\sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0 - t_n\right)\mathbf{x}_n = 0 \tag{4.33}$$

根据 (4.32) 式，并利用我们对 t_n 的目标编码方案的选择，我们得到了偏置项的形式

$$w_0 = -\mathbf{w}^{\mathrm{T}}\mathbf{m} \tag{4.34}$$

其中我们使用了

$$\sum_{n=1}^{N} t_n = N_1\frac{N}{N_1} - N_2\frac{N}{N_2} = 0 \tag{4.35}$$

并且 $\mathbf{m}$ 是总数据集的均值，由下式给出

$$\mathbf{m} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n = \frac{1}{N}\left(N_1\mathbf{m}_1 + N_2\mathbf{m}_2\right) \tag{4.36}$$

经过一些简单的代数运算，并再次利用 t_n 的选择，第二个方程 (4.33) 变为

$$\left(\mathbf{S}_{\mathrm{W}}+\frac{N_1N_2}{N}\mathbf{S}_{\mathrm{B}}\right)\mathbf{w}=N\left(\mathbf{m}_1-\mathbf{m}_2\right) \tag{4.37}$$

其中 $\mathbf{S}_{\mathrm{W}}$ 由 (4.28) 式定义，$\mathbf{S}_{\mathrm{B}}$ 由 (4.27) 式定义，并且我们已经用 (4.34) 式代入了偏置项。利用 (4.27) 式，我们注意到 $\mathbf{S}_{\mathrm{B}}\mathbf{w}$ 总是沿着 $(\mathbf{m}_2-\mathbf{m}_1)$ 的方向。因此，我们可以写成

$$\mathbf{w}\propto\mathbf{S}_{\mathrm{W}}^{-1}\left(\mathbf{m}_2-\mathbf{m}_1\right) \tag{4.38}$$

其中我们忽略了不相关的尺度因子。因此，权向量与从费舍尔准则中发现的结果一致。此外，我们还找到了由 (4.34) 式给出的偏置值 w_0 的表达式。这告诉我们，如果 $y(\mathbf{x})=\mathbf{w}^{\mathrm{T}}(\mathbf{x}-\mathbf{m})>0$，一个新的向量 $\mathbf{x}$ 应该被分类为属于 $\mathcal{C}_1$ 类，否则属于 $\mathcal{C}_2$ 类。

4.1.6 费舍尔 (Fisher's) 多类判别式

我们现在考虑将费舍尔判别式推广到 $K>2$ 个类的情形，并且我们假设输入空间 $\mathbf{x}$ 的维数 D 大于类的数量 K。接下来，我们引入 $D'>1$ 个线性“特征” $y_k=\mathbf{w}_k^{\mathrm{T}}\mathbf{x}$，其中 $k=1,\ldots,D'$。这些特征值可以方便地组合起来形成一个向量 $\mathbf{y}$。类似地，权向量集合 $\{\mathbf{w}_k\}$ 可以被视为矩阵 $\mathbf{W}$ 的列，因此

$$\mathbf{y}=\mathbf{W}^{\mathrm{T}}\mathbf{x} \tag{4.39}$$

注意，在 $\mathbf{y}$ 的定义中，我们再次没有包含任何偏置参数。将类内协方差矩阵推广到 K 个类的情况，由 (4.28) 式得出

$$\mathbf{S}_{\mathrm{W}}=\sum_{k=1}^{K}\mathbf{S}_k \tag{4.40}$$

其中

$$\mathbf{S}_k=\sum_{n\in\mathcal{C}_k}\left(\mathbf{x}_n-\mathbf{m}_k\right)\left(\mathbf{x}_n-\mathbf{m}_k\right)^{\mathrm{T}} \tag{4.41}$$

$$\mathbf{m}_k=\frac{1}{N_k}\sum_{n\in\mathcal{C}_k}\mathbf{x}_n \tag{4.42}$$

N_k 是 $\mathcal{C}_k$ 类中的模式数量。为了找到类间协方差矩阵的推广形式，我们遵循 Duda 和 Hart (1973) 的做法，首先考虑总协方差矩阵

$$\mathbf{S}_{\mathrm{T}}=\sum_{n=1}^{N}\left(\mathbf{x}_n-\mathbf{m}\right)\left(\mathbf{x}_n-\mathbf{m}\right)^{\mathrm{T}} \tag{4.43}$$

其中 $\mathbf{m}$ 是总数据集的均值

$$\mathbf{m}=\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n=\frac{1}{N}\sum_{k=1}^{K}N_k\mathbf{m}_k \tag{4.44}$$

$N = \sum_k N_k$ 是数据的总点数。总协方差矩阵可以分解为由 (4.40) 和 (4.41) 给出的类内协方差矩阵 $\mathbf{S}_{\mathrm{W}}$ 加上一个附加矩阵 $\mathbf{S}_{\mathrm{B}}$ 的和，我们将 $\mathbf{S}_{\mathrm{B}}$ 视作类间协方差的度量

$$\mathbf{S}_{\mathrm{T}} = \mathbf{S}_{\mathrm{W}} + \mathbf{S}_{\mathrm{B}} \tag{4.45}$$

其中

$$\mathbf{S}_{\mathrm{B}} = \sum_{k=1}^{K} N_k \left(\mathbf{m}_k - \mathbf{m}\right)\left(\mathbf{m}_k - \mathbf{m}\right)^{\mathrm{T}} \tag{4.46}$$

这些协方差矩阵是在原始的 **x**-空间中定义的。我们现在可以在投影后的 D' 维 **y**-空间中定义类似的矩阵

$$\mathbf{s}_{\mathrm{W}} = \sum_{k=1}^{K} \sum_{n \in \mathcal{C}_k} \left(\mathbf{y}_n - \boldsymbol{\mu}_k\right)\left(\mathbf{y}_n - \boldsymbol{\mu}_k\right)^{\mathrm{T}} \tag{4.47}$$

和

$$\mathbf{s}_{\mathrm{B}} = \sum_{k=1}^{K} N_k \left(\boldsymbol{\mu}_k - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}_k - \boldsymbol{\mu}\right)^{\mathrm{T}} \tag{4.48}$$

其中

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{y}_n, \quad \boldsymbol{\mu} = \frac{1}{N} \sum_{k=1}^{K} N_k \boldsymbol{\mu}_k \tag{4.49}$$

我们再次希望构造一个标量，当类间协方差大而类内协方差小时，该标量也大。现在有许多可能的判别准则可供选择 (Fukunaga, 1990)。其中一个例子是

$$J(\mathbf{W}) = \mathrm{Tr}\left\{\mathbf{s}_{\mathbf{W}}^{-1}\mathbf{s}_{\mathrm{B}}\right\} \tag{4.50}$$

然后，这个判别准则可以重写为投影矩阵 **W** 的显式函数，形式为

$$J(\mathbf{W}) = \mathrm{Tr}\left\{\left(\mathbf{W}\mathbf{S}_{\mathbf{W}}\mathbf{W}^{\mathrm{T}}\right)^{-1}\left(\mathbf{W}\mathbf{S}_{\mathrm{B}}\mathbf{W}^{\mathrm{T}}\right)\right\}. \tag{4.51}$$

最大化这类准则是直截了当的，尽管涉及一些复杂的步骤，Fukunaga (1990) 对此进行了详细讨论。权值 **W** 由 $\mathbf{S}_{\mathrm{W}}^{-1}\mathbf{S}_{\mathrm{B}}$ 的对应于 D' 个最大特征值的特征向量来确定。

所有这类准则都有一个重要的共同结果，值得强调。我们首先从 (4.46) 式注意到，$\mathbf{S}_{\mathrm{B}}$ 由 K 个矩阵的和组成，其中每个矩阵都是两个向量的外积，因此秩为 1。此外，由于约束 (4.44) 的结果，这些矩阵中只有 $(K-1)$ 个是独立的。因此，$\mathbf{S}_{\mathrm{B}}$ 的秩至多等于 $(K-1)$，从而至多只有 $(K-1)$ 个非零特征值。这表明，投影到由 $\mathbf{S}_{\mathrm{B}}$ 的特征向量张成的 $(K-1)$ 维子空间上并不会改变 $J(\mathbf{W})$ 的值，因此通过这种方法，我们最多只能找到 $(K-1)$ 个线性“特征”(Fukunaga, 1990)。

4.1.7 感知器算法

线性判别模型的另一个例子是 Rosenblatt (1962) 的感知器 (perceptron)，它在模式识别算法的历史中占有重要地位。它对应于一个两类模型，其中输入向量 **x** 首先通过一个

固定的非线性变换被转换成特征向量 $\boldsymbol{\phi}(\mathbf{x})$，然后用它来构建一个广义线性模型，其形式为

$$y(\mathbf{x}) = f\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right) \tag{4.52}$$

其中非线性激活函数 $f(\cdot)$ 由一个阶梯函数给出，其形式为

$$f(a) = \begin{cases} +1, & a \geqslant 0 \\ -1, & a < 0. \end{cases} \tag{4.53}$$

向量 $\boldsymbol{\phi}(\mathbf{x})$ 通常会包含一个偏置分量 $\phi_0(\mathbf{x}) = 1$。在之前关于两类分类问题的讨论中，我们主要关注于目标编码方案 $t \in \{0, 1\}$，这在概率模型的背景下是合适的。然而，对于感知器，使用 $\mathcal{C}_1$ 类的目标值 $t = +1$ 和 $\mathcal{C}_2$ 类的目标值 $t = -1$ 更为方便，这与激活函数的选择相匹配。

用于确定感知器参数 $\mathbf{w}$ 的算法可以最容易地通过误差函数的最小化来推导。误差函数的一个自然选择可能是被错误分类的模式总数。然而，这不会导致一个简单的学习算法，因为误差是 $\mathbf{w}$ 的分段常数函数，在 $\mathbf{w}$ 的任何变化导致决策边界穿过某个数据点的地方存在不连续性。因此，不能应用基于使用误差函数梯度来改变 $\mathbf{w}$ 的方法，因为梯度在几乎所有地方都为零。

