

模式识别与机器学习

樊超

November 19, 2025

目 录

1	引言	1
1.1	例子：多项式曲线拟合	3
1.2	概率论	9
1.3	模型选择	25
1.4	维度灾难	27
1.5	决策理论	30
1.6	信息论	39
2	概率分布	47
2.1	二元变量	47
2.2	多项式变量	52
2.3	高斯分布	55
2.4	指数家族	83
2.5	非参数方法	89

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, ..., 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。

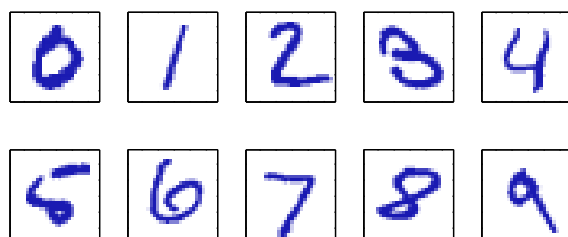


图 1.1 来自美国邮政编码（US zip codes）的手写数字示例

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 $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ ，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 $\mathbf{y}(\mathbf{x})$ ，它以一个数字图像 \mathbf{x} 作为输入，并生成一个输出向量 \mathbf{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 \mathbf{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??。

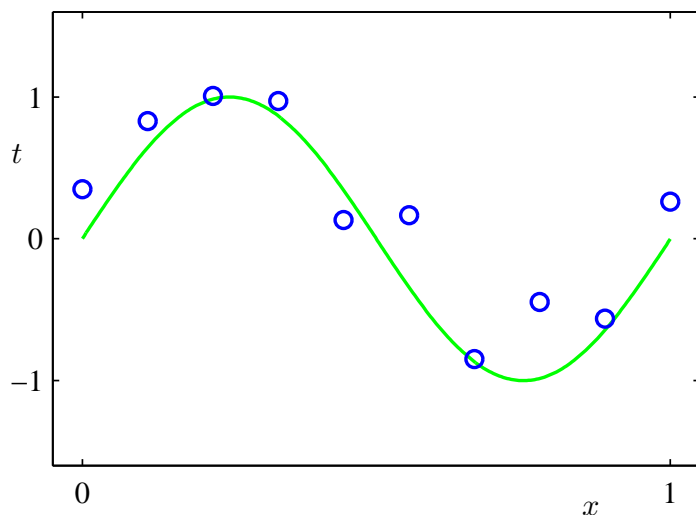


图 1.2 一个包含 $N = 10$ 个点的训练数据集的示意图，以蓝色圆点表示；每个点由输入变量 x 的一次观测及其对应的目标变量 t 构成。绿色曲线表示用于生成数据的函数 $\sin(2\pi x)$ 。我们的目标是在不知道这条绿色曲线的情况下，对某个新的 x 值预测对应的 t 值。

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 $\mathbf{x} \equiv (x_1, \dots, x_N)^T$ ，并且有与之对应的目标值观测，记作 $\mathbf{t} \equiv (t_1, \dots, t_N)^T$ 。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 ($n = 1, \dots, N$) 生成的；目标数据集 \mathbf{t} 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上进行泛化，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：

$$y(x, \mathbf{w}) = w_0 + w_1x + w_2x^2 + \cdots + w_Mx^M = \sum_{j=0}^M w_jx^j \quad (1.1)$$

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, \dots, 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。

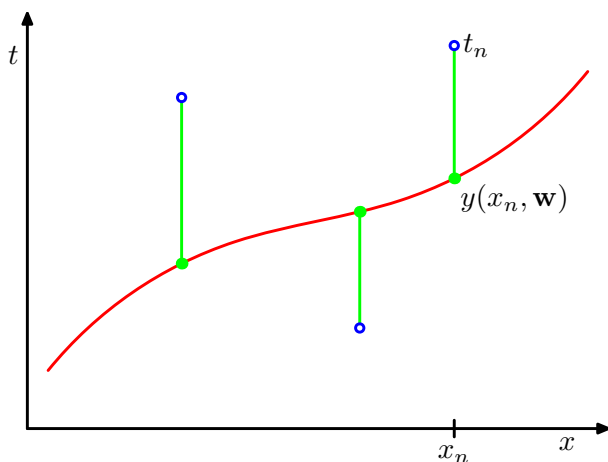


图 1.3 误差函数 (1.2) 对应于每个数据点与函数 $y(x, \mathbf{w})$ 之间偏差平方和的一半，其中这些偏差由垂直的绿色线段表示。

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

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

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。

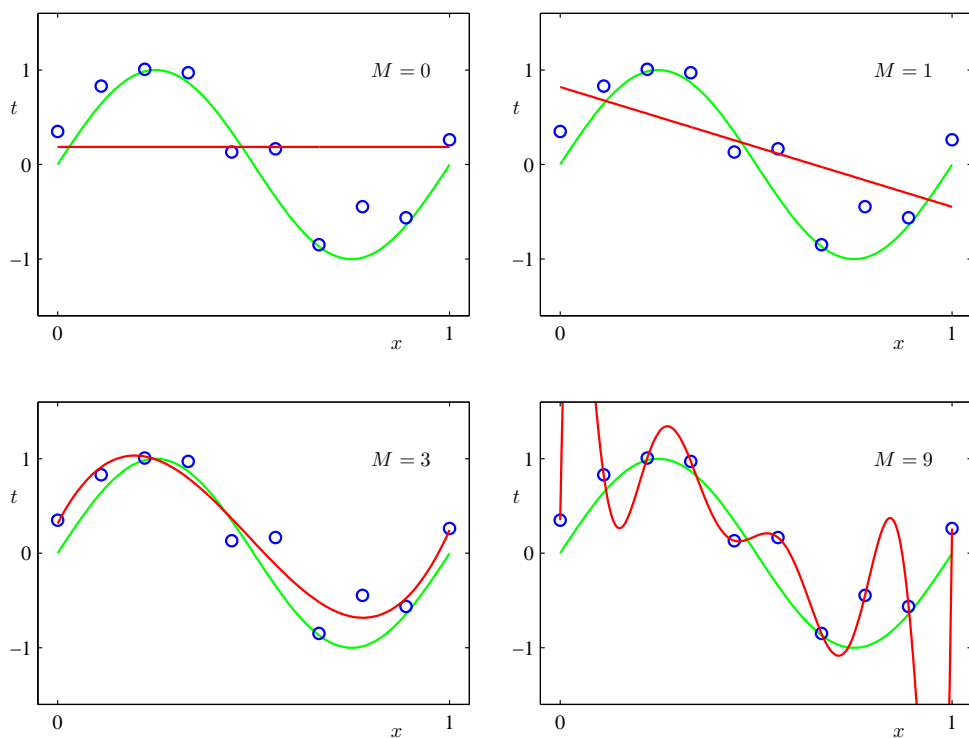


图 1.4 不同阶数 M 的多项式拟合结果示意图，以红色曲线表示，并拟合到图 1.2 所示的数据集上。

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 $E(\mathbf{w}^*)$ （由公式 1.2 给出），同时也可以计算测试数据集上的 $E(\mathbf{w}^*)$ 。有时使用均方根误差

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

$$E_{\text{RMS}} = \sqrt{\frac{2E(\mathbf{w}^*)}{N}} \quad (1.3)$$

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。

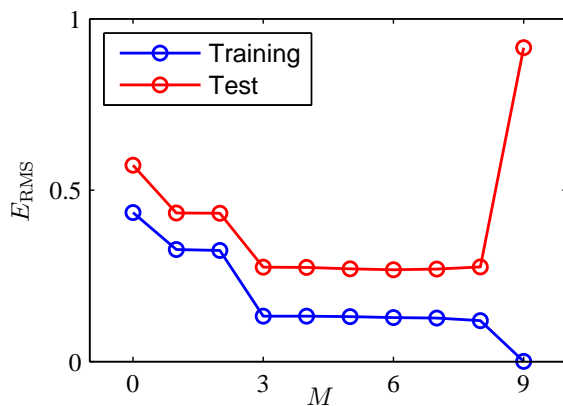


图 1.5 在不同 M 取值下, 训练集与独立测试集上的均方根误差 (由公式 1.3 定义) 曲线图。

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, \dots, 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}^* 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}^* 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^*	0.19	0.82	0.31	0.35
w_1^*		-1.27	7.99	232.37
w_2^*			-25.43	-5321.1
w_3^*			17.37	48568.31
w_4^*				-231639.30
w_5^*				640042.26
w_6^*				-1061800.52
w_7^*				-557682.99
w_8^*				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。

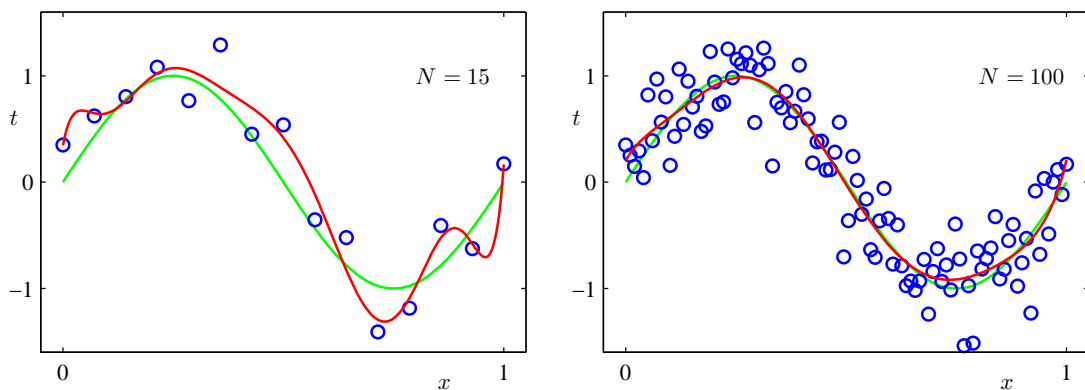


图 1.6 通过最小化平方和误差函数所得到的 $M = 9$ 多项式解的曲线图：左图为 $N = 15$ 个数据点的情况，右图为 $N = 100$ 个数据点的情况。可以看到，增大数据集的规模能够减轻过拟合问题。

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

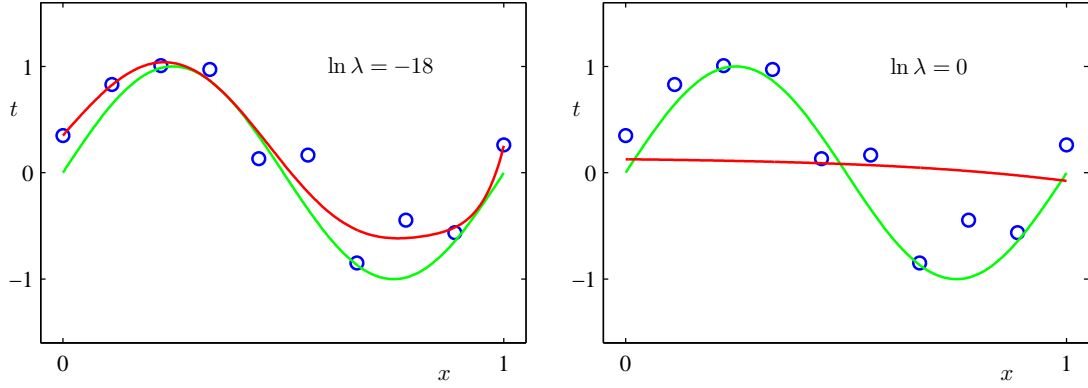


图 1.7 利用正则化误差函数 (1.4) 将 $M = 9$ 的多项式拟合到图 1.2 所示数据集的结果曲线图，其中正则化参数 λ 分别取 $\ln \lambda = -18$ 和 $\ln \lambda = 0$ 。无正则化的情况，即 $\lambda = 0$ （对应 $\ln \lambda = -\infty$ ），见图 1.4 的右下角。

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：

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

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）。

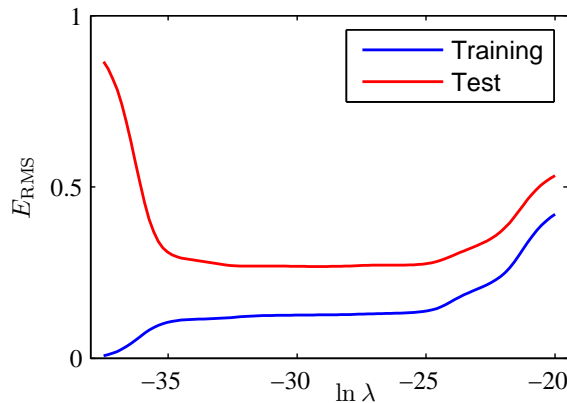


图 1.8 $M = 9$ 多项式情况下，均方根误差（公式 (1.3)）随 $\ln \lambda$ 变化的曲线图。

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^*	0.35	0.35	0.13
w_1^*	232.37	4.74	-0.05
w_2^*	-5321.83	-0.77	-0.06
w_3^*	48568.31	-31.97	-0.05
w_4^*	-231639.30	-3.89	-0.03
w_5^*	640042.26	55.28	-0.02
w_6^*	-1061800.52	41.32	-0.01
w_7^*	1042400.18	-45.92	-0.00
w_8^*	-557682.99	-91.53	0.00
w_9^*	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。

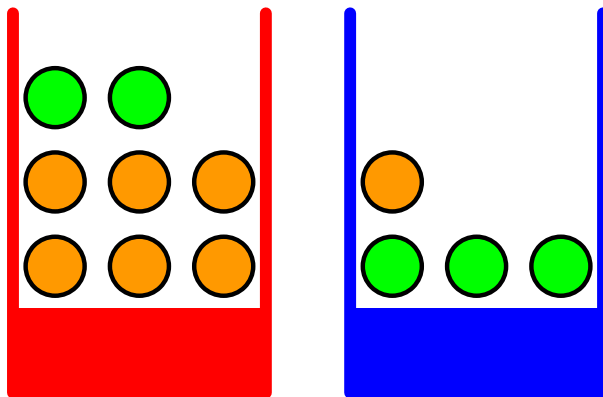


图 1.9 我们用一个简单的例子来引入概率 (probability) 的基本概念：两个有颜色的盒子，每个盒子里都装有水果（绿色表示苹果，橙色表示橙子）。

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 $i = 1, \dots, M$ ，而 Y 可以取值 y_j ，其中 $j = 1, \dots, L$ 。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

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

这里我们隐含地考虑极限 $N \rightarrow \infty$ 。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

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

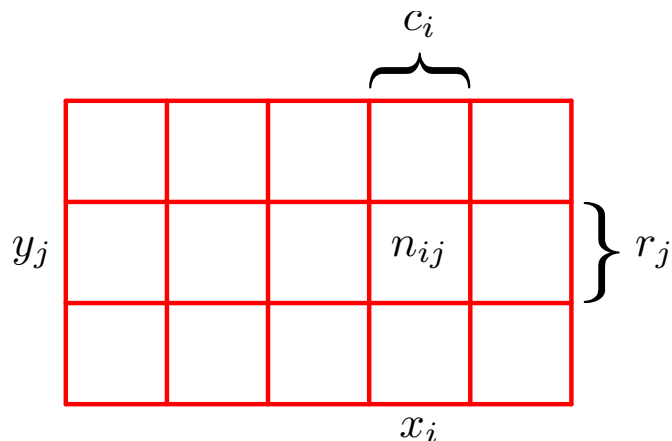


图 1.10 我们可以通过考虑两个随机变量来推导概率的加法法则和乘法法则。设随机变量 X 的取值为 x_i , 其中 $i = 1, \dots, M$; 随机变量 Y 的取值为 y_j , 其中 $j = 1, \dots, L$ 。在这个例子中, 我们取 $M = 5, L = 3$ 。如果考虑这两个变量的总实例数为 N , 那么当 $X = x_i$ 且 $Y = y_j$ 时的实例数记作 n_{ij} , 它表示数组对应单元格中的点数。第 i 列的点数 (对应 $X = x_i$) 记作 c_i , 第 j 行的点数 (对应 $Y = y_j$) 记作 r_j 。

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

$$p(X = x_i) = \sum_j p(X = x_i, Y = y_j) \quad (1.7)$$

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, 因为它通过与其他变量 (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 $p(Y = y_j | X = x_i)$, 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

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

由 (1.5)、(1.6) 和 (1.8) 可推得如下关系

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

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

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

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

这里 $p(X, Y)$ 是联合概率，读作“ X 和 Y 的概率”。类似地， $p(Y | X)$ 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：

$$p(Y | X) = \frac{p(X | Y) p(Y)}{p(X)} \quad (1.12)$$

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：

$$p(Y | X) = \frac{p(X | Y) p(Y)}{\sum_Y p(X | Y) p(Y)} \quad (1.13)$$

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 $N \rightarrow \infty$ 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 $p(X | Y = 1)$ 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

$$p(B = r) = 4/10 \quad (1.14)$$

$$p(B = b) = 6/10 \quad (1.15)$$

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 $p(F = a | B = b) = 3/4$ 。事实上，我们可

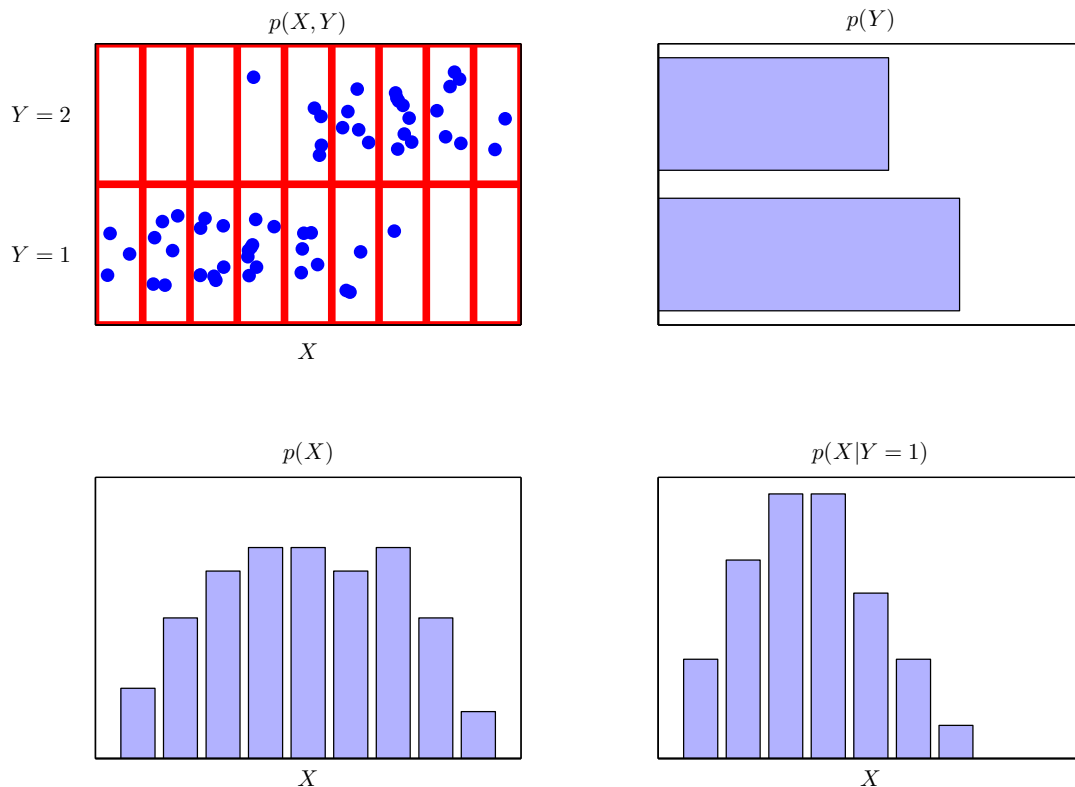


图 1.11 关于两个变量分布的示意图：变量 X 取 9 个可能值，变量 Y 取 2 个可能值。左上图展示了从这两个变量的联合概率分布中抽取的 60 个样本点。其余图则展示了边缘分布 $p(X)$ 和 $p(Y)$ 的直方图估计，以及条件分布 $p(X|Y=1)$ ，该条件分布对应于左上图的最下一行。

以写出在给定选中盒子的情况下，水果种类四个条件概率：

$$p(F = a \mid B = r) = 1/4 \quad (1.16)$$

$$p(F = o \mid B = r) = 3/4 \quad (1.17)$$

$$p(F = a \mid B = b) = 3/4 \quad (1.18)$$

$$p(F = o \mid B = b) = 1/4 \quad (1.19)$$

同样需要注意，这些概率是归一化的，因此满足

$$p(F = a \mid B = r) + p(F = o \mid B = r) = 1 \quad (1.20)$$

$$p(F = a \mid B = b) + p(F = o \mid B = b) = 1 \quad (1.21)$$

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：

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

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

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

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

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 $p(B | F)$, 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 $p(Y | X) = p(Y)$, 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 $p(F | B) = p(F)$, 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

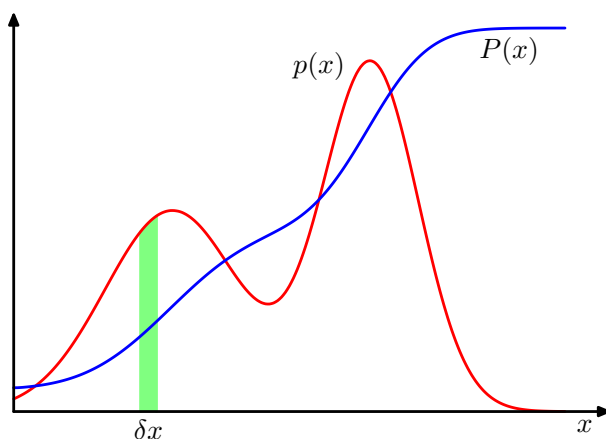


图 1.12 离散变量的概率概念可以扩展为连续变量 x 上的概率密度 $p(x)$, 此时 x 落在区间 $(x, x + \delta x)$ 内的概率为 $p(x)\delta x$, 当 $\delta x \rightarrow 0$ 。概率密度可以表示为累积分布函数 $P(x)$ 的导数。

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 $\delta x \rightarrow 0$ 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

$$p(a \in (a, b)) = \int_a^b p(x) dx \quad (1.24)$$

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：

$$p(x) \geq 0 \quad (1.25)$$

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

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

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

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：

$$P(z) = \int_{-\infty}^z p(x) dx \quad (1.28)$$

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, \dots, 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） $p(\mathbf{x}) = p(x_1, \dots, x_D)$ ，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

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

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

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

$$p(x) = \int p(x, y) dy \quad (1.31)$$

$$p(x, y) = p(y | x) p(x) \quad (1.32)$$

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 $\Delta \rightarrow 0$ 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

$$\mathbb{E}[f] = \sum_x p(x) f(x) \quad (1.33)$$

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:

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

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:

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

在第 11 章讨论采样方法时, 我们将广泛使用这一结果。当 $N \rightarrow \infty$ 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:

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

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

$$\mathbb{E}[f | y] = \int p(x | y) f(x) dx, \quad (1.37)$$

在连续变量情况下有类似的定义。函数 $f(x)$ 的方差定义为

$$\text{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] \quad (1.38)$$

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:

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

特别地, 变量 x 本身的方差可以写作

$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 \quad (1.40)$$

对于两个随机变量 x 和 y ，协方差定义为

$$\begin{aligned}\text{cov}[x, y] &= \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}\quad (1.41)$$

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 \mathbf{x} 和 \mathbf{y} ，协方差是一个矩阵：

$$\begin{aligned}\text{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x}, \mathbf{y}}[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^T - \mathbb{E}[\mathbf{y}^T]\}] \\ &= \mathbb{E}_{\mathbf{x}, \mathbf{y}}[\mathbf{x}\mathbf{y}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^T].\end{aligned}\quad (1.42)$$

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 $\text{cov}[\mathbf{x}] \equiv \text{cov}[\mathbf{x}, \mathbf{x}]$ 。

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程度。

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，而是在我们期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 \mathbf{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 \mathbf{w} 的假设。观测到的数据 $\mathcal{D} = \{t_1, \dots, t_N\}$ 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

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

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 \mathbf{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 \mathbf{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 \mathbf{w} 的概率分布，它对 \mathbf{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

$$\text{后验分布} \propto \text{似然函数} \times \text{先验分布} \quad (1.44)$$

这里所有的量都被看作是关于 \mathbf{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 \mathbf{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：

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

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， \mathbf{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 \mathbf{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 \mathbf{w} 被设定为使似然函数 $p(\mathcal{D}|\mathbf{w})$ 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 \mathbf{w} 。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 $\mathbf{X} = \{x_1, \dots, x_N\}$ 。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， \mathbf{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}_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 \mathbf{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。如果采用经典的极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

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

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

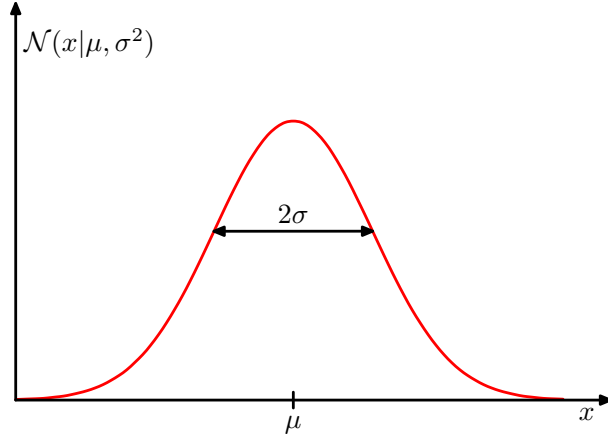


图 1.13 一维高斯分布的示意图，展示了均值 μ 和标准差 σ 。

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

此外，很容易证明高斯分布是归一化的，因此

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

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：

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

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：

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

由 (1.49) 和 (1.50) 可得， x 的方差为

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

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

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

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 $\mathbf{x} = (x_1, \dots, x_N)^T$ ，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 $(x_1, \dots, x_D)^T$ 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：

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

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。

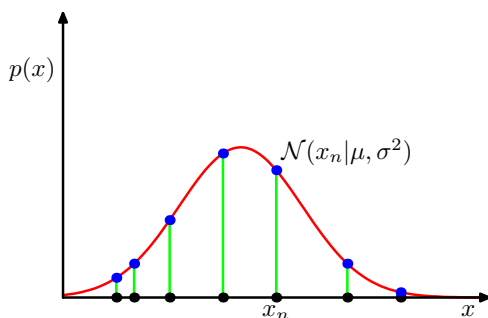


图 1.14 高斯分布似然函数的示意图由红色曲线表示。黑点表示数据集 x_n ，公式 1.53 给出的似然函数对应于这些蓝色数值的乘积。最大化似然的过程就是通过调整高斯分布的均值和方差，使得这个乘积最大化。

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：

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

对 (1.54) 式关于 μ 最大化，可得最大似然解：

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n \quad (1.55)$$

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

$$\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2 \quad (1.56)$$

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_{ML}^2 都是数据集取值 x_1, \dots, 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：

$$\mathbb{E}[\mu_{\text{ML}}] = \mu \quad (1.57)$$

$$\mathbb{E}[\sigma_{\text{ML}}^2] = \frac{N-1}{N} \sigma^2 \quad (1.58)$$

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。

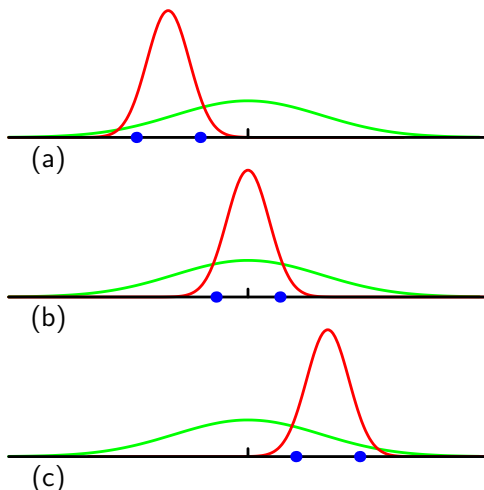


图 1.15 示意图展示了在使用最大似然方法确定高斯分布方差时偏差产生的原因。绿色曲线表示生成数据的真实高斯分布，三条红色曲线表示分别拟合三个数据集得到的高斯分布，每个数据集包含两个蓝色数据点，拟合采用最大似然结果 1.55 和 1.56。在三个数据集上取平均后，均值是正确的，但方差被系统性低估，因为它是相对于样本均值而不是相对于真实均值来计算的。

由 (1.58) 可得，下列方差参数的估计是无偏的：

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

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 $N \rightarrow \infty$ 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 $\mathbf{x} = (x_1, \dots, x_N)^T$ 及其对应的目标值 $\mathbf{t} = (t_1, \dots, t_N)^T$, 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:

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

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。

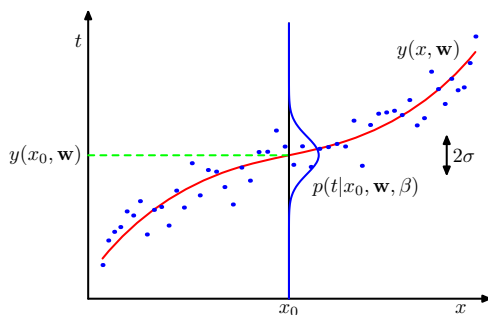


图 1.16 高斯条件分布的示意图, 表示在给定 x 的情况下 t 的分布 (公式 (1.60))。其中, 均值由多项式函数 $y(x, \mathbf{w})$ 给出, 精度由参数 β 决定, 并且与方差的关系为 $\beta^{-1} = \sigma^2$

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

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

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

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

首先考虑多项式系数的最大似然解, 记为 \mathbf{w}_{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：

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

同样地，我们可以先确定控制均值的参数向量 \mathbf{w}_{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：

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

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：

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

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：

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

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：

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

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 \mathbf{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 \mathbf{x} 和 \mathbf{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：

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

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：

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

其中，均值和方差分别为

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

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

这里矩阵 \mathbf{S} 给定为

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

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

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 \mathbf{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。

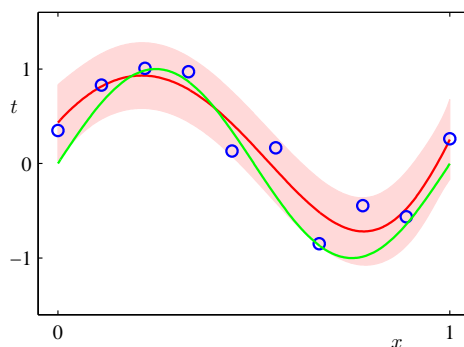


图 1.17 基于贝叶斯方法对 $M = 9$ 多项式进行曲线拟合所得到的预测分布 (predictive distribution), 其固定参数为 $\alpha = 5 \times 10^{-3}$ 和 $\beta = 11.1$ (对应已知的噪声方差)。红色曲线表示预测分布的均值, 红色区域表示均值上下一个标准差的范围。

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。

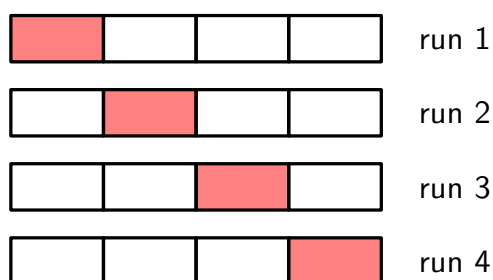


图 1.18 S 折交叉验证的方法, 这里以 $S = 4$ 的情况为例进行说明。该方法将可用数据划分为 S 个组 (在最简单的情况下, 这些组大小相等)。然后使用其中 $S-1$ 个组来训练模型, 并在剩下的一个组上进行评估。这个过程会对所有 S 种可能的留出组重复进行, 这里用红色方块表示。最终将 S 次运行得到的性能指标取平均。

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：

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

这里， $p(\mathcal{D}|\mathbf{w}_{\text{ML}})$ 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。