因此，我们考虑另一种误差函数，称为感知器准则 (perceptron criterion)。为了推导出它，我们注意到我们正在寻找一个权向量 $\mathbf{w}$，使得 $\mathcal{C}_1$ 类中的模式 $\mathbf{x}_n$ 满足 $\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) > 0$，而 $\mathcal{C}_2$ 类中的模式 $\mathbf{x}_n$ 满足 $\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) < 0$。使用 $t \in \{-1, +1\}$ 的目标编码方案，这意味着我们希望所有模式都满足 $\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\, t_n > 0$。感知器准则将零误差与任何正确分类的模式相关联，而对于一个错误分类的模式 $\mathbf{x}_n$，它试图最小化 $-\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\, t_n$。因此，感知器准则由下式给出

$$E_{\mathrm{P}}(\mathbf{w}) = -\sum_{n \in \mathcal{M}} \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n t_n \tag{4.54}$$

其中 $\mathcal{M}$ 表示所有被错误分类的模式的集合。与特定错误分类模式相关的误差贡献是 $\mathbf{w}$ 的线性函数，在模式被错误分类的 $\mathbf{w}$ 空间区域中如此，而在模式被正确分类的区域中为零。因此，总误差函数是分段线性的。

我们现在将随机梯度下降算法应用于这个误差函数。权向量 $\mathbf{w}$ 的变化由下式给出

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_{\mathrm{P}}(\mathbf{w}) = \mathbf{w}^{(\tau)} + \eta \boldsymbol{\phi}_n t_n \tag{4.55}$$

其中 η 是学习率参数，τ 是算法步数的索引。因为如果我们将 $\mathbf{w}$ 乘以一个常数，感知器函数 $y(\mathbf{x}, \mathbf{w})$ 保持不变，所以我们可以在不失一般性的前提下将学习率参数 η 设为 1。请注意，随着权向量在训练过程中演变，被错误分类的模式集合会发生变化。

感知器学习算法有一个简单的解释，如下所述。我们依次循环遍历训练模式，对于每个模式 $\mathbf{x}_n$，我们评估感知器函数 (4.52)。如果模式被正确分类，则权向量保持不变；如果它被错误分类，那么对于 $\mathcal{C}_1$ 类，我们将向量 $\boldsymbol{\phi}(\mathbf{x}_n)$ 添加到当前的权向量估计 $\mathbf{w}$ 上，而对于 $\mathcal{C}_2$ 类，我们从 $\mathbf{w}$ 中减去向量 $\boldsymbol{\phi}(\mathbf{x}_n)$。图 4.7 说明了感知器学习算法。

如果我们考虑感知器学习算法中单次更新的影响，我们会看到一个错误分类模式对误差的贡献将会减少，因为根据 (4.55) 式，我们有

$$-\mathbf{w}^{(\tau+1)\mathrm{T}}\boldsymbol{\phi}_n t_n = -\mathbf{w}^{(\tau)\mathrm{T}}\boldsymbol{\phi}_n t_n - \left(\boldsymbol{\phi}_n t_n\right)^{\mathrm{T}}\boldsymbol{\phi}_n t_n < -\mathbf{w}^{(\tau)\mathrm{T}}\boldsymbol{\phi}_n t_n \tag{4.56}$$

其中我们已设置 $\eta = 1$，并使用了 $\|\boldsymbol{\phi}_n t_n\|^2 > 0$。当然，这并不意味着其他被错误分类的模式对误差函数的贡献也会减少。此外，权向量的改变可能会导致一些先前正确分类的模式变得错误分类。因此，感知器学习规则并不能保证在每一步都减少总误差函数。

然而，感知器收敛定理表明，如果存在一个精确解（换句话说，如果训练数据集是线性可分的），那么感知器学习算法保证能在有限步数内找到一个精确解。该定理的证明可以在例如 Rosenblatt (1962)、Block (1962)、Nilsson (1965)、Minsky and Papert (1969)、Hertz et al. (1991) 和 Bishop (1995a) 中找到。但需要注意的是，达到收敛所需的步数可能仍然相当多，而且在实际应用中，在达到收敛之前，我们将无法区分一个不可分问题和一个只是收敛缓慢的问题。

即使数据集是线性可分的，也可能存在许多解，找到哪一个解将取决于参数的初始化以及数据点的呈现顺序。此外，对于非线性可分的数据集，感知器学习算法将永远不会收敛。

除了学习算法上的困难之外，感知器不提供概率输出，也不能轻易地推广到 $K > 2$ 个类别。然而，其最主要的局限性在于（与本章和前一章讨论的所有模型一样），它是基于固定基函数的线性组合。关于感知器局限性的更详细讨论可以在 Minsky and Papert (1969) 和 Bishop (1995a) 中找到。

Rosenblatt 制造了感知器的模拟硬件实现，基于电机驱动的可变电阻器来实现自适应参数 w_j。这些在图 4.8 中有所展示。输入来自一个基于光电传感器阵列的简单摄像系统，而基函数 ϕ 可以通过各种方式选择，例如基于输入图像中随机选择的像素子集的简单固定函数。典型的应用包括学习区分简单的形状或字符。

在感知器被开发的同时，Widrow 及其同事正在探索一个密切相关的系统，称为 adaline，它是“adaptive linear element”（自适应线性元件）的缩写。该模型的函数形式与感知器相同，但采用了不同的训练方法 (Widrow and Hoff, 1960; Widrow and Lehr, 1990)。

4.2 概率生成模型

接下来，我们转向分类的概率观点，并展示具有线性决策边界的模型是如何从对数据分布的简单假设中产生的。在第 1.5.4 节中，我们讨论了判别式方法和生成式方法在分类上的区别。在这里，我们将采用一种生成式方法，其中我们对类条件概率密度 $p\left(\mathbf{x} \mid \mathcal{C}_k\right)$ 和类先验概率 $p\left(\mathcal{C}_k\right)$ 进行建模，然后使用贝叶斯定理计算后验概率 $p\left(\mathcal{C}_k \mid \mathbf{x}\right)$。

首先考虑两类的情况。$\mathcal{C}_1$ 类的后验概率可以写成

$$\begin{aligned} p\left(\mathcal{C}_1 \mid \mathbf{x}\right) &= \frac{p\left(\mathbf{x} \mid \mathcal{C}_1\right) p\left(\mathcal{C}_1\right)}{p\left(\mathbf{x} \mid \mathcal{C}_1\right) p\left(\mathcal{C}_1\right) + p\left(\mathbf{x} \mid \mathcal{C}_2\right) p\left(\mathcal{C}_2\right)} \\ &= \frac{1}{1+\exp(-a)} = \sigma(a) \end{aligned} \tag{4.57}$$

其中我们定义了

$$a = \ln \frac{p(\mathbf{x} \mid \mathcal{C}_1) p(\mathcal{C}_1)}{p(\mathbf{x} \mid \mathcal{C}_2) p(\mathcal{C}_2)} \tag{4.58}$$

$\sigma(a)$ 是由下式定义的逻辑 Sigmoid 函数

$$\sigma(a) = \frac{1}{1 + \exp(-a)} \tag{4.59}$$

它绘制在图 4.9 中。术语“Sigmoid”的意思是 S 形。这种类型的函数有时也被称为“挤压函数”(squashing function)，因为它将整个实轴映射到一个有限区间内。逻辑 Sigmoid 在前面章节中已经遇到过，并在许多分类算法中扮演着重要角色。它满足以下对称性

$$\sigma(-a) = 1 - \sigma(a) \tag{4.60}$$

这很容易验证。逻辑 Sigmoid 的逆函数由下式给出

$$a = \ln \left(\frac{\sigma}{1 - \sigma} \right) \tag{4.61}$$

被称为 logit 函数。它表示两类概率比值的对数 $\ln \left[p(\mathcal{C}_1 \mid \mathbf{x}) / p(\mathcal{C}_2 \mid \mathbf{x}) \right]$，也称为对数几率 (log odds)。

请注意，在 (4.57) 式中，我们只是将后验概率改写成了一种等价的形式，因此逻辑 Sigmoid 的出现可能看起来相当空泛。然而，如果 $a(\mathbf{x})$ 具有简单的函数形式，它将具有重要意义。我们很快会考虑 $a(\mathbf{x})$ 是 $\mathbf{x}$ 的线性函数的情况，在这种情况下，后验概率由一个广义线性模型控制。

对于 $K > 2$ 个类的情况，我们有

$$\begin{aligned} p(\mathcal{C}_k \mid \mathbf{x}) &= \frac{p(\mathbf{x} \mid \mathcal{C}_k) p(\mathcal{C}_k)}{\sum_j p(\mathbf{x} \mid \mathcal{C}_j) p(\mathcal{C}_j)} \\ &= \frac{\exp(a_k)}{\sum_j \exp(a_j)} \end{aligned} \tag{4.62}$$

这被称为归一化指数函数，可以看作是逻辑 Sigmoid 的多类推广。这里，量 a_k 定义为

$$a_k = \ln p(\mathbf{x} \mid \mathcal{C}_k) p(\mathcal{C}_k) \tag{4.63}$$

归一化指数函数也被称为 softmax 函数，因为它代表了'max' 函数的一个平滑版本，因为如果对于所有 $j \neq k$ 都有 $a_k \gg a_j$，那么 $p(\mathcal{C}_k \mid \mathbf{x}) \simeq 1$ 且 $p(\mathcal{C}_j \mid \mathbf{x}) \simeq 0$。

我们现在研究选择特定形式的类条件概率密度的后果，首先考察连续输入变量 $\mathbf{x}$，然后简要讨论离散输入的情况。

4.2.1 连续输入

让我们假设类条件概率密度是高斯分布，然后探究后验概率的最终形式。首先，我们假设所有类共享相同的协方差矩阵。因此，$\mathcal{C}_k$ 类的密度由下式给出

$$p(\mathbf{x} \mid \mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\} \tag{4.64}$$

我们首先考虑两类的情况。根据 (4.57) 和 (4.58) 式，我们有

$$p\left(\mathcal{C}_1 \mid \mathbf{x}\right)=\sigma\left(\mathbf{w}^{\mathrm{T}} \mathbf{x}+w_0\right) \tag{4.65}$$

其中我们定义了

$$\mathbf{w}=\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right) \tag{4.66}$$

$$w_0=-\frac{1}{2} \boldsymbol{\mu}_1^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1+\frac{1}{2} \boldsymbol{\mu}_2^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2+\ln \frac{p\left(\mathcal{C}_1\right)}{p\left(\mathcal{C}_2\right)}. \tag{4.67}$$

我们看到，来自高斯密度指数的 $\mathbf{x}$ 的二次项相互抵消了（由于共同协方差矩阵的假设），从而使得逻辑 Sigmoid 函数的自变量成为 $\mathbf{x}$ 的线性函数。对于二维输入空间 $\mathbf{x}$ 的情况，结果如图 4.10 所示。最终的决策边界对应于后验概率 $p\left(\mathcal{C}_k \mid \mathbf{x}\right)$ 恒定的曲面，因此将由 $\mathbf{x}$ 的线性函数给出，故决策边界在输入空间中是线性的。先验概率 $p\left(\mathcal{C}_k\right)$ 仅通过偏置参数 w_0 参与进来，因此先验概率的变化只会导致决策边界和更一般的等后验概率平行等高线的平行平移。

对于一般 K 类的情况，根据 (4.62) 和 (4.63) 式，我们有

$$a_k(\mathbf{x})=\mathbf{w}_k^{\mathrm{T}} \mathbf{x}+w_{k0} \tag{4.68}$$

其中我们定义了

$$\mathbf{w}_k=\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k \tag{4.69}$$

$$w_{k0}=-\frac{1}{2} \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k+\ln p\left(\mathcal{C}_k\right) \tag{4.70}$$

我们看到，由于共享协方差导致的二次项抵消，$a_k(\mathbf{x})$ 再次是 $\mathbf{x}$ 的线性函数。由此产生的决策边界，对应于最小错误分类率，将出现在两个后验概率（两个最大的概率）相等的地方，因此将由 $\mathbf{x}$ 的线性函数定义，所以我们再次得到了一个广义线性模型。

如果放松共享协方差矩阵的假设，允许每个类条件概率密度 $p\left(\mathbf{x} \mid \mathcal{C}_k\right)$ 拥有自己的协方差矩阵 $\boldsymbol{\Sigma}_k$，那么之前的抵消将不再发生，我们将得到 $\mathbf{x}$ 的二次函数，从而产生一个二次判别式。线性判别边界和二次判别边界如图 4.11 所示。

4.2.2 最大似然解

一旦我们为类条件概率密度 $p\left(\mathbf{x} \mid \mathcal{C}_k\right)$ 指定了参数化的函数形式，我们就可以使用最大似然法来确定参数的值，以及先验类概率 $p\left(\mathcal{C}_k\right)$。这需要一个包含 $\mathbf{x}$ 观测值及其相应类标签的数据集。

首先考虑两类的情况，每类都有一个共享协方差矩阵的高斯类条件概率密度，假设我们有一个数据集 $\left\{\mathbf{x}_n, t_n\right\}$，其中 $n=1, \ldots, N$。这里 $t_n=1$ 表示 $\mathcal{C}_1$ 类，$t_n=0$ 表示 $\mathcal{C}_2$ 类。我们用 $p\left(\mathcal{C}_1\right)=\pi$ 来表示先验类概率，因此 $p\left(\mathcal{C}_2\right)=1-\pi$。对于来自 $\mathcal{C}_1$ 类的数据点 $\mathbf{x}_n$，我们有 $t_n=1$，因此

$$p\left(\mathbf{x}_n, \mathcal{C}_1\right)=p\left(\mathcal{C}_1\right) p\left(\mathbf{x}_n \mid \mathcal{C}_1\right)=\pi \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}\right)$$

类似地，对于 $\mathcal{C}_2$ 类，我们有 $t_n = 0$，因此

$$p(\mathbf{x}_n, \mathcal{C}_2) = p(\mathcal{C}_2)\, p(\mathbf{x}_n \mid \mathcal{C}_2) = (1-\pi)\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

因此，似然函数由下式给出

$$p(\mathbf{t} \mid \pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \left[\pi \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma})\right]^{t_n} \left[(1-\pi)\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma})\right]^{1-t_n} \tag{4.71}$$

其中 $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$。像往常一样，最大化似然函数的对数是很方便的。首先考虑关于 π 的最大化。对数似然函数中依赖于 π 的项是

$$\sum_{n=1}^{N} \{t_n \ln \pi + (1-t_n) \ln(1-\pi)\} \tag{4.72}$$

将关于 π 的导数设为零并重新整理，我们得到

$$\pi = \frac{1}{N}\sum_{n=1}^{N} t_n = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2} \tag{4.73}$$

其中 N_1 表示 $\mathcal{C}_1$ 类中数据点的总数，N_2 表示 $\mathcal{C}_2$ 类中数据点的总数。因此，π 的最大似然估计正如预期的那样，就是 $\mathcal{C}_1$ 类中的点所占的比例。这个结果很容易推广到多类情况，其中 $\mathcal{C}_k$ 类先验概率的最大似然估计同样是由分配给该类的训练集点的比例给出。

现在考虑关于 $\boldsymbol{\mu}_1$ 的最大化。同样，我们可以从对数似然函数中挑出依赖于 $\boldsymbol{\mu}_1$ 的项，得到

$$\sum_{n=1}^{N} t_n \ln \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = -\frac{1}{2}\sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) + \text{ const.} \tag{4.74}$$

将关于 $\boldsymbol{\mu}_1$ 的导数设为零并重新整理，我们得到

$$\boldsymbol{\mu}_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n \mathbf{x}_n \tag{4.75}$$

这仅仅是分配给 $\mathcal{C}_1$ 类的所有输入向量 $\mathbf{x}_n$ 的均值。通过类似的论证，$\boldsymbol{\mu}_2$ 的相应结果由下式给出

$$\boldsymbol{\mu}_2 = \frac{1}{N_2}\sum_{n=1}^{N} (1-t_n)\, \mathbf{x}_n \tag{4.76}$$

这同样是分配给 $\mathcal{C}_2$ 类的所有输入向量 $\mathbf{x}_n$ 的均值。

最后，考虑共享协方差矩阵 $\boldsymbol{\Sigma}$ 的最大似然解。从对数似然函数中挑出依赖于 $\boldsymbol{\Sigma}$ 的项，我们有

$$\begin{aligned} &-\frac{1}{2}\sum_{n=1}^{N} t_n \ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) \\ &-\frac{1}{2}\sum_{n=1}^{N} (1-t_n) \ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} (1-t_n) (\mathbf{x}_n - \boldsymbol{\mu}_2)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_2) \\ &= -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{N}{2}\mathrm{Tr}\left\{\boldsymbol{\Sigma}^{-1}\mathbf{S}\right\} \end{aligned} \tag{4.77}$$

其中我们定义了

$$\mathbf{S} = \frac{N_1}{N}\mathbf{S}_1 + \frac{N_2}{N}\mathbf{S}_2 \tag{4.78}$$

$$\mathbf{S}_1 = \frac{1}{N_1}\sum_{n\in\mathcal{C}_1}(\mathbf{x}_n - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_1)^{\mathrm{T}} \tag{4.79}$$

$$\mathbf{S}_2 = \frac{1}{N_2}\sum_{n\in\mathcal{C}_2}(\mathbf{x}_n - \boldsymbol{\mu}_2)(\mathbf{x}_n - \boldsymbol{\mu}_2)^{\mathrm{T}} \tag{4.80}$$

利用高斯分布最大似然解的标准结果，我们看到 $\boldsymbol{\Sigma} = \mathbf{S}$，这代表了与两个类分别相关的协方差矩阵的加权平均。

这一结果很容易扩展到 K 类问题，以获得相应参数的最大似然解，其中每个类条件密度都是具有共享协方差矩阵的高斯分布。请注意，用高斯分布拟合类别的方法对异常值不鲁棒，因为高斯分布的最大似然估计本身就不是鲁棒的。

4.2.3 离散特征

现在让我们考虑离散特征值 x_i 的情况。为简单起见，我们首先考察二值特征值 $x_i \in \{0, 1\}$，稍后讨论推广到更一般的离散特征。如果有 D 个输入，那么一个一般的分布将对应于每个类的一个 2^D 个数字的表，其中包含 $2^D - 1$ 个独立变量（由于求和约束）。因为这随着特征数量呈指数增长，我们可能会寻求一种限制性更强的表示。这里我们将采用朴素贝叶斯假设，其中特征值在给定类别 $\mathcal{C}_k$ 的条件下被视为独立的。因此，我们有如下形式的类条件分布

$$p(\mathbf{x} \mid \mathcal{C}_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1 - \mu_{ki})^{1-x_i} \tag{4.81}$$

其中，对于每个类别，包含 D 个独立参数。将其代入 (4.63) 式，得到

$$a_k(\mathbf{x}) = \sum_{i=1}^{D}\{x_i \ln \mu_{ki} + (1 - x_i)\ln(1 - \mu_{ki})\} + \ln p(\mathcal{C}_k) \tag{4.82}$$

这再次是输入值 x_i 的线性函数。对于 $K = 2$ 类的情况，我们可以选择考虑由 (4.57) 给出的逻辑 Sigmoid 公式。对于每个可取 $M > 2$ 个状态的离散变量，可以得到类似的结论。