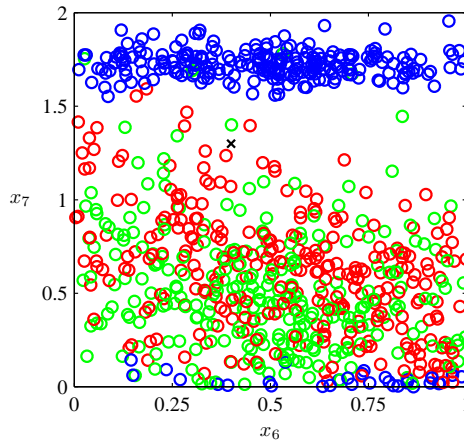


图 1.19 油流数据中输入变量 x_6 和 x_7 的散点图，其中红色表示“homogenous”类别，绿色表示“annular”类别，蓝色表示“laminar”类别。我们的目标是对标记为“x”的新测试点进行分类。

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

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。

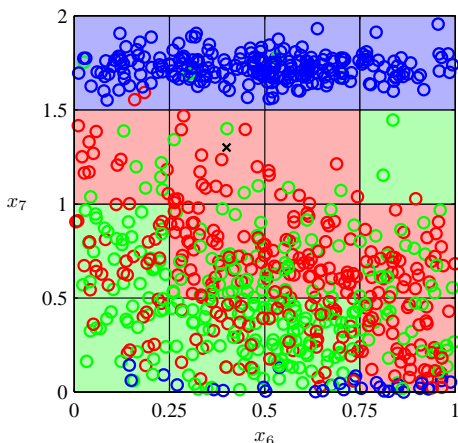


图 1.20 分类问题的一种简单解法示意图：将输入空间划分为若干单元格，并将新的测试点分配到与其所在单元格中占多数的类别。正如我们很快将看到的，这种过于简单的方法存在严重缺陷。

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）。

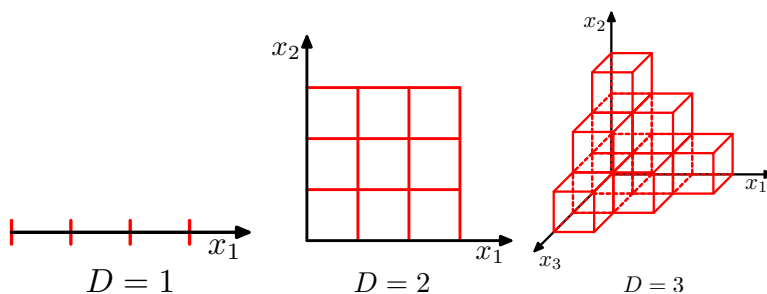


图 1.21 维度灾难的示意图，展示了规则网格的区域数量如何随着空间维度 D 的增加而指数式增长。为了清晰起见，在 $D = 3$ 的情况下仅展示了部分立方区域。

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：

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

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：

$$V_D(r) = K_D r^D \quad (1.75)$$

其中常数 K_D 只依赖于维度 D 。因此，所需的体积分数为

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

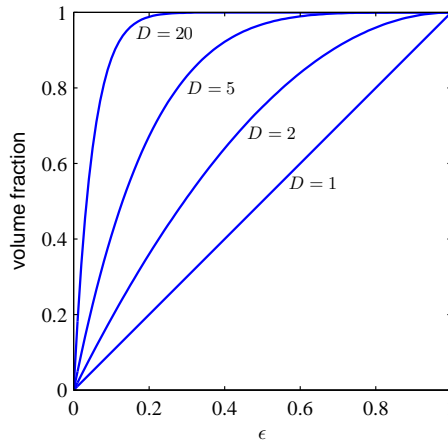


图 1.22 球体体积在半径范围 $r = 1 - \epsilon$ 到 $r = 1$ 内所占比例随空间维度 D 变化的曲线图。

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

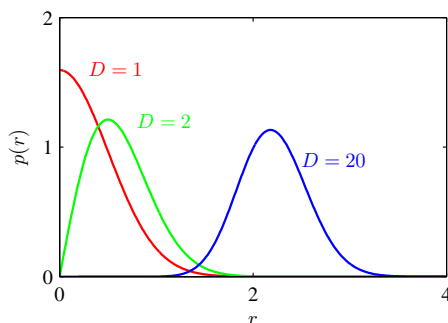


图 1.23 高斯分布关于半径 r 的概率密度随维度 D 变化的曲线图。在高维空间中，高斯分布的大部分概率质量集中在某一特定半径处的薄壳内。

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：

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

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：

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, C_2) + p(\mathbf{x} \in \mathcal{R}_2, C_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, C_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, C_1) d\mathbf{x} \end{aligned} \quad (1.78)$$

我们可以自由选择决策规则，将每个点 \mathbf{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 \mathbf{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 \mathbf{x} ，有 $p(\mathbf{x}, C_1) > p(\mathbf{x}, C_2)$ ，那么就应当把该 \mathbf{x} 判为类别 C_1 。根据概率的乘法法则 (product rule)，我们有 $p(\mathbf{x}, C_k) = p(C_k|\mathbf{x}) p(\mathbf{x})$ 。由于 $p(\mathbf{x})$ 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 \mathbf{x} 判为其后验概率 $p(C_k|\mathbf{x})$ 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。

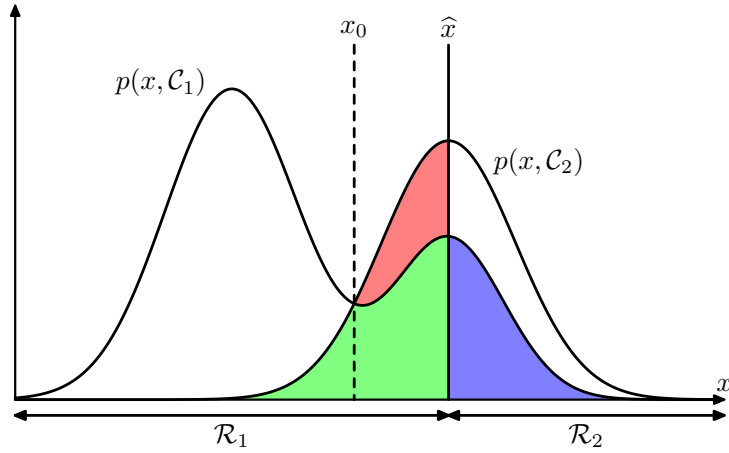


图 1.24 关于两个类别的联合概率 $p(x, C_k)$ 随 x 变化的示意图，同时给出了决策边界 $x = \hat{x}$ 。当 $x \geq \hat{x}$ 时，被分类为类别 C_2 ，因此属于决策区域 \mathcal{R}_2 ；而当 $x < \hat{x}$ 时，被分类为类别 C_1 ，属于决策区域 \mathcal{R}_1 。错误来源于蓝色、绿色和红色区域：在 $x < \hat{x}$ 时，错误来自类别 C_2 的点被误分类为 C_1 （由红色和绿色区域表示）；而在 $x \geq \hat{x}$ 时，错误来自类别 C_1 的点被误分类为 C_2 （由蓝色区域表示）。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 $p(x, C_1)$ 与 $p(x, C_2)$ 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 $p(C_k|x)$ 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：

$$\begin{aligned} p(\text{correct}) &= \sum_{k=1}^K p(\mathbf{x} \in \mathcal{R}_k, C_k) \\ &= \sum_{k=1}^K \int_{\mathcal{R}_k} p(\mathbf{x}, C_k) d\mathbf{x} \end{aligned} \quad (1.79)$$

当决策区域 \mathcal{R}_k 的选择使得每个 \mathbf{x} 都被分配给满足 $p(\mathbf{x}, C_k)$ 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： $p(\mathbf{x}, C_k) = p(C_k|\mathbf{x}) p(\mathbf{x})$ ，并注意到因子 $p(\mathbf{x})$ 对所有项都是相同的，于是可以得到：每个 \mathbf{x} 应该被判为其后验概率 $p(C_k|\mathbf{x})$ 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 \mathbf{x} ，其真实类别为 C_k ，而我们将其判为类别 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。

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

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 \mathbf{x} ，我们对真实类别的不确定性由联合分布 $p(\mathbf{x}, C_k)$ 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：

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, C_k) d\mathbf{x} \quad (1.80)$$

对于每个 \mathbf{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 \mathbf{x} 应该最小化 $\sum_k L_{kj} p(\mathbf{x}, C_k)$ 。如前所述，我们可以利用乘法规则 $p(\mathbf{x}, C_k) = p(C_k | \mathbf{x}) p(\mathbf{x})$ 消去公共因子 $p(\mathbf{x})$ 。因此，能够最小化期望损失的判别规则是：将每个新的 \mathbf{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 $p(C_k | \mathbf{x})$ ，这一决策就变得非常简单。

$$\sum_k L_{kj} p(C_k | \mathbf{x}) \quad (1.81)$$

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 $p(C_k|x)$ 显著小于 1，或者等价地说，此时各联合分布 $p(x, C_k)$ 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 $p(C_k|x) \leq \theta$ 时，就拒绝对输入 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。

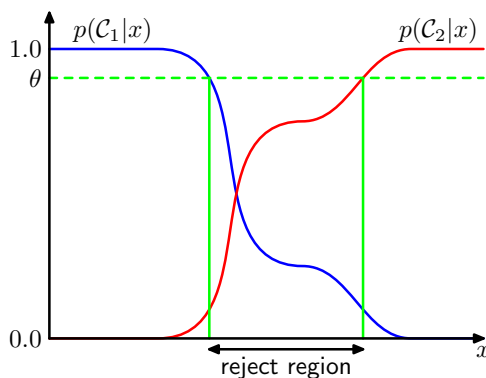


图 1.26 拒绝选项的示意图：对于输入 x ，如果两类后验概率中较大的一个小于或等于某个阈值 θ ，则该输入将被拒绝分类。

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 $p(C_k|x)$ 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 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 C_k 的类条件密度 $p(x|C_k)$ 。同时单独推断先验类别概率 $p(C_k)$ 。然后使用贝叶斯定理，其形式为

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)} \quad (1.82)$$

以此来求解后验类别概率 $p(C_k|x)$ 。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

$$p(\mathbf{x}) = \sum_j p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j) \quad (1.83)$$

等价地，我们也可以直接对联合分布 $p(\mathbf{x}, \mathcal{C}_k)$ 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 \mathbf{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 $p(\mathcal{C}_k|\mathbf{x})$ ，然后再利用决策理论将每个新的 \mathbf{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 $f(\mathbf{x})$ ，称为判别函数，它将每个输入 \mathbf{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 \mathbf{x} 和 \mathcal{C}_k 的联合分布。对于许多应用来说， \mathbf{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 $p(\mathbf{x}|\mathcal{C}_k)$ 。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 $p(\mathbf{x})$ 。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 $p(\mathcal{C}_k|\mathbf{x})$ 的情况下去寻找联合分布 $p(\mathbf{x}, \mathcal{C}_k)$ ，就可能对计算资源的浪费，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 $f(\mathbf{x})$ ，它将每个 \mathbf{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 $p(\mathcal{C}_k|\mathbf{x})$ 。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。

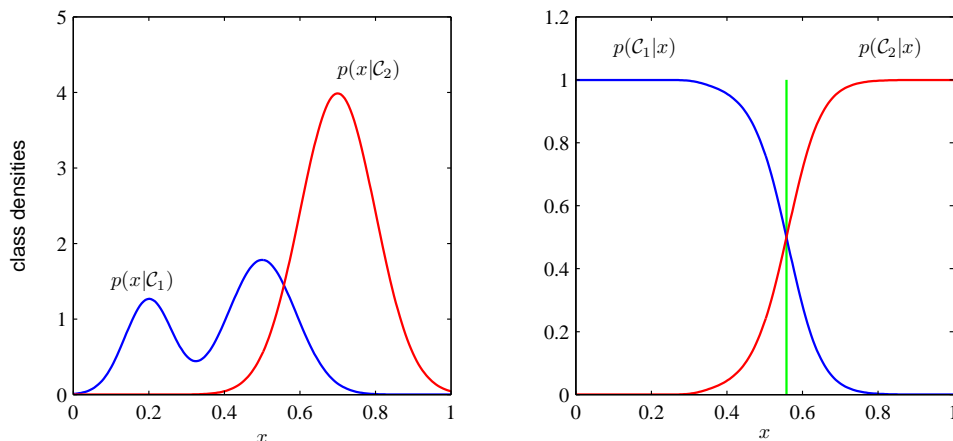


图 1.27 两个类别在单一输入变量 x 下的类条件密度示例（左图），以及对应的后验概率（右图）。需要注意的是，左图中蓝色表示的类条件密度 $p(x|C_1)$ 的左侧峰值对后验概率没有影响。右图中的竖直绿色线表示在 x 上给出最小误分类率的决策边界。

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 \mathbf{x}_I ）与血液数据的输入分布（记作 \mathbf{x}_B ）相互独立，于是有

$$p(\mathbf{x}_I, \mathbf{x}_B | \mathcal{C}_k) = p(\mathbf{x}_I | \mathcal{C}_k) p(\mathbf{x}_B | \mathcal{C}_k) \quad (1.84)$$

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

$$\begin{aligned} p(\mathcal{C}_k | \mathbf{x}_I, \mathbf{x}_B) &\propto p(\mathbf{x}_I, \mathbf{x}_B | \mathcal{C}_k) p(\mathcal{C}_k) \\ &\propto p(\mathbf{x}_I | \mathcal{C}_k) p(\mathbf{x}_B | \mathcal{C}_k) p(\mathcal{C}_k) \\ &\propto \frac{p(\mathcal{C}_k | \mathbf{x}_I) p(\mathcal{C}_k | \mathbf{x}_B)}{p(\mathcal{C}_k)} \end{aligned} \quad (1.85)$$

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 $p(\mathbf{x}_I, \mathbf{x}_B)$ 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 (Loss functions for regression)

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 \mathbf{x} 选择一个关于 t 的具体估计 $y(\mathbf{x})$ 。假设这样做会带来一个损失 $L(t, y(\mathbf{x}))$ 。那么平均（期望）损失可以表示为

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

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 $L(t, y(\mathbf{x})) = \{y(\mathbf{x}) - t\}^2$ 。在这种情况下，期望损失可以写作

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

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

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

解出 $y(\mathbf{x})$ ，并结合概率的求和法则和乘法法则，我们得到这是在给定 \mathbf{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 $y(\mathbf{x}) = \mathbb{E}[\mathbf{t} | \mathbf{x}]$ 。

$$y(\mathbf{x}) = \frac{\int t p(\mathbf{x}, t) dt}{p(\mathbf{x})} = \int t p(t | \mathbf{x}) dt = \mathbb{E}_t[t | \mathbf{x}] \quad (1.89)$$

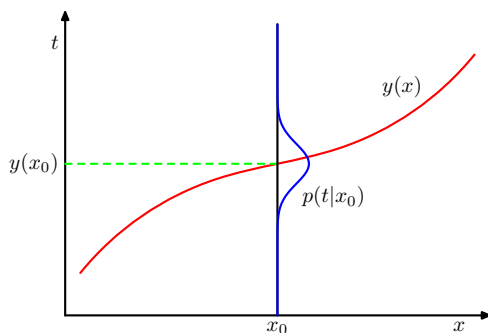


图 1.28 使期望平方损失最小的回归函数 $y(x)$ 由条件分布 $p(t|x)$ 的均值给出。

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：

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

为使记号更简洁，我们以 $\mathbb{E}[t | \mathbf{x}]$ 表示 $\mathbb{E}_t[t | \mathbf{x}]$ 。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

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

我们要求解的函数 $y(\mathbf{x})$ 只出现在第一项中。当 $y(\mathbf{x}) = \mathbb{E}[t | \mathbf{x}]$ 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 $p(t | \mathbf{x})$ ，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 $p(t | \mathbf{x})$ ，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 $p(t | \mathbf{x})$ 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