4.2.4 指数族

正如我们所见，对于高斯分布的输入和离散输入，后验类概率都由具有逻辑 Sigmoid ($K = 2$ 类) 或 Softmax ($K \geqslant 2$ 类) 激活函数的广义线性模型给出。这些是一个更一般结果的特例，该结果通过假设类条件概率密度 $p(\mathbf{x} \mid \mathcal{C}_k)$ 是指数族分布的成员而获得。

使用指数族成员的 (2.194) 形式，我们看到 $\mathbf{x}$ 的分布可以写成如下形式：

$$p(\mathbf{x} \mid \boldsymbol{\lambda}_k) = h(\mathbf{x})g(\boldsymbol{\lambda}_k)\exp\left\{\boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}. \tag{4.83}$$

现在我们将注意力限制在满足 $\mathbf{u}(\mathbf{x}) = \mathbf{x}$ 的这类分布的子类上。然后，我们利用 (2.236) 引入一个尺度参数 s，从而得到形式受限的指数族类条件概率密度：

$$p(\mathbf{x} \mid \boldsymbol{\lambda}_k, s) = \frac{1}{s} h\left(\frac{1}{s}\mathbf{x}\right) g(\boldsymbol{\lambda}_k) \exp\left\{\frac{1}{s}\boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{x}\right\}. \tag{4.84}$$

请注意，我们允许每个类别拥有自己的参数向量 $\boldsymbol{\lambda}_k$，但我们假设这些类别共享相同的尺度参数 s。

对于两类问题，我们将类条件概率密度的此表达式代入 (4.58) 式，可以看到后验类概率再次由作用于线性函数 $a(\mathbf{x})$ 的逻辑 Sigmoid 函数给出，该线性函数为

$$a(\mathbf{x}) = (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)^{\mathrm{T}}\mathbf{x} + \ln g(\boldsymbol{\lambda}_1) - \ln g(\boldsymbol{\lambda}_2) + \ln p(\mathcal{C}_1) - \ln p(\mathcal{C}_2). \tag{4.85}$$

类似地，对于 K 类问题，我们将类条件概率密度表达式代入 (4.63) 式，得到

$$a_k(\mathbf{x}) = \boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{x} + \ln g(\boldsymbol{\lambda}_k) + \ln p(\mathcal{C}_k) \tag{4.86}$$

因此，它再次是 $\mathbf{x}$ 的线性函数。

4.3 概率判别模型

对于两类分类问题，我们已经看到，对于广泛选择的类条件分布 $p(\mathbf{x} \mid \mathcal{C}_k)$，$\mathcal{C}_1$ 类的后验概率可以写成作用于 $\mathbf{x}$ 的线性函数的逻辑 Sigmoid 函数。类似地，对于多类情况，$\mathcal{C}_k$ 类的后验概率由 $\mathbf{x}$ 的线性函数的 softmax 变换给出。对于特定选择的类条件概率密度 $p(\mathbf{x} \mid \mathcal{C}_k)$，我们使用最大似然法来确定密度的参数以及类先验 $p(\mathcal{C}_k)$，然后使用贝叶斯定理来找到后验类概率。

然而，另一种方法是明确使用广义线性模型的函数形式，并直接通过最大似然法确定其参数。我们将看到，有一个有效算法可以找到这样的解，称为迭代重加权最小二乘法（iterative reweighted least squares, 或 IRLS）。

通过分别拟合类条件概率密度和类先验，然后应用贝叶斯定理来找到广义线性模型参数的间接方法，是生成式建模的一个例子，因为我们可以利用这样的模型并从边缘分布 $p(\mathbf{x})$ 中抽取 $\mathbf{x}$ 的值来生成合成数据。在直接方法中，我们最大化一个通过条件分布 $p(\mathcal{C}_k \mid \mathbf{x})$ 定义的似然函数，这代表了一种判别式训练形式。判别式方法的一个优点是，通常需要确定的自适应参数会更少，这一点我们很快就会看到。它也可能带来更高的预测性能，特别是当类条件密度假设对真实分布的近似效果较差时。

4.3.1 固定基函数

到目前为止，在本章中我们考虑的分类模型是直接使用原始输入向量 $\mathbf{x}$ 进行工作的。然而，如果我们首先使用基函数向量 $\boldsymbol{\phi}(\mathbf{x})$ 对输入进行一个固定的非线性变换，那么所有这些算法同样适用。由此产生的决策边界在特征空间 $\boldsymbol{\phi}$ 中是线性的，而这些决策边界在原始 $\mathbf{x}$ 空间中对应于非线性决策边界，如图 4.12 所示。在特征空间 $\boldsymbol{\phi}(\mathbf{x})$ 中线性可分的类，在原始观测空间 $\mathbf{x}$ 中不一定是线性可分的。请注意，正如我们在讨论回归

的线性模型时所讨论的，其中一个基函数通常设置为常数，例如 $\phi_0(\mathbf{x}) = 1$，因此对应的参数 w_0 扮演着偏置的角色。在本章的其余部分，我们将包含一个固定的基函数变换 $\boldsymbol{\phi}(\mathbf{x})$，因为这将突出与第 3 章讨论的回归模型的一些有用相似之处。

对于许多实际感兴趣的问题，类条件概率密度 $p(\mathbf{x} \mid \mathcal{C}_k)$ 之间存在显著的重叠。这对应于后验概率 $p(\mathcal{C}_k \mid \mathbf{x})$，对于至少某些 $\mathbf{x}$ 值而言，它们既不是 0 也不是 1。在这种情况下，最优解是通过准确建模后验概率，然后应用标准决策理论获得的，如第 1 章所讨论。请注意，非线性变换 $\boldsymbol{\phi}(\mathbf{x})$ 不能消除这种类别重叠。事实上，它们可能会增加重叠的程度，或者在原始观测空间中不存在重叠的地方创建重叠。然而，适当选择的非线性可以使建模后验概率的过程更容易。

这种固定基函数模型具有重要的局限性，这些局限性将在后面的章节中通过允许基函数本身适应数据来解决。尽管存在这些局限性，带有固定非线性基函数的模型在应用中仍扮演着重要角色，对这些模型的讨论将引入理解其更复杂对应物所需的许多关键概念。

4.3.2 逻辑回归

我们从考虑二分类问题开始，对广义线性模型进行处理。在我们对第 4.2 节中生成方法的讨论中，我们看到在相当一般的假设下，类别 $\mathcal{C}_1$ 的后验概率可以写成作用在特征向量 ϕ 的线性函数上的逻辑 S 形函数，即

$$p(\mathcal{C}_1 \mid \phi) = y(\phi) = \sigma\left(\mathbf{w}^{\mathrm{T}}\phi\right) \tag{4.87}$$

其中 $p(\mathcal{C}_2 \mid \phi) = 1 - p(\mathcal{C}_1 \mid \phi)$。这里的 $\sigma(\cdot)$ 是由 (4.59) 定义的逻辑 S 形函数。在统计学的术语中，该模型被称为逻辑回归，但需要强调的是，这是一个用于分类而非回归的模型。

对于一个 M 维的特征空间 ϕ，该模型有 M 个可调参数。相比之下，如果我们使用最大似然法拟合高斯类别条件密度，我们将为均值使用 $2M$ 个参数，为（共享的）协方差矩阵使用 $M(M+1)/2$ 个参数。再加上类别先验 $p(\mathcal{C}_1)$，参数总数为 $M(M+5)/2+1$ 个，它随 M 二次增长，与逻辑回归中参数数量对 M 的线性依赖形成对比。对于较大的 M 值，直接使用逻辑回归模型具有明显的优势。

我们现在使用最大似然法来确定逻辑回归模型的参数。为此，我们将利用逻辑 S 形函数的导数，它可以方便地用 S 形函数本身来表达

$$\frac{d\sigma}{da} = \sigma(1-\sigma) \tag{4.88}$$

对于一个数据集 $\{\phi_n, t_n\}$，其中 $t_n \in \{0,1\}$ 且 $\phi_n = \phi(\mathbf{x}_n)$，$n = 1, \ldots, N$，似然函数可以写成

$$p(\mathbf{t} \mid \mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} \{1 - y_n\}^{1-t_n} \tag{4.89}$$

其中 $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$ 且 $y_n = p(\mathcal{C}_1 \mid \phi_n)$。像往常一样，我们可以通过取似然的负对数来定义一个误差函数，从而得到以下形式的交叉熵误差函数

$$E(\mathbf{w}) = -\ln p(\mathbf{t} \mid \mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln (1 - y_n)\} \tag{4.90}$$

其中 $y_n = \sigma(a_n)$ 且 $a_n = \mathbf{w}^{\mathrm{T}}\phi_n$。对 $E(\mathbf{w})$ 关于 $\mathbf{w}$ 取梯度，我们得到

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\, \phi_n \tag{4.91}$$

这里我们使用了 (4.88) 的结果。我们看到，涉及逻辑 S 形函数导数的因子被抵消了，从而使得对数似然的梯度形式得以简化。特别是，数据点 n 对梯度的贡献由目标值与模型预测值之间的“误差”$y_n - t_n$ 乘以基函数向量 ϕ_n 给出。此外，与 (3.13) 比较表明，这与线性回归模型的平方和误差函数的梯度具有完全相同的形式。

如果需要，我们可以利用结果 (4.91) 给出一个序列算法，其中模式被逐个呈现，并使用 (3.22) 更新每个权重向量，其中 ∇E_n 是 (4.91) 中的第 n^{th} 项。

值得注意的是，对于线性可分的数据集，最大似然法可能会表现出严重的过拟合。这是因为最大似然解发生在对应于 $\sigma = 0.5$（等价于 $\mathbf{w}^{\mathrm{T}}\phi = 0$）的超平面将两类分开，并且 $\mathbf{w}$ 的幅值趋于无穷大时。在这种情况下，逻辑 S 形函数在特征空间中变得无限陡峭，对应于一个赫维赛德阶跃函数，从而将来自每个类别 k 的每个训练点都分配给后验概率 $p(\mathcal{C}_k \mid \mathbf{x}) = 1$。此外，通常存在一个解的连续体，因为任何分离超平面都会在训练数据点处产生相同的后验概率，正如稍后在图 10.13 中将看到的。最大似然法无法在这些解中偏爱任何一个，实践中找到哪个解将取决于所选的优化算法和参数初始化。请注意，即使数据点的数量相对于模型中参数的数量很大，只要训练数据集是线性可分的，这个问题就会出现。可以通过纳入先验并找到 $\mathbf{w}$ 的 MAP 解，或者等效地通过向误差函数添加正则化项来避免这种奇异性。

4.3.3 迭代重加权最小二乘法

在第 3 章讨论的线性回归模型的情况下，在高斯噪声模型的假设下，最大似然解可得到闭式解。这是因为对数似然函数对参数向量 $\mathbf{w}$ 具有二次依赖性。对于逻辑回归，由于逻辑 S 形函数的非线性，不再存在闭式解。然而，与二次形式的偏离并不大。确切地说，误差函数是凹函数，正如我们稍后将看到的，因此它具有唯一的最小值。此外，可以使用基于牛顿-拉夫逊（Newton-Raphson）迭代优化方案的高效迭代技术来最小化误差函数，该方案使用对数似然函数的局部二次近似。用于最小化函数 $E(\mathbf{w})$ 的牛顿-拉夫逊更新形式为 (Fletcher, 1987; Bishop and Nabney, 2008)

$$\mathbf{w}^{(\text{new})} = \mathbf{w}^{(\text{old})} - \mathbf{H}^{-1}\nabla E(\mathbf{w}) \tag{4.92}$$

其中 $\mathbf{H}$ 是 Hessian 矩阵，其元素包含 $E(\mathbf{w})$ 关于 $\mathbf{w}$ 分量的二阶导数。

让我们首先将牛顿-拉夫逊方法应用于具有平方和误差函数 (3.12) 的线性回归模型 (3.3)。该误差函数的梯度和 Hessian 矩阵由下式给出

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \left(\mathbf{w}^{\mathrm{T}}\phi_n - t_n\right)\phi_n = \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w} - \mathbf{\Phi}^{\mathrm{T}}\mathbf{t} \tag{4.93}$$

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \phi_n\phi_n^{\mathrm{T}} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi} \tag{4.94}$$

其中 $\mathbf{\Phi}$ 是 $N \times M$ 设计矩阵，其第 n^{th} 行由 ϕ_n^{T} 给出。牛顿-拉夫逊更新然后采用以下形式

$$\begin{aligned}\mathbf{w}^{(\text{new})} &= \mathbf{w}^{(\text{old})} - \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\left\{\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w}^{(\text{old})} - \mathbf{\Phi}^{\mathrm{T}}\mathbf{t}\right\} \\ &= \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}\end{aligned} \tag{4.95}$$

我们认识到这就是标准最小二乘解。请注意，在这种情况下，误差函数是二次的，因此牛顿-拉夫逊公式在一步中即可给出精确解。

现在让我们将牛顿-拉夫逊更新应用于逻辑回归模型的交叉熵误差函数 (4.90)。从 (4.91) 我们看到，该误差函数的梯度和 Hessian 矩阵由下式给出

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\phi_n = \mathbf{\Phi}^{\mathrm{T}}(\mathbf{y} - \mathbf{t}) \tag{4.96}$$

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1 - y_n)\phi_n\phi_n^{\mathrm{T}} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi} \tag{4.97}$$

其中我们使用了 (4.88)。此外，我们引入了 $N \times N$ 对角矩阵 $\mathbf{R}$，其元素为

$$R_{nn} = y_n(1 - y_n) \tag{4.98}$$

我们看到 Hessian 矩阵不再是常数，而是通过权重矩阵 $\mathbf{R}$ 依赖于 $\mathbf{w}$，这对应于误差函数不再是二次的这一事实。利用逻辑 S 形函数的形式所带来的性质 $0 < y_n < 1$，我们看到对于任意向量 $\mathbf{u}$，有 $\mathbf{u}^{\mathrm{T}}\mathbf{H}\mathbf{u} > 0$，因此 Hessian 矩阵 $\mathbf{H}$ 是正定的。由此可知，误差函数是 $\mathbf{w}$ 的凹函数，因此具有唯一的最小值。

对于逻辑回归模型，牛顿-拉夫逊更新公式变为

$$\begin{aligned}\mathbf{w}^{(\text{new})} &= \mathbf{w}^{(\text{old})} - \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}(\mathbf{y} - \mathbf{t}) \\ &= \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi}\right)^{-1}\left\{\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi}\mathbf{w}^{(\text{old})} - \mathbf{\Phi}^{\mathrm{T}}(\mathbf{y} - \mathbf{t})\right\} \\ &= \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{z}\end{aligned} \tag{4.99}$$

其中 $\mathbf{z}$ 是一个 N 维向量，其元素为

$$\mathbf{z} = \mathbf{\Phi}\mathbf{w}^{(\text{old})} - \mathbf{R}^{-1}(\mathbf{y} - \mathbf{t}) \tag{4.100}$$

我们看到，更新公式 (4.99) 采用了加权最小二乘问题的正规方程组的形式。由于加权矩阵 $\mathbf{R}$ 不是常数，而是取决于参数向量 $\mathbf{w}$，我们必须迭代地应用正规方程组，每次都使用新的权重向量 $\mathbf{w}$ 来计算修正后的加权矩阵 $\mathbf{R}$。因此，该算法被称为迭代重加权

最小二乘法（iterative reweighted least squares），简称 IRLS (Rubin, 1983)。与加权最小二乘问题中一样，对角加权矩阵 $\mathbf{R}$ 的元素可以解释为方差，因为逻辑回归模型中 t 的均值和方差由下式给出

$$\mathbb{E}[t] = \sigma(\mathbf{x}) = y \tag{4.101}$$

$$\text{var}[t] = \mathbb{E}\left[t^2\right] - \mathbb{E}[t]^2 = \sigma(\mathbf{x}) - \sigma(\mathbf{x})^2 = y(1-y) \tag{4.102}$$

其中我们使用了 $t \in \{0, 1\}$ 时 $t^2 = t$ 的性质。实际上，我们可以将 IRLS 解释为在变量 $a = \mathbf{w}^{\mathrm{T}}\phi$ 空间中线性化问题的解。然后，对应于 $\mathbf{z}$ 的第 n^{th} 个元素的量 z_n 可以在该空间中被简单地解释为一个有效目标值，该值是通过在当前工作点 $\mathbf{w}^{(\text{old})}$ 周围对逻辑 S 形函数进行局部线性近似获得的

$$\begin{aligned} a_n(\mathbf{w}) &\simeq a_n\left(\mathbf{w}^{(\text{old})}\right) + \left.\frac{\mathrm{d}a_n}{\mathrm{d}y_n}\right|_{\mathbf{w}^{(\text{old})}} (t_n - y_n) \\ &= \phi_n^{\mathrm{T}}\mathbf{w}^{(\text{old})} - \frac{(y_n - t_n)}{y_n(1-y_n)} = z_n. \end{aligned} \tag{4.103}$$

4.3.4 多类别逻辑回归

在我们关于多类别分类的生成模型的讨论中，我们已经看到，对于一大类分布，后验概率由特征变量的线性函数的 Softmax 变换给出，因此

$$p(\mathcal{C}_k \mid \phi) = y_k(\phi) = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \tag{4.104}$$

其中“激活”a_k 由下式给出

$$a_k = \mathbf{w}_k^{\mathrm{T}}\boldsymbol{\phi} \tag{4.105}$$

在那里，我们使用最大似然法分别确定类别条件密度和类别先验，然后使用贝叶斯定理找到相应的后验概率，从而隐式地确定了参数 $\{\mathbf{w}_k\}$。在这里，我们考虑使用最大似然法直接确定该模型的参数 $\{\mathbf{w}_k\}$。为此，我们需要 y_k 关于所有激活 a_j 的导数。它们由下式给出

$$\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j) \tag{4.106}$$

其中 I_{kj} 是单位矩阵的元素。接下来我们写出似然函数。这最容易通过使用 1-of-K 编码方案来完成，在该方案中，属于类别 $\mathcal{C}_k$ 的特征向量 ϕ_n 的目标向量 $\mathbf{t}_n$ 是一个二元向量，除了元素 k 等于一之外，所有元素都为零。则似然函数由下式给出

$$p(\mathbf{T} \mid \mathbf{w}_1, \ldots, \mathbf{w}_K) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(\mathcal{C}_k \mid \phi_n)^{t_{nk}} = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}} \tag{4.107}$$

其中 $y_{nk} = y_k(\phi_n)$，而 $\mathbf{T}$ 是一个 $N \times K$ 目标变量矩阵，其元素为 t_{nk}。取负对数则得到

$$E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\ln p(\mathbf{T} \mid \mathbf{w}_1, \ldots, \mathbf{w}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk} \tag{4.108}$$

这被称为多类别分类问题的交叉熵误差函数。

我们现在对误差函数关于其中一个参数向量 $\mathbf{w}_j$ 取梯度。利用 Softmax 函数导数的结果 (4.106)，我们得到

$$\nabla_{\mathbf{w}_j} E\left(\mathbf{w}_1, \ldots, \mathbf{w}_K\right) = \sum_{n=1}^{N} \left(y_{nj} - t_{nj}\right) \phi_n \tag{4.109}$$