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

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 $q \rightarrow 0$ 时则是条件众数。

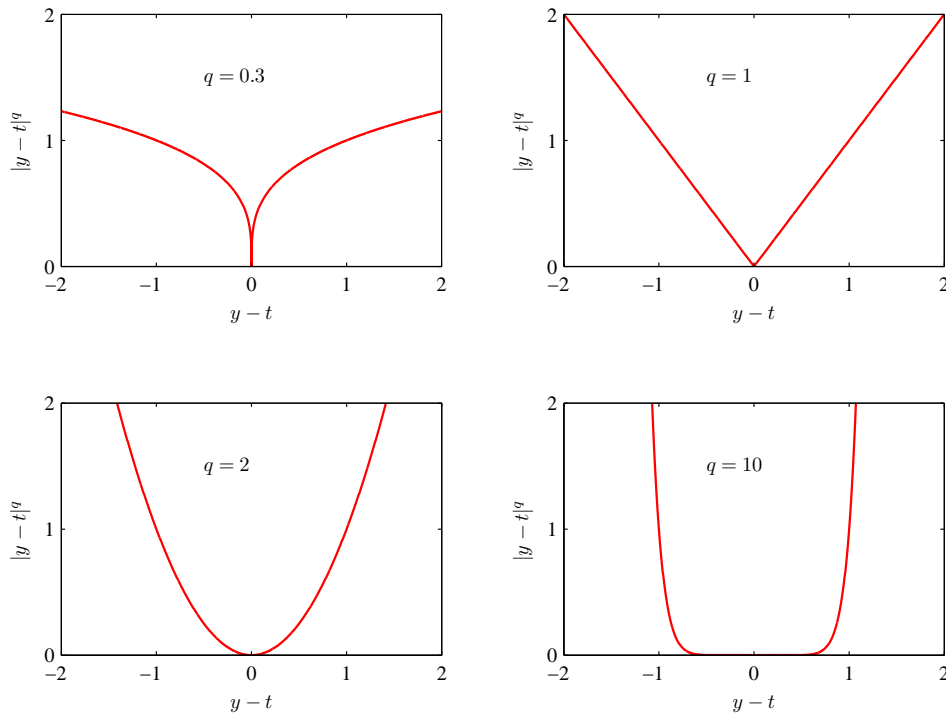


图 1.29 不同 q 取值下量 $L_q = |y - t|^q$ 的曲线图。

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

$$h(x) = -\log_2 p(x) \quad (1.92)$$

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

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

这个重要的量被称为随机变量 x 的熵（entropy）。注意到 $\lim_{p \rightarrow 0} p \ln p = 0$ ，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力。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 $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$ 。在这种情况下，熵为

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

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, 111110, 111111。此时所需传输的平均码长为

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

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

$$W = \frac{N!}{\prod_i n_i!} \quad (1.94)$$

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：

$$H = \frac{1}{N} \ln W = \frac{1}{N} \ln N! - \frac{1}{N} \sum_i \ln n_i! \quad (1.95)$$

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

$$\ln N! \simeq N \ln N - N \quad (1.96)$$

由此可得

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

这里我们利用了 $\sum_i n_i = N$ 。其中 $p_i = \lim_{N \rightarrow \infty} \left(\frac{n_i}{N} \right)$ 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

$$H[p] = - \sum_i p(x_i) \ln p(x_i) \quad (1.98)$$

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 $0 \leq p_i \leq 1$ ，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：

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

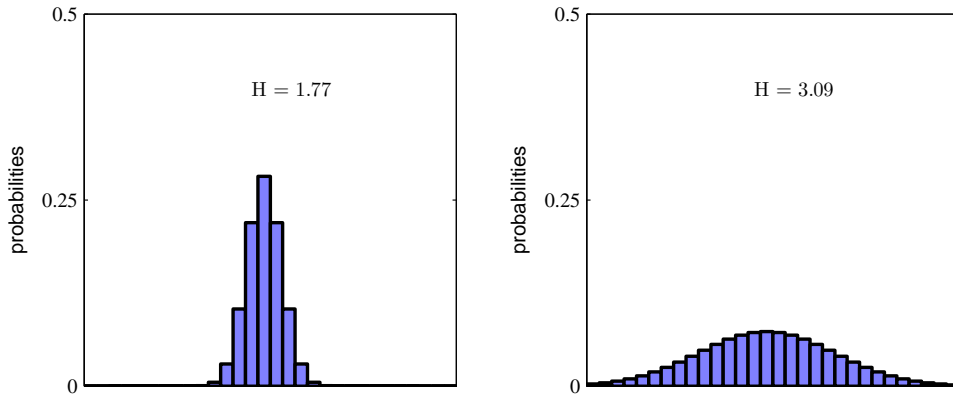


图 1.30 两个在 30 个区间上的概率分布的直方图，用来说明更宽的分布具有更大的熵 H 。最大熵出现在均匀分布的情况下，此时 $H = -\ln\left(\frac{1}{30}\right) = 3.40$ 。

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 $H = \ln M$ 。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

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

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 $p(x)$ 连续的前提下，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

$$\int_{i\Delta}^{(i+1)\Delta} p(x) dx = p(x_i) \Delta \quad (1.101)$$

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

$$H_\Delta = -\sum_i p(x_i) \Delta \ln(p(x_i) \Delta) = -\sum_i p(x_i) \Delta \ln p(x_i) - \ln \Delta \quad (1.102)$$

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 $\Delta \rightarrow 0$ 。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

$$\lim_{\Delta \rightarrow 0} \left\{ \sum_i p(x_i) \Delta \ln p(x_i) \right\} = - \int p(x) \ln p(x) dx \quad (1.103)$$

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 $\Delta \rightarrow 0$ 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

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

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：

$$\int p(x) dx = 1 \quad (1.105)$$

$$\int x p(x) dx = \mu \quad (1.106)$$

$$\int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx = \sigma^2 \quad (1.107)$$

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：

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

使用变分法，令该泛函对 $p(x)$ 的变分导数为零，可得

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

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

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

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

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

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 $p(x, y)$ ，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 $-\ln p(y | x)$ 。因此，确定 y 所需的平均额外信息量可以写为

$$H[y|x] = - \iint p(y, x) \ln p(y | x) dx dy \quad (1.111)$$

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：

$$H[x, y] = H[y|x] + H[x] \quad (1.112)$$

其中 $H[x, y]$ 是联合分布 $p(x, y)$ 的微分熵，而 $H[x]$ 是边缘分布 $p(x)$ 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

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

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 $KL(p \parallel q) \neq KL(q \parallel p)$ 。

我们现在来证明 KL 散度满足 $KL(p \parallel q) \geq 0$ ，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 $\mathbf{x} = \lambda a + (1 - \lambda)b$ ， $0 \leq \lambda \leq 1$ 。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 $f(\lambda a + (1 - \lambda)b)$ 。凸性意味着

$$f(\lambda a + (1 - \lambda)b) \leq \lambda f(a) + (1 - \lambda)f(b) \quad (1.114)$$

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

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

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，

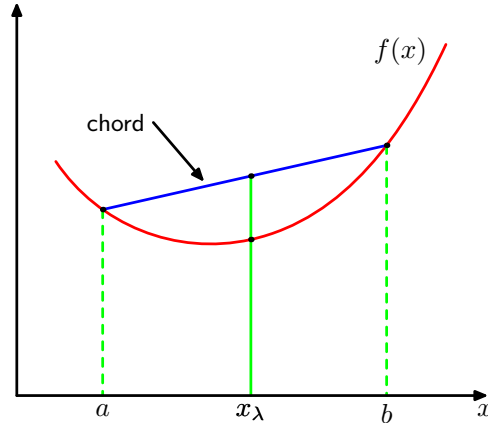


图 1.31 凸函数 $f(x)$ 的定义是：任意弦（蓝色部分）都位于函数曲线（红色部分）的上方或重合。

那么式 (1.115) 可以写作

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

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

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

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

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

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 $\boldsymbol{\theta}$ 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

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

在式 (1.119) 的右边，第二项与参数 $\boldsymbol{\theta}$ 无关，而第一项正是分布 $q(\mathbf{x}|\boldsymbol{\theta})$ 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：

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

这被称为变量 \mathbf{x} 和 \mathbf{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

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

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 $\mathbf{x}_1, \dots, \mathbf{x}_N$ 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 $\mathbf{x}_1, \dots, \mathbf{x}_N$ 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

$$p(x = 1 \mid \mu) = \mu \quad (2.1)$$

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

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

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

$$\mathbb{E}[x] = \mu \quad (2.3)$$

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

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

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

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

$$\ln p(\mathcal{D} | \mu) = \sum_{n=1}^N \ln p(x_n | \mu) = \sum_{n=1}^N \{x_n \ln \mu + (1 - x_n) \ln(1 - \mu)\} \quad (2.6)$$

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

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

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

$$\mu_{\text{ML}} = \frac{m}{N} \quad (2.8)$$

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 ($\mu_{\text{ML}} = 1$)。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：

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

其中

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

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。

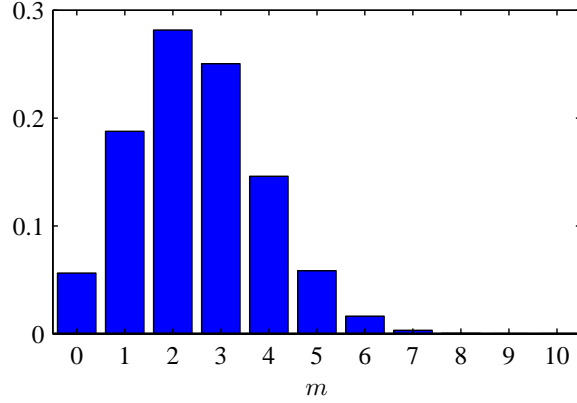


图 2.1 直方图绘制了二项分布 (2.9) 作为 m 的函数，其中 $N = 10$ 且 $\mu = 0.25$ 。

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 $m = x_1 + \dots + x_N$ ，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：

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

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

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

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

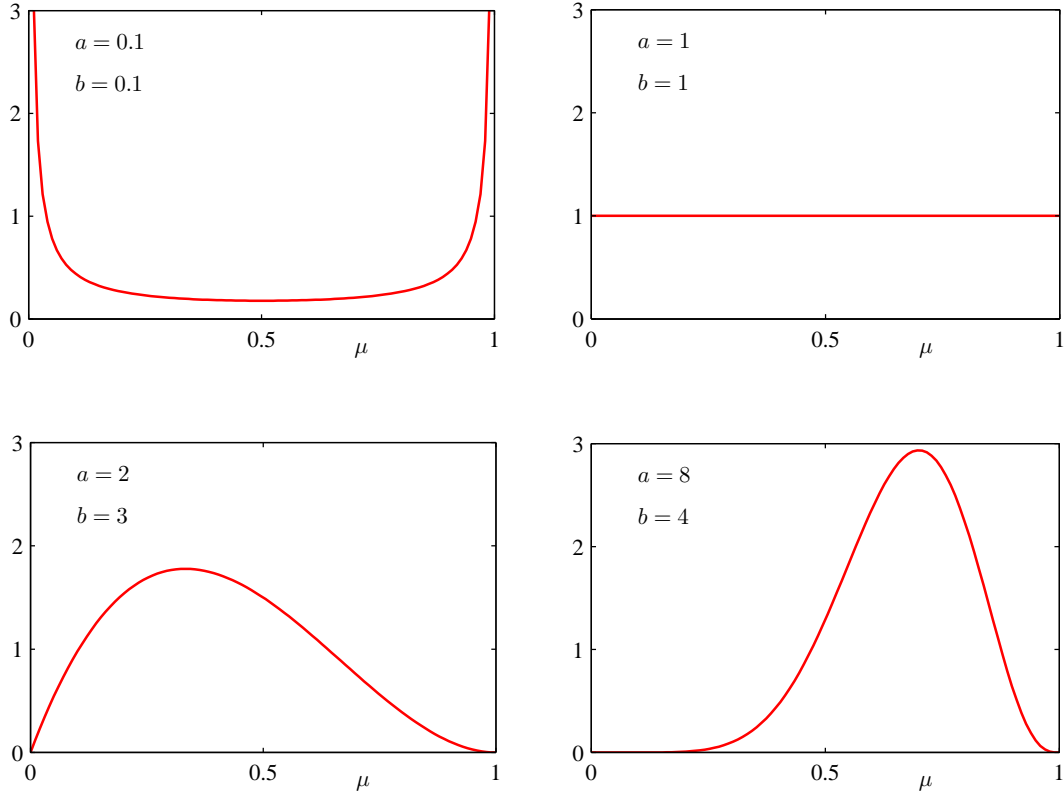


图 2.2 贝塔分布 ($\text{Beta}(\mu | a, b)$) 的曲线图, 展示了式 (2.13) 所定义的分佈随参数 (μ) 的变化情况, 其中 (a) 和 (b) 取不同的超参数值。

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

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

贝塔分布的均值和方差为

$$\mathbb{E}[\mu] = \frac{a}{a+b} \quad (2.15)$$

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

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

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

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

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)与先验分佈在 (μ) 上具有相同的函数形式, 这反映了该先验分佈相对于似然函数的共轭性质。事实上, 它依然是一个贝塔分佈, 其归一化系数可通过与式(2.13) 对比得

到：

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

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。

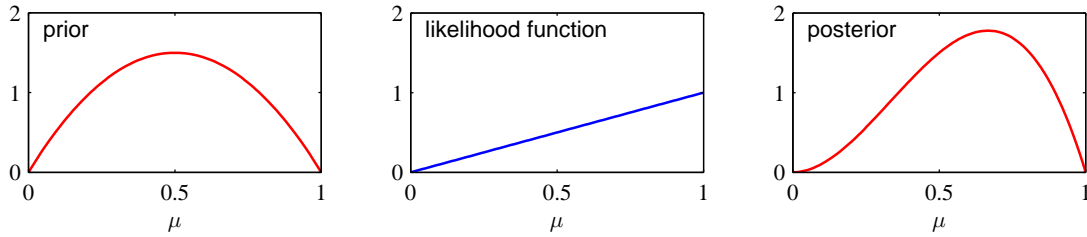


图 2.3 顺序贝叶斯推断过程中的单步示意图。先验分布由参数 $a = 2$ 、 $b = 2$ 的 Beta 分布给出；似然函数由公式 (2.9) 给出，其中 $N = m = 1$ ，对应于一次 $x = 1$ 的观测，因此后验分布为参数 $a = 3$ 、 $b = 2$ 的 Beta 分布。

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：

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

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

我们得到

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

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 $m, l \rightarrow \infty$ 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 $a \rightarrow \infty$ 或 $b \rightarrow \infty$ 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：

$$\mathbb{E}_{\theta}[\theta] = \mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\theta}[\theta | \mathcal{D}]] \quad (2.21)$$

其中

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

$$\mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\theta}[\theta | \mathcal{D}]] \equiv \int \left\{ \int \theta p(\theta | \mathcal{D}) d\theta \right\} p(\mathcal{D}) d\mathcal{D} \quad (2.23)$$

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

$$\text{var}_{\theta}[\theta] = \mathbb{E}_{\mathcal{D}}[\text{var}_{\theta}[\theta | \mathcal{D}]] + \text{var}_{\mathcal{D}}[\mathbb{E}_{\theta}[\theta | \mathcal{D}]] \quad (2.24)$$

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

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

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

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

其中 $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)^T$ ，并且参数满足约束条件 $\mu_k \geq 0$ 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：

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

并且有期望

$$\mathbb{E}[\mathbf{x} | \boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x} | \boldsymbol{\mu}) \mathbf{x} = (\mu_1, \dots, \mu_K)^T = \boldsymbol{\mu}. \quad (2.28)$$

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

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

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：

$$m_k = \sum_n x_{nk}, \quad (2.30)$$

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 $\ln p(\mathcal{D} | \boldsymbol{\mu})$ 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：

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

对式 (2.31) 关于 μ_k 求导并令其为零，得到

$$\mu_k = -\frac{m_k}{\lambda}. \quad (2.32)$$

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

$$\mu_k^{\text{ML}} = \frac{m_k}{N}, \quad (2.33)$$

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, \dots, m_K 的联合分布。根据式 (2.29)，该分布可写为

$$\text{Mult}(m_1, m_2, \dots, m_K | \boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \dots m_K} \prod_{k=1}^K \mu_k^{m_k}, \quad (2.34)$$