其中我们使用了 $\sum_k t_{nk} = 1$ 的性质。我们再次看到与线性模型的平方和误差函数和逻辑回归模型的交叉熵误差所发现的梯度具有相同的形式，即误差 $(y_{nj} - t_{nj})$ 乘以基函数 ϕ_n 的乘积。同样，我们可以使用它来制定一个序列算法，其中模式被逐个呈现，并使用 (3.22) 更新每个权重向量。

我们已经看到，对于数据点 n，线性回归模型的对数似然函数关于参数向量 $\mathbf{w}$ 的导数采用误差 $y_n - t_n$ 乘以特征向量 ϕ_n 的形式。类似地，对于逻辑 S 形激活函数和交叉熵误差函数 (4.90) 的组合，以及 Softmax 激活函数和多类别交叉熵误差函数 (4.108) 的组合，我们再次获得了这种相同的简单形式。正如我们将在第 4.3.6 节中看到的，这是一个更一般结果的例子。

为了找到一个批处理算法，我们再次诉诸牛顿-拉夫逊更新，以获得多类别问题相应的 IRLS 算法。这需要评估 Hessian 矩阵，该矩阵包含大小为 $M \times M$ 的块，其中块 j, k 由下式给出

$$\nabla_{\mathbf{w}_k} \nabla_{\mathbf{w}_j} E\left(\mathbf{w}_1, \ldots, \mathbf{w}_K\right) = -\sum_{n=1}^{N} y_{nk} \left(I_{kj} - y_{nj}\right) \phi_n \phi_n^{\mathrm{T}} \tag{4.110}$$

与两类别问题一样，多类别逻辑回归模型的 Hessian 矩阵是正定的，因此误差函数再次具有唯一的最小值。多类别情况下 IRLS 的实践细节可以在 Bishop and Nabney (2008) 中找到。

4.3.5 概率单位回归（Probit Regression）

我们已经看到，对于指数族描述的广泛类别条件分布，所得的后验类别概率由作用在特征变量的线性函数上的逻辑（或 Softmax）变换给出。然而，并非所有类别条件密度的选择都会产生这种简单的后验概率形式（例如，如果类别条件密度是使用高斯混合建模的）。这表明可能值得探索其他类型的判别概率模型。然而，对于本章的目的，我们将回到两类别情况，并仍停留在广义线性模型的框架内，因此

$$p(t = 1 \mid a) = f(a) \tag{4.111}$$

其中 $a = \mathbf{w}^{\mathrm{T}} \phi$，而 $f(\cdot)$ 是激活函数。

激励连接函数（link function）的替代选择的一种方法是考虑一个带噪声的阈值模型，如下所示。对于每个输入 ϕ_n，我们评估 $a_n = \mathbf{w}^{\mathrm{T}} \phi_n$，然后我们根据以下规则设置目标值

$$\begin{cases} t_n = 1 \text{ if } a_n \geqslant \theta \\ t_n = 0 \text{ otherwise} \end{cases} \tag{4.112}$$

如果 θ 的值是从概率密度 $p(\theta)$ 中抽取的，那么相应的激活函数将由累积分布函数给出

$$f(a) = \int_{-\infty}^{a} p(\theta)\mathrm{d}\theta \tag{4.113}$$

如图 4.13 所示。

作为一个具体的例子，假设密度 $p(\theta)$ 由一个零均值、单位方差的高斯分布给出。相应的累积分布函数由下式给出

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta \mid 0, 1)\mathrm{d}\theta \tag{4.114}$$

这被称为 Probit 函数（probit function）。它具有 S 形，并与逻辑 S 形函数在图 4.9 中进行了比较。请注意，使用更一般的高斯分布不会改变模型，因为这等效于线性系数 **w** 的重新缩放。许多数值软件包提供了对一个密切相关函数的评估，该函数定义为

$$\mathrm{erf}(a) = \frac{2}{\sqrt{\pi}} \int_{0}^{a} \exp\left(-\theta^2/2\right) \mathrm{d}\theta \tag{4.115}$$

被称为 erf 函数或误差函数（不要与机器学习模型的误差函数混淆）。它与概率单位函数的关系为

$$\Phi(a) = \frac{1}{2}\left\{1 + \frac{1}{\sqrt{2}}\,\mathrm{erf}(a)\right\}. \tag{4.116}$$

基于 Probit 激活函数的广义线性模型被称为 Probit 回归。

我们可以使用最大似然法来确定该模型的参数，只需对前面讨论的思想进行直接扩展即可。实际上，使用概率单位回归发现的结果往往与逻辑回归的结果相似。然而，当我们在第 4.5 节讨论逻辑回归的贝叶斯处理时，我们将发现概率单位模型的另一种用途。

在实际应用中可能出现的一个问题是异常值（outliers），它可能例如通过输入向量 **x** 的测量误差或目标值 t 的错误标记而产生。由于这些点可能位于理想决策边界的错误一侧很远的地方，它们会严重扭曲分类器。请注意，逻辑回归模型和概率单位回归模型在这方面表现不同，因为对于 $x \to \infty$，逻辑 S 形的尾部以 $\exp(-x)$ 的形式渐近衰减，而对于概率单位激活函数，它们以 $\exp\left(-x^2\right)$ 的形式衰减，因此概率单位模型可能对异常值显著更敏感。

然而，逻辑模型和概率单位模型都假设数据被正确标记。错误标记的影响很容易纳入概率模型中，方法是引入一个概率 ϵ，表示目标值 t 被翻转到错误值（Opper and Winther, 2000a），从而导致数据点 **x** 的目标值分布形式为

$$\begin{aligned} p(t \mid \mathbf{x}) &= (1-\epsilon)\sigma(\mathbf{x}) + \epsilon(1-\sigma(\mathbf{x})) \\ &= \epsilon + (1-2\epsilon)\sigma(\mathbf{x}) \end{aligned} \tag{4.117}$$

其中 $\sigma(\mathbf{x})$ 是输入向量 **x** 的激活函数。这里的 ϵ 可以预先设定，也可以被视为一个超参数，其值从数据中推断出来。

4.3.6 典则连接函数

对于具有高斯噪声分布的线性回归模型，对应于负对数似然的误差函数由 (3.12) 给出。如果我们对误差函数中来自数据点 n 的贡献关于参数向量 $\mathbf{w}$ 取导数，它采用误差 $y_n - t_n$ 乘以特征向量 ϕ_n 的形式，其中 $y_n = \mathbf{w}^{\mathrm{T}}\phi_n$。类似地，对于逻辑 S 形激活函数和交叉熵误差函数 (4.90) 的组合，以及 Softmax 激活函数和多类别交叉熵误差函数 (4.108) 的组合，我们再次获得了这种相同的简单形式。我们现在证明这是一个一般性结果，即假设目标变量的条件分布来自指数族，并伴随着激活函数的相应选择，该选择被称为典则连接函数。

我们再次使用指数族分布的受限形式 (4.84)。请注意，这里我们将指数族分布的假设应用于目标变量 t，这与第 4.2.4 节中将其应用于输入向量 $\mathbf{x}$ 形成对比。因此，我们考虑目标变量的条件分布形式为

$$p(t \mid \eta, s) = \frac{1}{s} h\left(\frac{t}{s}\right) g(\eta) \exp\left\{\frac{\eta t}{s}\right\}. \tag{4.118}$$

使用导致推导出结果 (2.226) 的相同推理思路，我们看到 t 的条件均值，我们将其表示为 y，由下式给出

$$y \equiv \mathbb{E}[t \mid \eta] = -s\frac{d}{d\eta} \ln g(\eta) \tag{4.119}$$

因此 y 和 η 必须相关，我们通过 $\eta = \psi(y)$ 来表示这种关系。

遵循 Nelder 和 Wedderburn (1972) 的定义，我们广义线性模型定义为 y 是输入（或特征）变量的线性组合的非线性函数的模型，因此

$$y = f\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\right) \tag{4.120}$$

其中 $f(\cdot)$ 在机器学习文献中被称为激活函数，而 $f^{-1}(\cdot)$ 在统计学中被称为连接函数（link function）。

现在考虑该模型的对数似然函数，它作为 η 的函数，由下式给出

$$\ln p(\mathbf{t} \mid \eta, s) = \sum_{n=1}^{N} \ln p\left(t_n \mid \eta, s\right) = \sum_{n=1}^{N} \left\{\ln g\left(\eta_n\right) + \frac{\eta_n t_n}{s}\right\} + \text{ const} \tag{4.121}$$

其中我们假设所有观测值共享一个共同的尺度参数（例如，对于高斯分布，它对应于噪声方差），因此 s 独立于 n。对数似然关于模型参数 $\mathbf{w}$ 的导数然后由下式给出

$$\begin{aligned} \nabla_{\mathbf{w}} \ln p(\mathbf{t} \mid \eta, s) &= \sum_{n=1}^{N} \left\{\frac{d}{d\eta_n} \ln g\left(\eta_n\right) + \frac{t_n}{s}\right\} \frac{d\eta_n}{dy_n} \frac{dy_n}{da_n} \nabla a_n \\ &= \sum_{n=1}^{N} \frac{1}{s} \left\{t_n - y_n\right\} \psi'\left(y_n\right) f'\left(a_n\right) \phi_n \end{aligned} \tag{4.122}$$

其中 $a_n = \mathbf{w}^{\mathrm{T}}\phi_n$，并且我们使用了 $y_n = f\left(a_n\right)$ 以及 $\mathbb{E}[t \mid \eta]$ 的结果 (4.119)。我们现在看到，如果我们对连接函数 $f^{-1}(y)$ 选择由下式给出的特定形式，就会有极大的简化

$$f^{-1}(y) = \psi(y) \tag{4.123}$$

这给出 $f(\psi(y)) = y$，因此 $f'(\psi)\psi'(y) = 1$。此外，因为 $a = f^{-1}(y)$，我们有 $a = \psi$，因此 $f'(a)\psi'(y) = 1$。在这种情况下，误差函数的梯度简化为

$$\nabla \ln E(\mathbf{w}) = \frac{1}{s}\sum_{n=1}^{N}\{y_n - t_n\}\phi_n \tag{4.124}$$