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, \dots, m_K 的 K 个组的不同方式的数量, 其形式为

$$\binom{N}{m_1 m_2 \dots m_K} = \frac{N!}{m_1! m_2! \dots m_K!}. \quad (2.35)$$

注意, 变量 m_k 满足约束条件

$$\sum_{k=1}^K m_k = N. \quad (2.36)$$

2.2.1 狄利克雷分布 (The Dirichlet Distribution)

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

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

其中 $0 \leq \mu_k \leq 1$ 且 $\sum_k \mu_k = 1$ 。这里的 $\alpha_1, \dots, \alpha_K$ 是该分布的参数, 记作 $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)^T$ 。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。

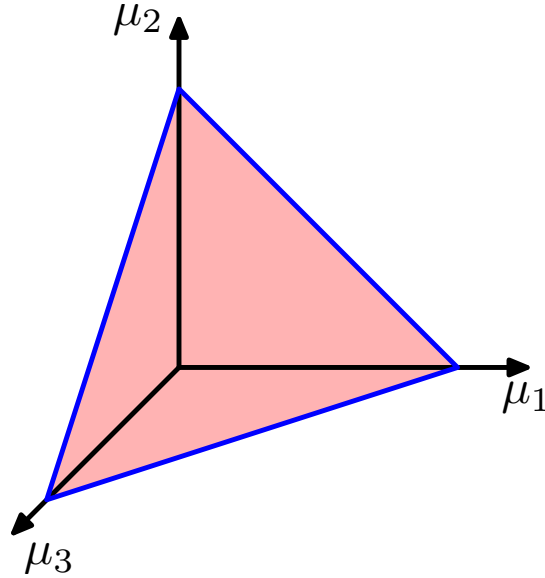


图 2.4 对三个变量 μ_1, μ_2, μ_3 的 Dirichlet 分布被限制在一个如图所示的单纯形 (有界线性流形) 中, 这是由约束条件 $0 \leq \mu_k \leq 1$ 和 $\sum_k \mu_k = 1$ 所导致的。

该分布的归一化形式为

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

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

$$\alpha_0 = \sum_{k=1}^K \alpha_k. \quad (2.39)$$

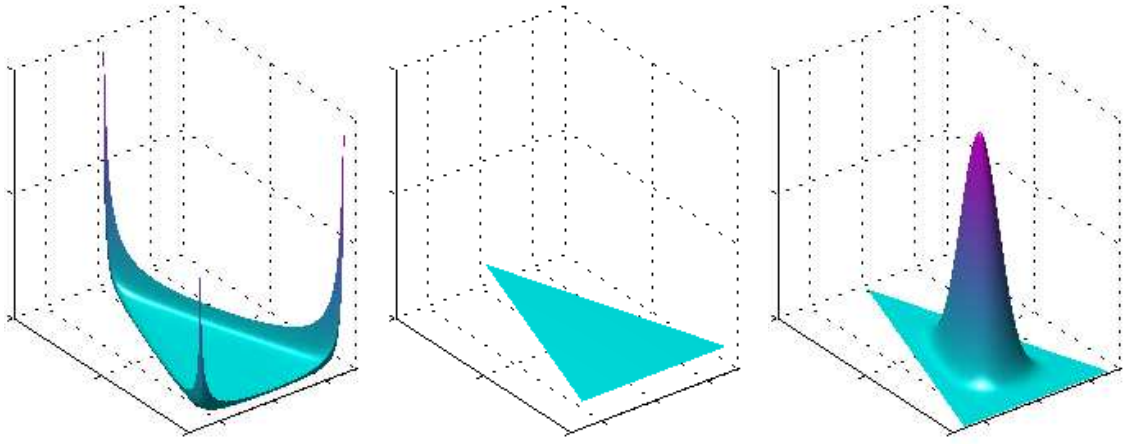


图 2.5 三元 Dirichlet 分布的图示中，两个水平坐标轴位于单纯形平面内，垂直轴表示密度值。左图中 $\{\alpha_k\} = 0.1$ ，中图中 $\{\alpha_k\} = 1$ ，右图中 $\{\alpha_k\} = 10$ 。

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：

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

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：

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

其中 $\mathbf{m} = (m_1, \dots, m_K)^T$ 。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

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

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

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

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, \dots, x_N ，它们在区间 $[0, 1]$ 上均匀分布，我们关心它们的平均值 $\frac{x_1 + x_2 + \dots + x_N}{N}$ 。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 $N \rightarrow \infty$ 时也将趋近于高斯分布（见图 2.1 中 $N = 10$ 的情况）。

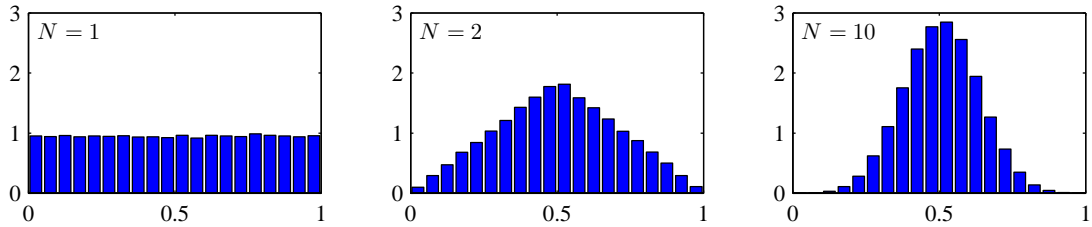


图 2.6 对于不同取值的 N ，绘制由 N 个均匀分布数值的均值所构成的直方图。可以观察到，随着 N 的增大，该分布趋近于高斯分布。

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：

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

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：

$$\boldsymbol{\Sigma} \mathbf{u}_i = \lambda_i \mathbf{u}_i, \quad (2.45)$$

其中 $i = 1, \dots, D$ 。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

$$\mathbf{u}_i^T \mathbf{u}_j = I_{ij}, \quad (2.46)$$

其中 I_{ij} 是单位矩阵的第 (i, j) 个元素, 满足

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

协方差矩阵 Σ 可以用其特征向量展开为

$$\Sigma = \sum_{i=1}^D \lambda_i \mathbf{u}_i \mathbf{u}_i^T, \quad (2.48)$$

而其逆矩阵 Σ^{-1} 则可写为

$$\Sigma^{-1} = \sum_{i=1}^D \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T. \quad (2.49)$$

将上述式子代入 (2.44), 可得该二次型为

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

其中我们定义

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

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 $\mathbf{y} = (y_1, \dots, y_D)^T$, 则有

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

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

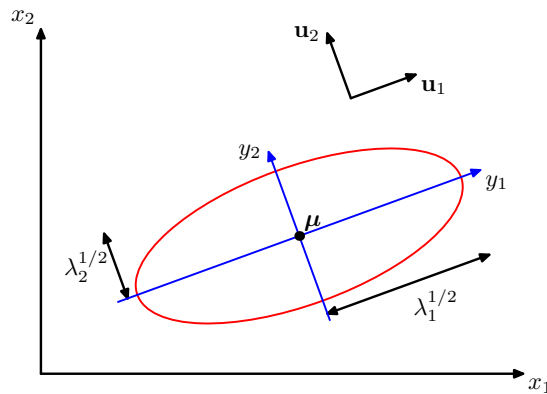


图 2.7 红色曲线表示二维空间 $\mathbf{x} = (x_1, x_2)$ 中高斯分布的等概率密度椭圆, 其密度为在 $\mathbf{x} = \boldsymbol{\mu}$ 处取值的 $\exp(-1/2)$ 。该椭圆的主轴由协方差矩阵的特征向量 \mathbf{u}_i 定义, 且对应的特征值为 λ_i 。

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

$$J_{ij} = \frac{\partial x_i}{\partial y_j} = U_{ji}, \quad (2.53)$$

其中 U_{ji} 是矩阵 \mathbf{U}^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

$$|\mathbf{J}|^2 = |\mathbf{U}^T|^2 = |\mathbf{U}^T|, |\mathbf{U}| = |\mathbf{U}^T \mathbf{U}| = |\mathbf{I}| = 1, \quad (2.54)$$

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

$$|\Sigma|^{1/2} = \prod_{j=1}^D \lambda_j^{1/2}. \quad (2.55)$$

因此，在 y_j 坐标系下，高斯分布可以写为

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

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

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

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 Σ 的意义。在高斯分布下， \mathbf{x} 的期望为

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

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

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}, \quad (2.59)$$

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 $\mathbb{E}[\mathbf{x}\mathbf{x}^T]$ 。该矩阵可写为

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

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 $\boldsymbol{\mu}\mathbf{z}^T$ 和 $\boldsymbol{\mu}^T\mathbf{z}$ 的交叉项会由于对称性而消失。项 $\boldsymbol{\mu}\boldsymbol{\mu}^T$ 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 $\mathbf{z}\mathbf{z}^T$ 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

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

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

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

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

$$\mathbb{E}[\mathbf{x}\mathbf{x}^T] = \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}. \quad (2.62)$$

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

$$\text{cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T]. \quad (2.63)$$

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

$$\text{cov}[\mathbf{x}] = \boldsymbol{\Sigma}. \quad (2.64)$$

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 $\boldsymbol{\Sigma} = \text{diag}(\sigma_i^2)$ ，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。

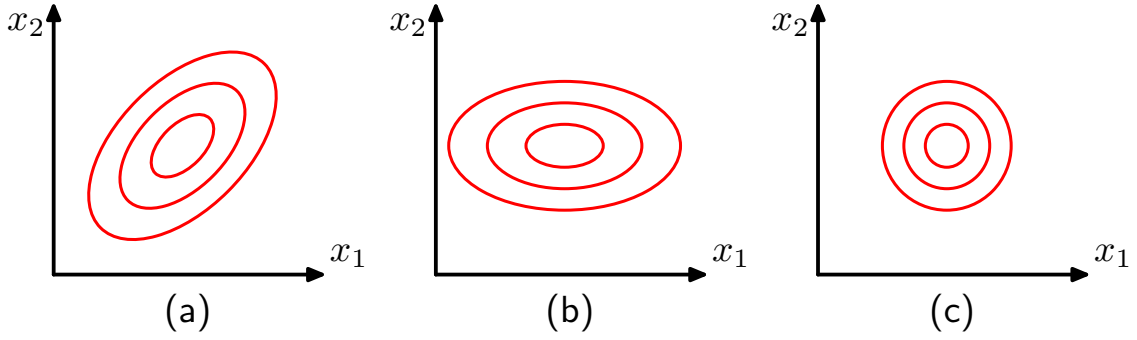


图 2.8 二维高斯分布的等概率密度轮廓：(a) 协方差矩阵为一般形式；(b) 协方差矩阵为对角形式，此时椭圆轮廓与坐标轴对齐；(c) 协方差矩阵与单位矩阵成比例，此时轮廓为同心圆。

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 $\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma})$ ，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

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}. \quad (2.65)$$

相应地，均值向量 $\boldsymbol{\mu}$ 也被划分为

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \quad (2.66)$$

而协方差矩阵 Σ 被分块写为

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}. \quad (2.67)$$

由于协方差矩阵是对称的, 即 $\Sigma^T = \Sigma$, 因此 Σ_{aa} 和 Σ_{bb} 都是对称矩阵, 且有 $\Sigma_{ba} = \Sigma_{ab}^T$. 在许多情况下, 使用协方差矩阵的逆会更加方便, 我们定义

$$\Lambda \equiv \Sigma^{-1}, \quad (2.68)$$

称 Λ 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:

$$\Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}. \quad (2.69)$$

由于对称矩阵的逆仍然是对称的, 因此 Λ_{aa} 和 Λ_{bb} 也是对称矩阵, 并且有 $\Lambda_{ab}^T = \Lambda_{ba}$. 需要强调的是, Λ_{aa} 并不等于 Σ_{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 $p(\mathbf{x}_a | \mathbf{x}_b)$ 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

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

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 $p(\mathbf{x}_a | \mathbf{x}_b)$ 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 $p(\mathbf{x}_a | \mathbf{x}_b)$ 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 $\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \Sigma)$, 其指数项可写为

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

其中 “const” 表示与 \mathbf{x} 无关的常数项, 并且利用了 Σ 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 Σ^{-1} ; 线性项的系数为 $\Sigma^{-1}\boldsymbol{\mu}$, 从而可以求出 $\boldsymbol{\mu}$ 。

现在将这一方法应用到条件高斯分布 $p(\mathbf{x}_a | \mathbf{x}_b)$ 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

$$-\frac{1}{2} \mathbf{x}_a^T \boldsymbol{\Lambda}_{aa} \mathbf{x}_a. \quad (2.72)$$

由此我们可以直接得到条件分布的协方差（或逆精度）为

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

现在考虑式 (2.70) 中所有关于 \mathbf{x}_a 的一次项

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

其中我们使用了 $\boldsymbol{\Lambda}_{ba}^T = \boldsymbol{\Lambda}_{ab}$ 。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

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

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

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix}, \quad (2.76)$$

其中

$$\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}. \quad (2.77)$$

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

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

并结合 (2.76)，得到

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

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

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

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

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

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 $p(\mathbf{x}_a | \mathbf{x}_b)$ 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 $p(\mathbf{x}_a | \mathbf{x}_b)$ 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：

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

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

$$-\frac{1}{2}\mathbf{x}_b^T \Lambda_{bb} \mathbf{x}_b + \mathbf{x}_b^T \mathbf{m} = -\frac{1}{2}(\mathbf{x}_b - \Lambda_{bb}^{-1} \mathbf{m})^T \Lambda_{bb} (\mathbf{x}_b - \Lambda_{bb}^{-1} \mathbf{m}) + \frac{1}{2}\mathbf{m}^T \Lambda_{bb}^{-1} \mathbf{m} \quad (2.84)$$

其中我们定义了

$$\mathbf{m} = \Lambda_{bb} \boldsymbol{\mu}_b - \Lambda_{ba} (\mathbf{x}_a - \boldsymbol{\mu}_a) \quad (2.85)$$

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：

$$\int \exp \left\{ -\frac{1}{2}(\mathbf{x}_b - \Lambda_{bb}^{-1} \mathbf{m})^T \Lambda_{bb} (\mathbf{x}_b - \Lambda_{bb}^{-1} \mathbf{m}) \right\} d\mathbf{x}_b \quad (2.86)$$

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

$$\begin{aligned} & \frac{1}{2} [\Lambda_{bb} \boldsymbol{\mu}_b - \Lambda_{ba} (\mathbf{x}_a - \boldsymbol{\mu}_a)]^T \Lambda_{bb}^{-1} [\Lambda_{bb} \boldsymbol{\mu}_b - \Lambda_{ba} (\mathbf{x}_a - \boldsymbol{\mu}_a)] \\ & - \frac{1}{2} \mathbf{x}_a^T \Lambda_{aa} \mathbf{x}_a + \mathbf{x}_a^T (\Lambda_{aa} \boldsymbol{\mu}_a + \Lambda_{ab} \boldsymbol{\mu}_b) + \text{const} \\ & = -\frac{1}{2} \mathbf{x}_a^T (\Lambda_{aa} - \Lambda_{ab} \Lambda_{bb}^{-1} \Lambda_{ba}) \mathbf{x}_a \\ & + \mathbf{x}_a^T (\Lambda_{aa} - \Lambda_{ab} \Lambda_{bb}^{-1} \Lambda_{ba})^{-1} \boldsymbol{\mu}_a + \text{const} \end{aligned} \quad (2.87)$$

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

$$\Sigma_a = (\Lambda_{aa} - \Lambda_{ab} \Lambda_{bb}^{-1} \Lambda_{ba})^{-1} \quad (2.88)$$

类似地，均值由下式给出

$$\Sigma_a (\Lambda_{aa} - \Lambda_{ab} \Lambda_{bb}^{-1} \Lambda_{ba}) \boldsymbol{\mu}_a = \boldsymbol{\mu}_a \quad (2.89)$$