对于高斯分布 $s = \beta^{-1}$，而对于逻辑模型 $s = 1$。

4.4 拉普拉斯近似

在第 4.5 节中，我们将讨论逻辑回归的贝叶斯处理。正如我们将看到的，这比第 3.3 节和第 3.5 节中讨论的线性回归模型的贝叶斯处理更复杂。特别是，我们无法对参数向量 **w** 进行精确积分，因为后验分布不再是高斯分布。因此，有必要引入某种形式的近似。在本书的后面，我们将考虑一系列基于解析近似和数值采样的技术。

这里我们介绍一个简单但广泛使用的框架，称为拉普拉斯近似（Laplace approximation），其目的是为定义在一组连续变量上的概率密度找到一个高斯近似。首先考虑单个连续变量 z 的情况，假设分布 $p(z)$ 由下式定义

$$p(z) = \frac{1}{Z}f(z) \tag{4.125}$$

其中 $Z = \int f(z)\mathrm{d}z$ 是归一化系数。我们将假设 Z 的值是未知的。在拉普拉斯方法中，目标是找到一个高斯近似 $q(z)$，它以分布 $p(z)$ 的众数为中心。第一步是找到 $p(z)$ 的一个众数，换句话说，找到一个点 z_0，使得 $p'(z_0) = 0$，或者等价地

$$\left.\frac{df(z)}{dz}\right|_{z=z_0} = 0. \tag{4.126}$$

高斯分布具有这样的性质：它的对数是变量的二次函数。因此，我们考虑以众数 z_0 为中心的 $\ln f(z)$ 的泰勒展开，因此

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2}A(z - z_0)^2 \tag{4.127}$$

其中

$$A = -\left.\frac{d^2}{dz^2}\ln f(z)\right|_{z=z_0} \tag{4.128}$$

请注意，泰勒展开中的一阶项没有出现，因为 z_0 是分布的局部最大值。取指数，我们得到

$$f(z) \simeq f(z_0)\exp\left\{-\frac{A}{2}(z - z_0)^2\right\} \tag{4.129}$$

然后我们可以利用高斯分布归一化的标准结果来获得一个归一化分布 $q(z)$，因此

$$q(z) = \left(\frac{A}{2\pi}\right)^{1/2}\exp\left\{-\frac{A}{2}(z - z_0)^2\right\}. \tag{4.130}$$

拉普拉斯近似如图 4.14 所示。请注意，高斯近似只有在它的精度 $A > 0$ 时才定义良好，换句话说，驻点 z_0 必须是一个局部最大值，这样 $f(z)$ 在点 z_0 处的二阶导数才是负的。

我们可以将拉普拉斯方法扩展到近似定义在 M 维空间 $\mathbf{z}$ 上的分布 $p(\mathbf{z}) = f(\mathbf{z})/Z$。在驻点 $\mathbf{z}_0$ 处，梯度 $\nabla f(\mathbf{z})$ 将消失。围绕该驻点展开，我们有

$$\ln f(\mathbf{z}) \simeq \ln f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0) \tag{4.131}$$

其中 $M \times M$ Hessian 矩阵 $\mathbf{A}$ 定义为

$$\mathbf{A} = -\nabla\nabla \ln f(\mathbf{z})|_{\mathbf{z}=\mathbf{z}_0} \tag{4.132}$$

而 ∇ 是梯度算子。对两边取指数，我们得到

$$f(\mathbf{z}) \simeq f(\mathbf{z}_0)\exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\}. \tag{4.133}$$

分布 $q(\mathbf{z})$ 与 $f(\mathbf{z})$ 成比例，并且可以使用归一化多元高斯分布的标准结果 (2.43) 通过观察找到适当的归一化系数，给出

$$q(\mathbf{z}) = \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}}\exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\} = \mathcal{N}\left(\mathbf{z} \mid \mathbf{z}_0, \mathbf{A}^{-1}\right) \tag{4.134}$$

其中 $|\mathbf{A}|$ 表示 $\mathbf{A}$ 的行列式。当其精度矩阵 $\mathbf{A}$ 是正定时，这个高斯分布将定义良好，这意味着驻点 $\mathbf{z}_0$ 必须是局部最大值，而不是最小值或鞍点。

为了应用拉普拉斯近似，我们首先需要找到众数 $\mathbf{z}_0$，然后在该众数处评估 Hessian 矩阵。在实践中，众数通常通过运行某种形式的数值优化算法找到 (Bishop and Nabney, 2008)。实践中遇到的许多分布将是多峰的，因此根据考虑的众数，会有不同的拉普拉斯近似。请注意，为了应用拉普拉斯方法，真实分布的归一化常数 Z 不需要是已知的。由于中心极限定理，随着观测数据点数量的增加，模型的后验分布预计将越来越好地被高斯分布近似，因此我们期望拉普拉斯近似在数据点数量相对较大的情况下最有用。

拉普拉斯近似的一个主要弱点是，由于它基于高斯分布，因此它只直接适用于实变量。在其他情况下，可以将拉普拉斯近似应用于变量的变换。例如，如果 $0 \leqslant \tau < \infty$，那么我们可以考虑 $\ln \tau$ 的拉普拉斯近似。然而，拉普拉斯框架最严重的限制是它纯粹基于真实分布在变量的特定值处的方面，因此可能无法捕捉重要的全局属性。在第 10 章中，我们将考虑采用更全局视角的替代方法。

4.4.1 模型对比和贝叶斯信息准则（BIC）

除了近似分布 $p(\mathbf{z})$ 之外，我们还可以获得归一化常数 Z 的近似值。使用近似 (4.133)，我们有

$$\begin{aligned} Z &= \int f(\mathbf{z})\mathrm{d}\mathbf{z} \\ &\simeq f(\mathbf{z}_0)\int \exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\}\mathrm{d}\mathbf{z} \\ &= f(\mathbf{z}_0)\frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}} \end{aligned} \tag{4.135}$$

这里我们注意到被积函数是高斯分布，并使用了归一化高斯分布的标准结果 (2.43)。我们可以使用结果 (4.135) 来获得模型证据的近似值，正如第 3.4 节所讨论的，它在贝叶斯模型比较中起着核心作用。

考虑一个数据集 $\mathcal{D}$ 和一组具有参数 $\{\boldsymbol{\theta}_i\}$ 的模型 $\{\mathcal{M}_i\}$。对于每个模型，我们定义一个似然函数 $p(\mathcal{D} \mid \boldsymbol{\theta}_i, \mathcal{M}_i)$。如果我们在参数上引入一个先验 $p(\boldsymbol{\theta}_i \mid \mathcal{M}_i)$，那么我们感兴趣的是计算各种模型的模型证据 $p(\mathcal{D} \mid \mathcal{M}_i)$。从现在开始，我们省略对 $\mathcal{M}_i$ 的条件依赖以保持符号简洁。根据贝叶斯定理，模型证据由下式给出

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \tag{4.136}$$

将 $f(\boldsymbol{\theta}) = p(\mathcal{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$ 和 $Z = p(\mathcal{D})$ 对应起来，并应用结果 (4.135)，我们得到

$$\ln p(\mathcal{D}) \simeq \ln p\left(\mathcal{D} \mid \boldsymbol{\theta}_{\mathrm{MAP}}\right) + \underbrace{\ln p\left(\boldsymbol{\theta}_{\mathrm{MAP}}\right) + \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln |\mathbf{A}|}_{\text{奥卡姆因子}} \tag{4.137}$$

其中 $\boldsymbol{\theta}_{\mathrm{MAP}}$ 是后验分布众数处的 $\boldsymbol{\theta}$ 值，而 $\mathbf{A}$ 是负对数后验的二阶导数的 Hessian 矩阵

$$\mathbf{A} = -\nabla\nabla \ln p\left(\mathcal{D} \mid \boldsymbol{\theta}_{\mathrm{MAP}}\right) p\left(\boldsymbol{\theta}_{\mathrm{MAP}}\right) = -\nabla\nabla \ln p\left(\boldsymbol{\theta}_{\mathrm{MAP}} \mid \mathcal{D}\right) \tag{4.138}$$

(4.137) 右侧的第一项表示使用优化后的参数评估的对数似然，而剩余的三项构成了“奥卡姆因子”（Occam factor），它惩罚模型复杂度。

如果我们假设参数上的高斯先验分布是宽泛的，并且 Hessian 矩阵具有满秩，那么我们可以使用以下公式非常粗略地近似 (4.137)

$$\ln p(\mathcal{D}) \simeq \ln p\left(\mathcal{D} \mid \boldsymbol{\theta}_{\mathrm{MAP}}\right) - \frac{1}{2} M \ln N \tag{4.139}$$

其中 N 是数据点的数量，M 是 $\boldsymbol{\theta}$ 中的参数数量，我们省略了加性常数。这被称为贝叶斯信息准则（Bayesian Information Criterion，BIC）或 Schwarz 准则 (Schwarz, 1978)。请注意，与 (1.73) 给出的 AIC 相比，这更严厉地惩罚了模型复杂度。

复杂性度量，例如 AIC 和 BIC，具有易于评估的优点，但也可能给出误导性的结果。特别是，Hessian 矩阵具有满秩的假设通常是无效的，因为许多参数不是“充分确定的”。我们可以使用结果 (4.137) 从拉普拉斯近似开始获得模型证据的更准确估计，正如我们在第 5.7 节中在神经网络的背景下所说明的那样。

4.5 贝叶斯逻辑回归

我们现在转向逻辑回归的贝叶斯处理。逻辑回归的精确贝叶斯推断是难解的（intractable）。特别是，后验分布的评估需要对先验分布和一个似然函数的乘积进行归一化，而似然函数本身又包含逻辑 S 形函数的乘积，每个数据点一个。预测分布的评估同样难解。在这里，我们考虑将拉普拉斯近似应用于贝叶斯逻辑回归问题 (Spiegelhalter and Lauritzen, 1990; MacKay, 1992b)。

4.5.1 拉普拉斯近似

回想一下第 4.4 节，拉普拉斯近似是通过找到后验分布的众数，然后拟合一个以该众数为中心的高斯分布来获得的。这需要评估对数后验的二阶导数，这等价于找到 Hessian 矩阵。

因为我们寻求后验分布的高斯表示，所以自然地从高斯先验开始，我们将其写成一般形式

$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w} \mid \mathbf{m}_0, \mathbf{S}_0\right) \tag{4.140}$$

其中 $\mathbf{m}_0$ 和 $\mathbf{S}_0$ 是固定超参数。关于 $\mathbf{w}$ 的后验分布由下式给出

$$p(\mathbf{w} \mid \mathbf{t}) \propto p(\mathbf{w}) p(\mathbf{t} \mid \mathbf{w}) \tag{4.141}$$

其中 $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$。对两边取对数，并使用 (4.140) 代入先验分布，使用 (4.89) 代入似然函数，我们得到

$$\begin{aligned} \ln p(\mathbf{w} \mid \mathbf{t}) = & -\frac{1}{2}\left(\mathbf{w}-\mathbf{m}_0\right)^{\mathrm{T}} \mathbf{S}_0^{-1}\left(\mathbf{w}-\mathbf{m}_0\right) \\ & +\sum_{n=1}^{N}\left\{t_n \ln y_n + \left(1-t_n\right) \ln\left(1-y_n\right)\right\} + \text{ const} \end{aligned} \tag{4.142}$$

其中 $y_n = \sigma\left(\mathbf{w}^{\mathrm{T}} \phi_n\right)$。为了获得后验分布的高斯近似，我们首先最大化后验分布以得到 MAP（最大后验）解 $\mathbf{w}_{\text{MAP}}$，它定义了高斯分布的均值。协方差然后由负对数似然的二阶导数矩阵的逆给出，其形式为

$$\mathbf{S}_N = -\nabla\nabla \ln p(\mathbf{w} \mid \mathbf{t}) = \mathbf{S}_0^{-1} + \sum_{n=1}^{N} y_n\left(1-y_n\right) \phi_n \phi_n^{\mathrm{T}} \tag{4.143}$$

因此，后验分布的高斯近似采用以下形式

$$q(\mathbf{w}) = \mathcal{N}\left(\mathbf{w} \mid \mathbf{w}_{\text{MAP}}, \mathbf{S}_N\right) \tag{4.144}$$

在获得了后验分布的高斯近似之后，剩下的任务是相对于这个分布进行边缘化以进行预测。

4.5.2 预测分布

给定一个新的特征向量 $\phi(\mathbf{x})$，类别 $\mathcal{C}_1$ 的预测分布是通过对后验分布 $p(\mathbf{w} \mid \mathbf{t})$ 进行边缘化获得的，该分布本身被高斯分布 $q(\mathbf{w})$ 所近似，因此

$$p\left(\mathcal{C}_1 \mid \phi, \mathbf{t}\right) = \int p\left(\mathcal{C}_1 \mid \phi, \mathbf{w}\right) p(\mathbf{w} \mid \mathbf{t}) \mathrm{d}\mathbf{w} \simeq \int \sigma\left(\mathbf{w}^{\mathrm{T}} \phi\right) q(\mathbf{w}) \mathrm{d}\mathbf{w} \tag{4.145}$$

类别 $\mathcal{C}_2$ 的相应概率由 $p\left(\mathcal{C}_2 \mid \phi, \mathbf{t}\right) = 1 - p\left(\mathcal{C}_1 \mid \phi, \mathbf{t}\right)$ 给出。为了评估预测分布，我们首先注意到函数 $\sigma\left(\mathbf{w}^{\mathrm{T}} \phi\right)$ 仅通过其在 ϕ 上的投影依赖于 $\mathbf{w}$。记 $a = \mathbf{w}^{\mathrm{T}} \phi$，我们有

$$\sigma\left(\mathbf{w}^{\mathrm{T}} \phi\right) = \int \delta\left(a - \mathbf{w}^{\mathrm{T}} \phi\right) \sigma(a) \mathrm{d}a \tag{4.146}$$

其中 $\delta(\cdot)$ 是狄拉克 δ 函数。由此我们得到

$$\int \sigma\left(\mathbf{w}^{\mathrm{T}}\phi\right) q(\mathbf{w})\mathrm{d}\mathbf{w} = \int \sigma(a)p(a)\mathrm{d}a \tag{4.147}$$

其中

$$p(a) = \int \delta\left(a - \mathbf{w}^{\mathrm{T}}\phi\right) q(\mathbf{w})\mathrm{d}\mathbf{w} \tag{4.148}$$

我们可以通过注意到 δ 函数对 $\mathbf{w}$ 施加了线性约束，并通过积分排除所有正交于 ϕ 的方向，从而从联合分布 $q(\mathbf{w})$ 中形成一个边缘分布来评估 $p(a)$。由于 $q(\mathbf{w})$ 是高斯分布，我们从第 2.3.2 节中知道边缘分布也将是高斯分布。我们可以通过取矩，并交换 a 和 $\mathbf{w}$ 上的积分顺序来评估该分布的均值和协方差，因此

$$\mu_a = \mathbb{E}[a] = \int p(a)a\ \mathrm{d}a = \int q(\mathbf{w})\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\ \mathrm{d}\mathbf{w} = \mathbf{w}_{\mathrm{MAP}}^{\mathrm{T}}\boldsymbol{\phi} \tag{4.149}$$

其中我们使用了变分后验分布 $q(\mathbf{w})$ 的结果 (4.144)。类似地

$$\begin{aligned}\sigma_a^2 = \mathrm{var}[a] &= \int p(a)\left\{a^2 - \mathbb{E}[a]^2\right\}\mathrm{d}a \\ &= \int q(\mathbf{w})\left\{\left(\mathbf{w}^{\mathrm{T}}\phi\right)^2 - \left(\mathbf{m}_N^{\mathrm{T}}\phi\right)^2\right\}\mathrm{d}\mathbf{w} = \phi^{\mathrm{T}}\mathbf{S}_N\phi\end{aligned} \tag{4.150}$$

请注意，a 的分布与线性回归模型的预测分布 (3.58) 采用相同的形式，噪声方差设置为零。因此，我们对预测分布的变分近似变为

$$p\left(\mathcal{C}_1 \mid \mathbf{t}\right) = \int \sigma(a)p(a)\mathrm{d}a = \int \sigma(a)\mathcal{N}\left(a \mid \mu_a, \sigma_a^2\right)\mathrm{d}a. \tag{4.151}$$

该结果也可以通过利用第 2.3.2 节中给出的高斯分布边缘的结果直接推导出来。

对 a 的积分表示高斯分布与逻辑 S 形函数的卷积，无法解析评估。然而，我们可以通过利用由 (4.59) 定义的逻辑 S 形函数 $\sigma(a)$ 和由 (4.114) 定义的概率单位函数 $\Phi(a)$ 之间的高度相似性来获得一个良好的近似 (Spiegelhalter and Lauritzen, 1990; MacKay, 1992b; Barber and Bishop, 1998a)。为了获得对逻辑函数的最佳近似，我们需要重新缩放水平轴，以便我们用 $\Phi(\lambda a)$ 来近似 $\sigma(a)$。我们可以通过要求这两个函数在原点处具有相同的斜率来找到一个合适的 λ 值，这给出 $\lambda^2 = \pi/8$。对于这个 λ 的选择，逻辑 S 形函数和概率单位函数的相似性如图 4.9 所示。

使用概率单位函数的优点在于，它与高斯分布的卷积可以解析地用另一个概率单位函数来表示。具体来说，我们可以证明

$$\int \Phi(\lambda a)\mathcal{N}\left(a \mid \mu, \sigma^2\right)\mathrm{d}a = \Phi\left(\frac{\mu}{\left(\lambda^{-2} + \sigma^2\right)^{1/2}}\right). \tag{4.152}$$

我们现在将近似 $\sigma(a) \simeq \Phi(\lambda a)$ 应用于该等式两边出现的概率单位函数，从而得到逻辑 S 形函数与高斯分布的卷积的近似

$$\int \sigma(a)\mathcal{N}\left(a \mid \mu, \sigma^2\right)\mathrm{d}a \simeq \sigma\left(\kappa\left(\sigma^2\right)\mu\right) \tag{4.153}$$

其中我们定义了

$$\kappa\left(\sigma^2\right)=\left(1+\pi\sigma^2/8\right)^{-1/2} \tag{4.154}$$

将此结果应用于 (4.151)，我们得到以下形式的近似预测分布

$$p\left(\mathcal{C}_1 \mid \phi, \mathbf{t}\right)=\sigma\left(\kappa\left(\sigma_a^2\right)\mu_a\right) \tag{4.155}$$

其中 μ_a 和 σ_a^2 分别由 (4.149) 和 (4.150) 定义，而 $\kappa\left(\sigma_a^2\right)$ 由 (4.154) 定义。

请注意，对应于 $p\left(\mathcal{C}_1 \mid \phi, \mathbf{t}\right)=0.5$ 的决策边界由 $\mu_a=0$ 给出，这与使用 $\mathbf{w}$ 的 MAP 值获得的决策边界相同。因此，如果决策标准基于最小化误分类率，并且先验概率相等，那么对 $\mathbf{w}$ 的边缘化没有影响。然而，对于更复杂的决策标准，它将发挥重要作用。在变分推断的背景下，逻辑 sigmoid 模型在后验分布的高斯近似下的边缘化将在图 10.13 中进行说明。