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

$$\begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \quad (2.90)$$

利用 (2.76) 式, 我们得到

$$(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1} = \Sigma_{aa} \quad (2.91)$$

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

$$\mathbb{E}[\mathbf{x}_a] = \boldsymbol{\mu}_a \quad (2.92)$$

$$\text{cov}[\mathbf{x}_a] = \Sigma_{aa}. \quad (2.93)$$

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 $\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \Sigma)$, 其中 $\Lambda \equiv \Sigma^{-1}$ 以及

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

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}. \quad (2.95)$$

条件分布:

$$p(\mathbf{x}_a | \mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a | \boldsymbol{\mu}_{a|b}, \Lambda_{aa}^{-1}), \quad (2.96)$$

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

边缘分布:

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a | \boldsymbol{\mu}_a, \Sigma_{aa}). \quad (2.98)$$

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。

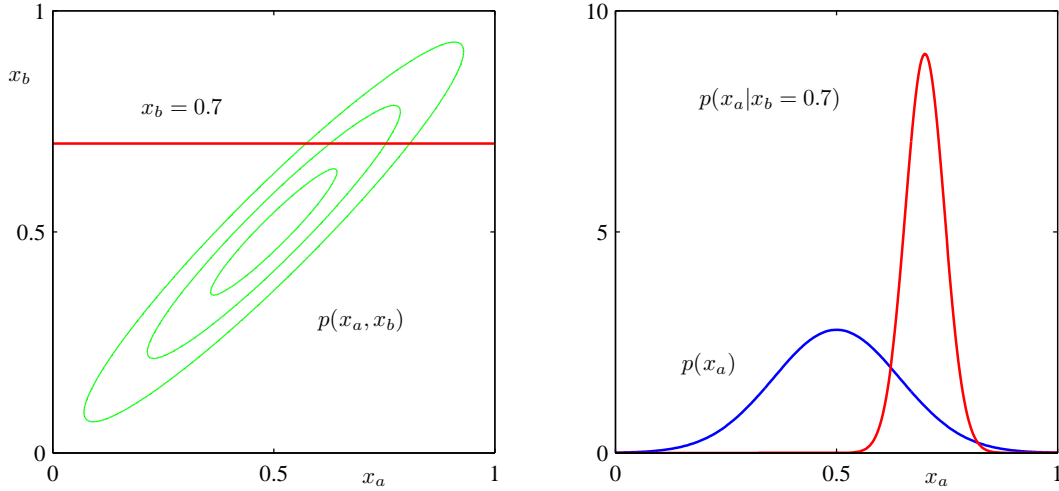


图 2.9 左图显示了关于两个变量的高斯分布 $p(x_a, x_b)$ 的等高线；右图显示了边缘分布 $p(x_a)$ （蓝色曲线）以及在 $x_b = 0.7$ 时的条件分布 $p(x_a | x_b)$ （红色曲线）。

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 $p(\mathbf{x}_a | \mathbf{x}_b)$ 和边缘分布 $p(\mathbf{x}_a)$ 的表达式。我们注意到，条件分布 $p(\mathbf{x}_a | \mathbf{x}_b)$ 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 $p(\mathbf{y} | \mathbf{x})$ ，其中 $p(\mathbf{y} | \mathbf{x})$ 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 $p(\mathbf{x} | \mathbf{y})$ 。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

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

$$p(\mathbf{y} | \mathbf{x}) = \mathcal{N}(\mathbf{y} | \mathbf{Ax} + \mathbf{b}, \mathbf{L}^{-1}) \quad (2.100)$$

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

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \quad (2.101)$$

然后考虑联合分布的对数

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

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

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

因此 \mathbf{z} 上的高斯分布的精度（逆协方差）矩阵为

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

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

$$\text{cov}[\mathbf{z}] = \mathbf{R}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^T \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T \end{pmatrix}. \quad (2.105)$$

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

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

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

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

再利用 (2.105)，得到

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

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

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

$$\text{cov}[\mathbf{y}] = \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T. \quad (2.110)$$

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 $p(\mathbf{x} | \mathbf{y})$ 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 $p(\mathbf{x} | \mathbf{y})$ 的均值和协方差为

$$\mathbb{E}[\mathbf{x} | \mathbf{y}] = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1} \{ \mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu} \}, \quad (2.111)$$

$$\text{cov}[\mathbf{x} | \mathbf{y}] = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}. \quad (2.112)$$

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 $p(\mathbf{x} | \mathbf{y})$ 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 $p(\mathbf{z}) = p(\mathbf{x})p(\mathbf{y} | \mathbf{x})$ 表示成了 $p(\mathbf{x} | \mathbf{y})p(\mathbf{y})$ 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

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \quad (2.113)$$

$$p(\mathbf{y} | \mathbf{x}) = \mathcal{N}(\mathbf{y} | \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \quad (2.114)$$

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

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} | \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \quad (2.115)$$

$$p(\mathbf{x} | \mathbf{y}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\Sigma} \{ \mathbf{A}^T \mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu} \}, \boldsymbol{\Sigma}) \quad (2.116)$$

其中

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^T \mathbf{L} \mathbf{A})^{-1} \quad (2.117)$$

2.3.4 高斯分布的最大似然估计

给定数据集 $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^T$ ，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

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

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

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

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

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

将该导数设为零，得到均值的最大似然估计解为

$$\boldsymbol{\mu}_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \quad (2.121)$$

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

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

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

如果我们在真实分布下对最大似然解取期望，得到

$$\mathbb{E}[\boldsymbol{\mu}_{\text{ML}}] = \boldsymbol{\mu}, \quad (2.123)$$

$$\mathbb{E}[\Sigma_{\text{ML}}] = \frac{N-1}{N} \Sigma. \quad (2.124)$$

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

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

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

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 $\boldsymbol{\mu}_{\text{ML}}^{(N)}$ 。如果把最后一个数据点 \mathbf{x}_N 的贡献单独拆出来，可得

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

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 $\boldsymbol{\mu}_{\text{ML}}^{(N-1)}$ 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” $(\mathbf{x}_N - \boldsymbol{\mu}_{\text{ML}}^{(N-1)})$ 的方向移动一小步（步长为 $1/N$ ），得到新的估计 $\boldsymbol{\mu}_{\text{ML}}^{(N)}$ 。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 $f(\theta)$:

$$f(\theta) \equiv \mathbb{E}[z | \theta] = \int z p(z | \theta) dz, \quad (2.127)$$

如图 2.10 所示。这样定义的函数称为回归函数。

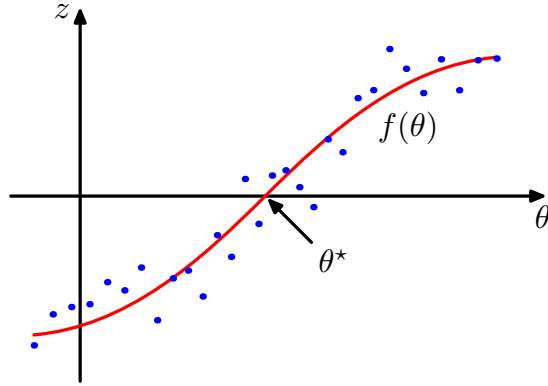


图 2.10 两个相关随机变量 z 和 θ 的示意图, 同时给出了回归函数 $f(\theta)$, 该函数由条件期望 $\mathbb{E}[z | \theta]$ 定义。Robbins-Monro 算法提供了一种通用的序列化方法, 用于求解此类函数的根 θ^* 。

我们的目标是寻找根 θ^* , 使得 $f(\theta^*) = 0$ 。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 θ^* 。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

$$\mathbb{E}[(z - f)^2 | \theta] < \infty, \quad (2.128)$$

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

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

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

$$\lim_{N \rightarrow \infty} a_N = 0, \quad (2.130)$$

$$\sum_{N=1}^{\infty} a_N = \infty, \quad (2.131)$$

$$\sum_{N=1}^{\infty} a_N^2 < \infty. \quad (2.132)$$

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 θ^* 。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

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

交换求导与求和顺序并取极限 $N \rightarrow \infty$ ，得到

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

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

$$\theta^{(N)} = \theta^{(N-1)} + a_{N-1} \frac{\partial}{\partial \theta^{(N-1)}} \ln p(x_N | \theta^{(N-1)}). \quad (2.135)$$

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 $\mu_{\text{ML}}^{(N)}$ ，随机变量 z 为

$$z = \frac{\partial}{\partial \mu_{\text{ML}}} \ln p(x | \mu_{\text{ML}}, \sigma^2) = \frac{1}{\sigma^2} (x - \mu_{\text{ML}}). \quad (2.136)$$

因此 z 的分布是以 $\mu - \mu_{\text{ML}}$ 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 μ 和 Σ 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 $\mathbf{X} = \{x_1, \dots, x_N\}$ 推断均值 μ 。似然函数（即给定 μ 时观测数据的概率，作为 μ 的函数）为

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

再次强调，似然函数 $p(\mathbf{X} | \mu)$ 并不是关于 μ 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

$$p(\mu) = \mathcal{N}(\mu | \mu_0, \sigma_0^2), \quad (2.138)$$

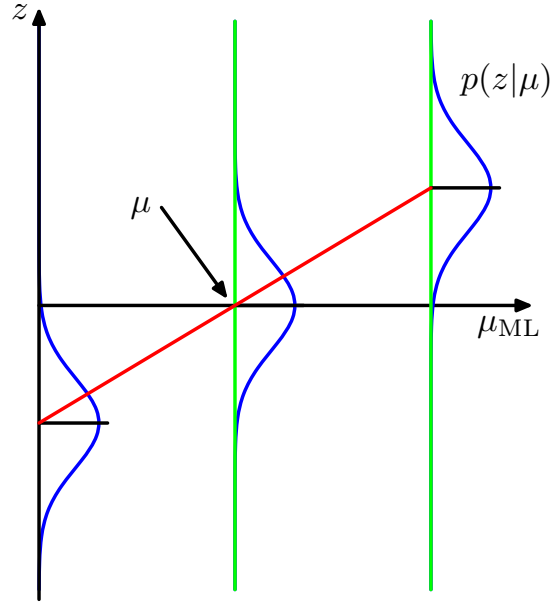


图 2.11 在高斯分布的情形下，令 θ 对应均值 μ ，则图 2.10 中所示的回归函数呈现为一条直线（红色所示）。此时随机变量 z 对应于对数似然函数的导数，表示为 $(x - \mu_{\text{ML}})/\sigma^2$ ，其期望（即定义回归函数的量）为 $(\mu - \mu_{\text{ML}})/\sigma^2$ ，同样是一条直线。该回归函数的根对应最大似然估计量 μ_{ML} 。

后验分布为

$$p(\mu | \mathbf{X}) \propto p(\mathbf{X} | \mu)p(\mu). \quad (2.139)$$

通过在指数中完成平方，简单推导可得后验分布为

$$p(\mu | \mathbf{X}) = \mathcal{N}(\mu | \mu_N, \sigma_N^2), \quad (2.140)$$

其中

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\text{ML}}, \quad (2.141)$$

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

这里 μ_{ML} 是样本均值给出的最大似然解

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n. \quad (2.143)$$

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 $N \rightarrow \infty$ 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 $N \rightarrow \infty$ 时，方差 $\sigma_N^2 \rightarrow 0$ ，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 $\sigma_0^2 \rightarrow \infty$ （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。

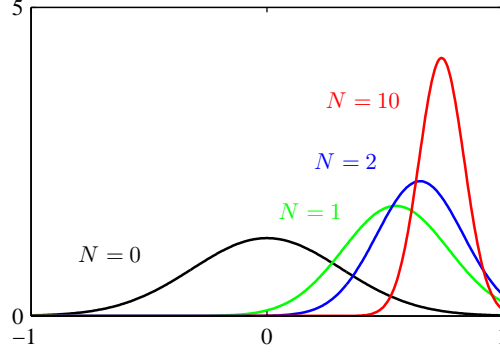


图 2.12 高斯分布均值 μ 的贝叶斯推断示意图，其中方差假定已知。图中曲线展示了对 μ 的先验分布（标记为 $N = 0$ 的曲线），该先验本身为高斯分布；同时还展示了随着数据点数量 N 增加，由式 (2.140) 给出的后验分布。数据点来自均值为 0.8、方差为 0.1 的高斯分布，先验均值取为 0。在先验和似然函数中，方差均设为真实值。

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

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

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

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

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

$$\text{Gam}(\lambda | a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda). \quad (2.146)$$

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 $a \geq 1$ 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

$$\mathbb{E}[\lambda] = \frac{a}{b}, \quad (2.147)$$

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

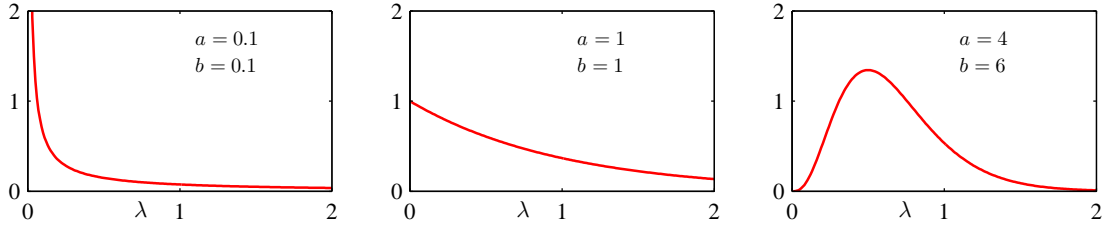


图 2.13 根据式 (2.146) 定义的伽马分布 $\text{Gam}(\lambda | a, b)$ 在不同参数 a 和 b 取值下的绘图。

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

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

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

$$a_N = a_0 + \frac{N}{2}, \quad (2.150)$$

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

这里 σ_{ML}^2 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 $N\sigma_{\text{ML}}^2/2$ ，其中 σ_{ML}^2 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

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

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 $p(\mu, \lambda) = p(\mu | \lambda)p(\lambda)$ ，我们可直接看出： $p(\mu | \lambda)$ 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

$$p(\mu, \lambda) = \mathcal{N}(\mu | \mu_0, (\beta\lambda)^{-1}) \text{Gam}(\lambda | a, b), \tag{2.154}$$

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。

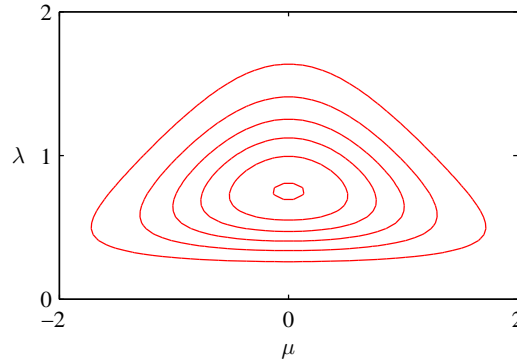


图 2.14 对参数取值 $\mu_0 = 0$ ， $\beta = 2$ ， $a = 5$ ， $b = 6$ 的正态-伽马分布 (??) 的等高线图。

对于 D 维变量 \mathbf{x} 的多元高斯分布 $\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ ，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：

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

其中 ν 称为自由度 (degrees of freedom)， \mathbf{W} 是 $D \times D$ 的尺度矩阵 (scale matrix)， $\text{Tr}(\cdot)$ 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 μ 和精度矩阵 Λ 均未知时，仿照单变量情形的推导，共轭先验为

$$p(\mu, \Lambda \mid \mu_0, \beta, \mathbf{W}, \nu) = \mathcal{N}(\mu \mid \mu_0, (\beta\Lambda)^{-1}) \mathcal{W}(\Lambda \mid \mathbf{W}, \nu), \quad (2.157)$$

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 $\text{Gam}(\tau \mid a, b)$ ，将精度 τ 边缘化（积分消掉）后，得到 x 的边缘分布为

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

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

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

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 $\nu \rightarrow \infty$ 时， $\text{St}(x \mid \mu, \lambda, \nu) \rightarrow \mathcal{N}(x \mid \mu, \lambda^{-1})$ ，即均值为 μ 、精度为 λ 的高斯分布。

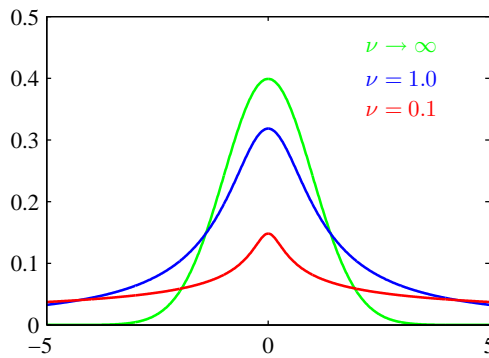


图 2.15 对不同的 ν 值绘制 Student's t 分布 (2.159)，其中 $\mu = 0$ 且 $\lambda = 1$ 。当 $\nu \rightarrow \infty$ 时，该分布趋近于均值为 μ 、精度为 λ 的高斯分布。

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。

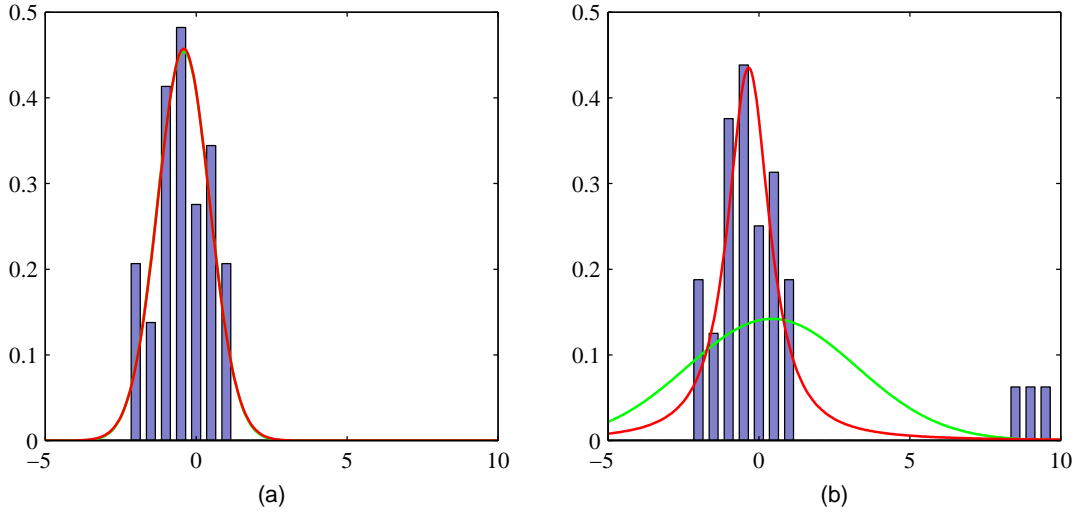


图 2.16 示例说明了与高斯分布相比，Student's t 分布的鲁棒性。(a) 对从高斯分布抽取的 30 个数据点绘制直方图，并分别给出基于 t 分布（红色曲线）和高斯分布（绿色曲线，几乎被红色曲线遮住）的最大似然拟合。由于 t 分布将高斯分布作为特殊情形，因此其结果与高斯分布几乎相同。(b) 使用同一数据集，但额外加入三个离群点，可以看到高斯分布（绿色曲线）受离群点影响严重，而 t 分布（红色曲线）则相对不受影响。

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

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

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

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

采用与单变量情形相同的技巧，对该积分求值得

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

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

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

这就是多元 Student's t 分布，它具有以下性质

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

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

$$\text{mode}[\mathbf{x}] = \boldsymbol{\mu} \quad (2.166)$$

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 $0 \leq \theta < 2\pi$ 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 $\mathcal{D} = \{\theta_1, \dots, \theta_N\}$ 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 $(\theta_1 + \dots + \theta_N)/N$ 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 $\mathbf{x}_1, \dots, \mathbf{x}_N$ 来描述，其中 $\|\mathbf{x}_n\| = 1$ ($n = 1, \dots, N$)，如图 2.17 所示。我们可以对向量 $\{\mathbf{x}_n\}$ 取平均，得到

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \quad (2.167)$$

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， $\bar{\mathbf{x}}$ 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 $\bar{\mathbf{x}} = (\bar{r} \cos \bar{\theta}, \bar{r} \sin \bar{\theta})$ 。代入 (2.167) 并分别对 x_1 和 x_2 分量取等，得到

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

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

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

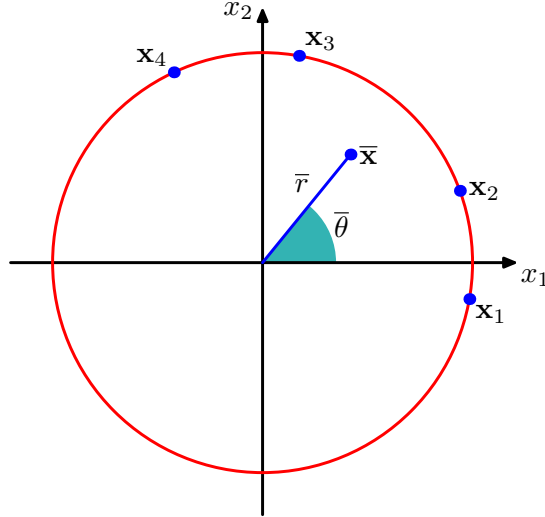


图 2.17 将周期变量的取值 θ_n 表示为位于单位圆上的二维向量 \mathbf{x}_n 的示意图，同时给出了这些向量的平均值 $\bar{\mathbf{x}}$ 。

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：

$$p(\theta) \geq 0 \quad (2.170)$$

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

$$p(\theta + 2\pi) = p(\theta) \quad (2.172)$$

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

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

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

$$x_1 = r \cos \theta, \quad x_2 = r \sin \theta \quad (2.174)$$

我们还将均值 $\boldsymbol{\mu}$ 映射到极坐标，写作

$$\mu_1 = r_0 \cos \theta_0, \quad \mu_2 = r_0 \sin \theta_0 \quad (2.175)$$

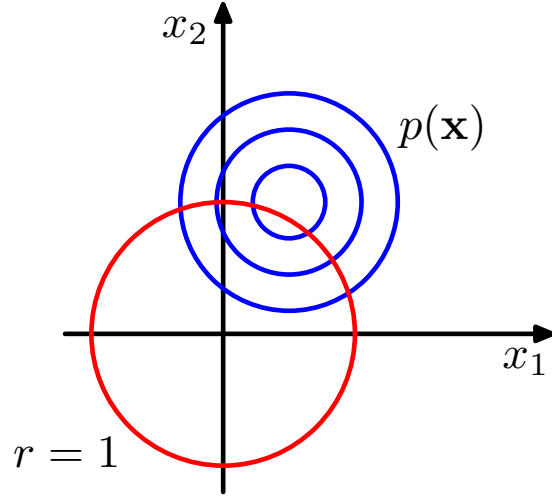


图 2.18 von Mises 分布可以通过考虑形如 (2.173) 的二维高斯分布推导而来，其密度等高线如蓝色所示，并对位于红色单位圆上的点进行条件化。

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

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

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

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

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

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

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

$$\ln p(\mathcal{D} | \theta_0, m) = -N \ln(2\pi) - N \ln I_0(m) + m \sum_{n=1}^N \cos(\theta_n - \theta_0) \tag{2.181}$$

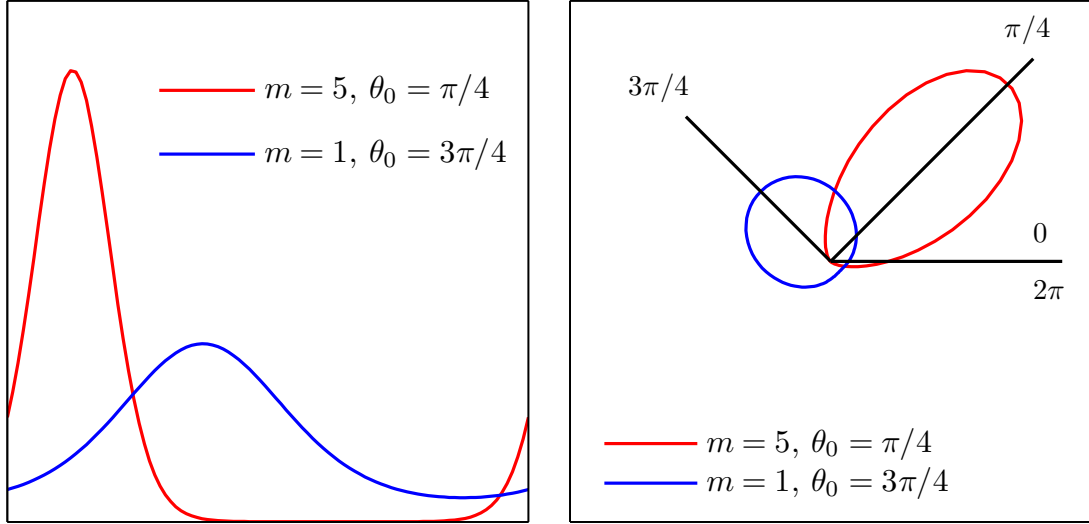


图 2.19 对两组不同参数取值绘制的 von Mises 分布，左图为笛卡尔坐标表示，右图为对应的极坐标表示。

将对 θ_0 的导数设为零，得到

$$\sum_{n=1}^N \sin(\theta_n - \theta_0) = 0 \quad (2.182)$$

为了求解 θ_0 ，我们利用三角恒等式

$$\sin(A - B) = \cos B \sin A - \cos A \sin B \quad (2.183)$$

由此得到

$$\theta_0^{\text{ML}} = \tan^{-1} \left\{ \frac{\sum_{n=1}^N \sin \theta_n}{\sum_{n=1}^N \cos \theta_n} \right\} \quad (2.184)$$

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

$$A(m) = \frac{1}{N} \sum_{n=1}^N \cos(\theta_n - \theta_0^{\text{ML}}) \quad (2.185)$$

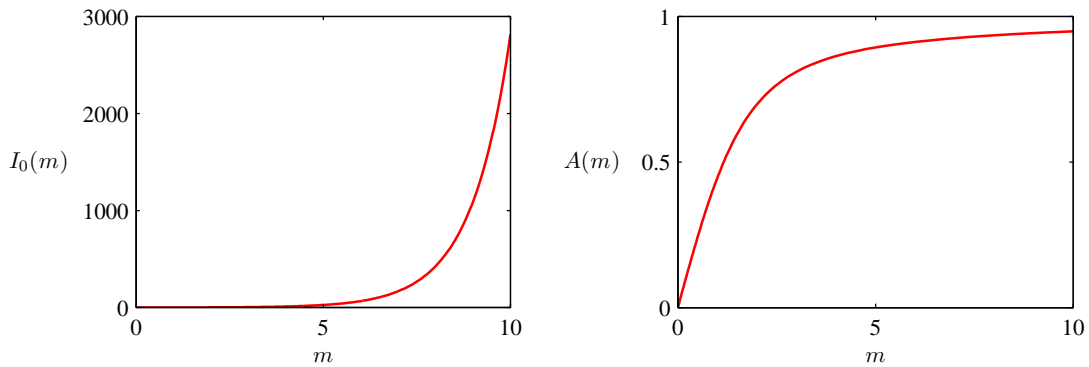


图 2.20 绘制由式 (2.180) 定义的贝塞尔函数 $I_0(m)$ ，以及由式 (2.186) 定义的函数 $A(m)$ 。

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

$$A(m) = \frac{I_1(m)}{I_0(m)} \quad (2.186)$$

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

$$A(m_{\text{ML}}) = \left(\frac{1}{N} \sum_{n=1}^N \cos \theta_n \right) \cos \theta_0^{\text{ML}} - \left(\frac{1}{N} \sum_{n=1}^N \sin \theta_n \right) \sin \theta_0^{\text{ML}} \quad (2.187)$$

(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 $(0, 2\pi)$ 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”（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（mixture distributions）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

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (2.188)$$

这被称为高斯混合（mixture of Gaussians）。每一个高斯密度 $\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ 都被称为混

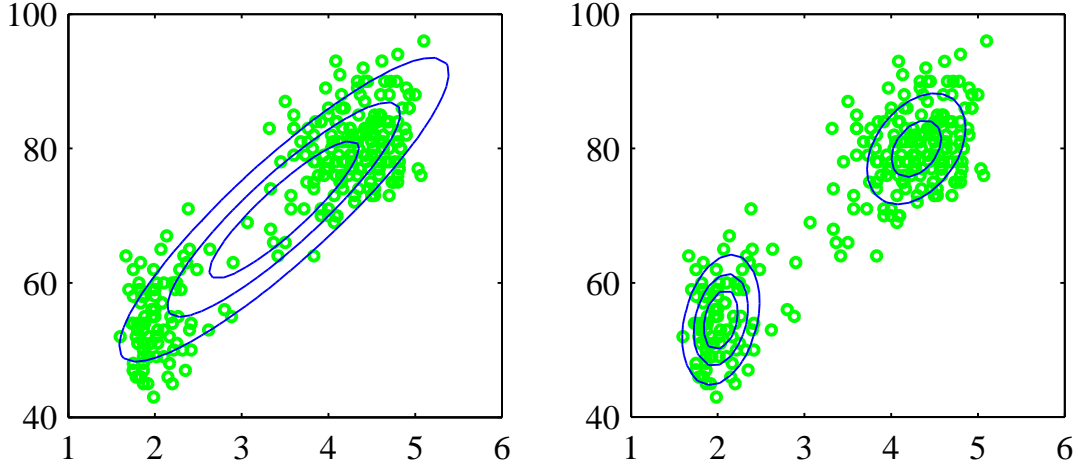


图 2.21 对“老忠实”数据的绘图，其中蓝色曲线表示等概率密度轮廓。左图为使用最大似然拟合得到的单个高斯分布。注意该分布无法捕捉数据中的两个聚集区域，并且将大部分概率质量放在这两个聚集之间的中间区域，而该区域的数据相对稀疏。右图为由两个高斯的线性组合构成的分布，利用第 9 章讨论的方法通过最大似然进行拟合，能够更好地刻画数据结构。

合的一个分量（component），它拥有自己的均值 μ_k 和协方差 Σ_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。

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（mixing coefficients）。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

$$\sum_{k=1}^K \pi_k = 1 \quad (2.189)$$

此外， $p(\mathbf{x}) \geq 0$ 的要求，连同 $\mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k) \geq 0$ 意味着 $\pi_k \geq 0$ （对所有 k 成立）。结合条件 (2.189)，我们得到

$$0 \leq \pi_k \leq 1 \quad (2.190)$$

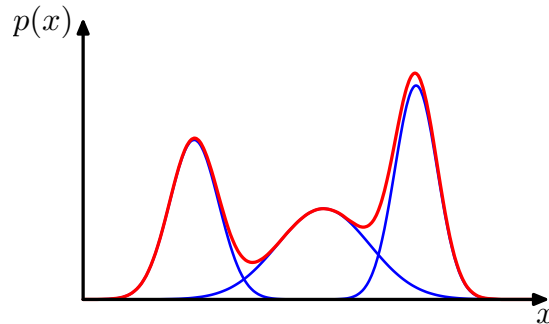


图 2.22 一维高斯混合分布示例，其中三个经过系数缩放的高斯分布以蓝色显示，它们的和以红色显示。

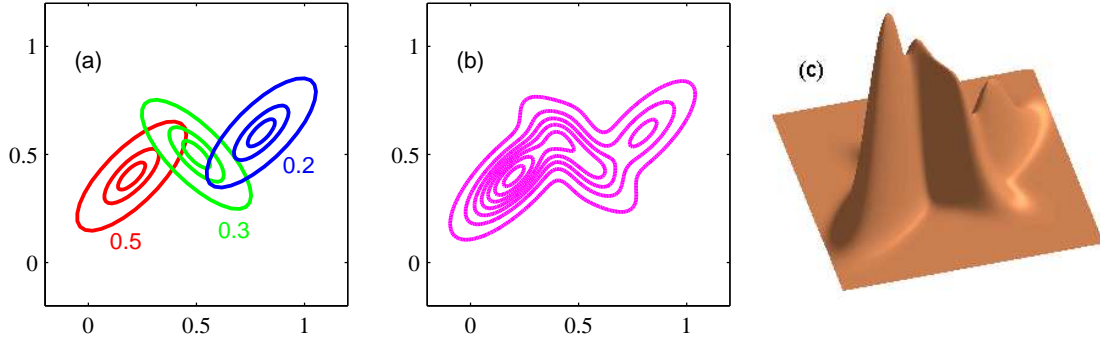


图 2.23 一维高斯混合分布示例，其中三个经过系数缩放的高斯分布以蓝色显示，它们的和以红色显示。

因此我们看到，混合系数满足作为概率的要求。

根据求和法则与乘积法则，边际密度为

$$p(\mathbf{x}) = \sum_{k=1}^K p(k) p(\mathbf{x} | k) \quad (2.191)$$

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 $\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = p(\mathbf{x} | k)$ 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 $p(k | \mathbf{x})$ （也称为责任（responsibilities））扮演着重要角色。根据贝叶斯定理，它们为

$$\begin{aligned} \gamma_k(\mathbf{x}) &\equiv p(k | \mathbf{x}) \\ &= \frac{p(k) p(\mathbf{x} | k)}{\sum_l p(l) p(\mathbf{x} | l)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_l \pi_l \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \end{aligned} \quad (2.192)$$

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 $\boldsymbol{\pi} \equiv \{\pi_1, \dots, \pi_K\}$ 、 $\boldsymbol{\mu} \equiv \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K\}$ 以及 $\boldsymbol{\Sigma} \equiv \{\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K\}$ 。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

$$\ln p(\mathbf{X} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \quad (2.193)$$

其中 $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ 。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化（expectation maximization）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：

$$p(\mathbf{x} | \boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \} \quad (2.194)$$

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

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \} d\mathbf{x} = 1 \quad (2.195)$$

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

$$p(x | \mu) = \text{Bern}(x | \mu) = \mu^x (1 - \mu)^{1-x} \quad (2.196)$$

将右端表示为对数的指数形式，得到

$$\begin{aligned} p(x | \mu) &= \exp \{ x \ln \mu + (1 - x) \ln(1 - \mu) \} \\ &= (1 - \mu) \exp \left\{ \ln \left(\frac{\mu}{1 - \mu} \right) x \right\} \end{aligned} \quad (2.197)$$

与 (2.194) 对比，可识别出

$$\eta = \ln \left(\frac{\mu}{1 - \mu} \right) \quad (2.198)$$

我们可以解出 μ 得到 $\mu = \sigma(\eta)$ ，其中

$$\sigma(\eta) = \frac{1}{1 + \exp(-\eta)} \quad (2.199)$$

被称为逻辑斯蒂克西格玛函数（logistic sigmoid function）。于是我们可以用标准形式 (2.194) 将伯努利分布写为

$$p(x | \eta) = \sigma(-\eta) \exp(\eta x) \quad (2.200)$$

这里利用了 $1 - \sigma(\eta) = \sigma(-\eta)$ ，这由 (2.199) 易证。与 (2.194) 对比可得

$$u(x) = x \quad (2.201)$$

$$h(x) = 1 \quad (2.202)$$

$$g(\eta) = \sigma(-\eta) \quad (2.203)$$

接下来考虑多项分布，对于单个观测 \mathbf{x} 它具有形式

$$p(\mathbf{x} | \boldsymbol{\mu}) = \prod_{k=1}^M \mu_k^{x_k} = \exp \left\{ \sum_{k=1}^M x_k \ln \mu_k \right\} \quad (2.204)$$

其中 $\mathbf{x} = (x_1, \dots, x_M)^T$ 。同样，我们可以将它写成标准形式 (2.194)：

$$p(\mathbf{x} | \boldsymbol{\eta}) = \exp(\boldsymbol{\eta}^T \mathbf{x}) \quad (2.205)$$

其中 $\eta_k = \ln \mu_k$ ，并定义 $\boldsymbol{\eta} = (\eta_1, \dots, \eta_M)^T$ 。再与 (2.194) 对比得到

$$\mathbf{u}(\mathbf{x}) = \mathbf{x} \quad (2.206)$$

$$h(\mathbf{x}) = 1 \quad (2.207)$$

$$g(\boldsymbol{\eta}) = 1 \quad (2.208)$$

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

$$\sum_{k=1}^M \mu_k = 1 \quad (2.209)$$

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 $\{\mu_k\}$ ($k = 1, \dots, M-1$) 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

$$0 \leq \mu_k \leq 1, \quad \sum_{k=1}^{M-1} \mu_k \leq 1 \quad (2.210)$$

利用约束 (2.209)，多项分布在此表示下变为

$$\begin{aligned} & \exp \left\{ \sum_{k=1}^M x_k \ln \mu_k \right\} \\ &= \exp \left\{ \sum_{k=1}^{M-1} x_k \ln \mu_k + \left(1 - \sum_{k=1}^{M-1} x_k \right) \ln \left(1 - \sum_{k=1}^{M-1} \mu_k \right) \right\} \\ &= \exp \left\{ \sum_{k=1}^{M-1} x_k \ln \left(\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j} \right) + \ln \left(1 - \sum_{k=1}^{M-1} \mu_k \right) \right\} \end{aligned} \quad (2.211)$$

现在令

$$\ln \left(\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j} \right) = \eta_k \quad (2.212)$$

对 k 求和后整理并回代，得到

$$\mu_k = \frac{\exp(\eta_k)}{1 + \sum_j \exp(\eta_j)} \quad (2.213)$$

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

$$p(\mathbf{x} | \boldsymbol{\eta}) = \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k) \right)^{-1} \exp(\boldsymbol{\eta}^T \mathbf{x}) \quad (2.214)$$

这是指数家族的标准形式，参数向量为 $\boldsymbol{\eta} = (\eta_1, \dots, \eta_{M-1})^T$ ，其中

$$\mathbf{u}(\mathbf{x}) = \mathbf{x} \quad (2.215)$$

$$h(\mathbf{x}) = 1 \quad (2.216)$$

$$g(\boldsymbol{\eta}) = \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k) \right)^{-1} \quad (2.217)$$

最后，我们考虑高斯分布。对于单变量高斯分布，有

$$p(x | \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\} \quad (2.218)$$

$$= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} x^2 + \frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} \mu^2 \right\} \quad (2.219)$$

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

$$\boldsymbol{\eta} = \begin{pmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{pmatrix} \quad (2.220)$$

$$\mathbf{u}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \quad (2.221)$$

$$h(x) = (2\pi)^{-1/2} \quad (2.222)$$

$$g(\boldsymbol{\eta}) = (-2\eta_2)^{1/2} \exp \left(\frac{\eta_1^2}{4\eta_2} \right) \quad (2.223)$$

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

$$\begin{aligned} & \nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \} d\mathbf{x} \\ & + g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \} \mathbf{u}(\mathbf{x}) d\mathbf{x} = 0 \end{aligned} \quad (2.224)$$

整理并再次利用 (2.195)，得到

$$-\frac{1}{g(\boldsymbol{\eta})} \nabla g(\boldsymbol{\eta}) = \int h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \} \mathbf{u}(\mathbf{x}) d\mathbf{x} = \mathbb{E}[\mathbf{u}(\mathbf{x})] \quad (2.225)$$

这里用到了 (2.194)。因此得到结果

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})] \quad (2.226)$$

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 $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ ，其似然函数为

$$p(\mathbf{X} | \boldsymbol{\eta}) = \left(\prod_{n=1}^N h(\mathbf{x}_n) \right) g(\boldsymbol{\eta})^N \exp \left\{ \boldsymbol{\eta}^T \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n) \right\} \quad (2.227)$$

将 $\ln p(\mathbf{X} | \boldsymbol{\eta})$ 关于 $\boldsymbol{\eta}$ 的梯度设为零，得到最大似然估计 $\boldsymbol{\eta}_{\text{ML}}$ 必须满足的条件

$$-\nabla \ln g(\boldsymbol{\eta}_{\text{ML}}) = \frac{1}{N} \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n) \quad (2.228)$$

原则上可通过求解得到 η_{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， $\mathbf{u}(x) = (x, x^2)^T$ ，因此需要保留 $\{x_n\}$ 的和以及 $\{x_n^2\}$ 的和。

如果考虑极限 $N \rightarrow \infty$ ，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 η_{ML} 将等于真实值 η 。

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 $p(\mathbf{x} | \eta)$ ，我们可以寻找一个与似然函数共轭的先验 $p(\eta)$ ，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：

$$p(\eta | \chi, \nu) = f(\chi, \nu) g(\eta)^\nu \exp \{ \nu \eta^T \chi \} \quad (2.229)$$

其中 $f(\chi, \nu)$ 为归一化系数， $g(\eta)$ 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

$$p(\eta | \mathbf{X}, \chi, \nu) \propto g(\eta)^{\nu+N} \exp \left\{ \eta^T \left(\sum_{n=1}^N \mathbf{u}(\mathbf{x}_n) + \nu \chi \right) \right\} \quad (2.230)$$

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 χ 。

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 $p(x | \lambda)$ 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 $\hat{h}(\eta) = h(\eta^2)$ 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

$$p_\eta(\eta) = p_\lambda(\lambda) \left| \frac{d\lambda}{d\eta} \right| = p_\lambda(\eta^2) \cdot 2\eta \propto \eta \quad (2.231)$$

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 $p(x | \lambda)$ 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

$$p(x | \mu) = f(x - \mu) \quad (2.232)$$

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 $\hat{x} = x + c$, 则

$$p(\hat{x} | \hat{\mu}) = f(\hat{x} - \hat{\mu}) \quad (2.233)$$

其中定义了 $\hat{\mu} = \mu + c$ 。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 $A \leq \mu \leq B$ 赋予与平移后的区间 $A - c \leq \mu \leq B - c$ 相同的概率质量。这意味着

$$\int_A^B p(\mu) d\mu = \int_{A-c}^{B-c} p(\mu) d\mu = \int_A^B p(\mu - c) d\mu \quad (2.234)$$

由于这对任意 A 和 B 都必须成立, 故有

$$p(\mu - c) = p(\mu) \quad (2.235)$$

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 $p(\mu | \mu_0, \sigma_0^2) = \mathcal{N}(\mu | \mu_0, \sigma_0^2)$, 取极限 $\sigma_0^2 \rightarrow \infty$ 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

$$p(x | \sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right) \quad (2.236)$$

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 $\hat{x} = cx$, 则

$$p(\hat{x} | \hat{\sigma}) = \frac{1}{\hat{\sigma}} f\left(\frac{\hat{x}}{\hat{\sigma}}\right) \quad (2.237)$$

其中定义了 $\hat{\sigma} = c\sigma$ 。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 $A \leq \sigma \leq B$ 与缩放后的区间 $A/c \leq \sigma \leq B/c$, 先验应对这两个区间赋予相等的概率质量。于是

$$\int_A^B p(\sigma) d\sigma = \int_{A/c}^{B/c} p(\sigma) d\sigma = \int_A^B p\left(\frac{\sigma}{c}\right) \frac{1}{c} d\sigma \quad (2.238)$$

由于这对任意 A 和 B 都必须成立, 故有

$$p(\sigma) = p\left(\frac{\sigma}{c}\right) \frac{1}{c} \quad (2.239)$$

从而 $p(\sigma) \propto 1/\sigma$ 。这又是一个非正常先验, 因为在 $0 \leq \sigma \leq \infty$ 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 $1 \leq \sigma \leq 10$ 、 $10 \leq \sigma \leq 100$ 以及 $100 \leq \sigma \leq 1000$ 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

$$\mathcal{N}(x | \mu, \sigma^2) \propto \sigma^{-1} \exp\{-(\tilde{x}/\sigma)^2\} \quad (2.240)$$

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 $\text{Gam}(\lambda | a_0, b_0)$, 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

$$p_i = \frac{n_i}{N\Delta_i} \quad (2.241)$$

易见 $\int p(x) dx = 1$ 。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。

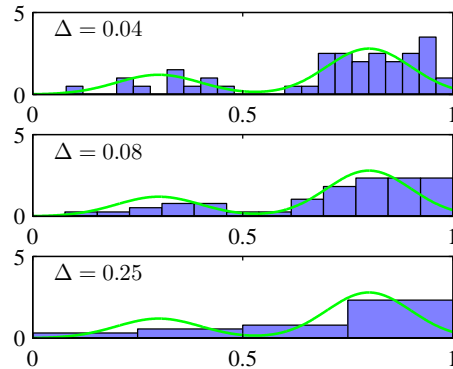


图 2.24 密度估计中直方图方法的示意图，其中 50 个数据点由绿色曲线所示的分布生成。基于式 (2.241) 并采用统一分箱宽度 Δ 的直方图密度估计在不同 Δ 取值下给出。

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

$$P = \int_{\mathcal{R}} p(\mathbf{x}) d\mathbf{x}. \quad (2.242)$$

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

$$\text{Bin}(K | N, P) = \frac{N!}{K!(N-K)!} P^K (1-P)^{N-K} \quad (2.243)$$

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 $\text{var}[K/N] = P(1-P)/N$ 。当 N 很大时，该分布将在均值附近急剧集中，因此

$$K \simeq NP \quad (2.244)$$

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

$$P \simeq p(\mathbf{x})V \quad (2.245)$$

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

$$p(\mathbf{x}) = \frac{K}{NV} \quad (2.246)$$

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 $N \rightarrow \infty$ 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

$$k(\mathbf{u}) = \begin{cases} 1, & |u_i| \leq 1/2, \quad i = 1, \dots, D \\ 0, & \text{otherwise} \end{cases} \quad (2.247)$$

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 $k((\mathbf{x} - \mathbf{x}_n)/h)$ 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

$$K = \sum_{n=1}^N k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right). \quad (2.248)$$

将此表达式代入 (2.246)，即可得到点 \mathbf{x} 处的估计密度

$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N \frac{1}{h^D} k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right) \quad (2.249)$$

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

$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N \frac{1}{(2\pi h^2)^{1/2}} \exp \left\{ -\frac{\|\mathbf{x} - \mathbf{x}_n\|^2}{2h^2} \right\} \quad (2.250)$$

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。

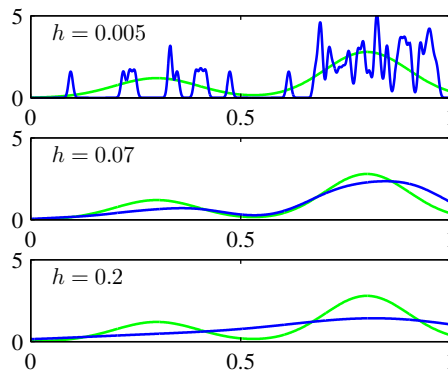


图 2.25 将核密度模型 (2.250) 应用于与图 2.24 中直方图方法相同的数据集。可以看到， h 起平滑参数作用：若 h 取值过小（上图），得到的密度模型噪声很大；若 h 取值过大（下图），则数据生成的真实分布（绿色曲线所示）本来的双峰结构被抹平。某个中间取值的 h （中图）则给出了最佳密度模型。

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：

$$k(\mathbf{u}) \geq 0, \quad (2.251)$$

$$\int k(\mathbf{u}) d\mathbf{u} = 1 \quad (2.252)$$

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。

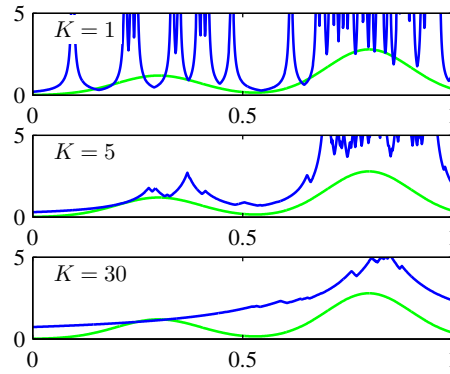


图 2.26 使用与图 2.25 和图 2.24 相同的数据集进行 K 近邻密度估计的示意图。可以看到，参数 K 决定平滑程度： K 较小时（上图），密度模型噪声很大； K 较大时（下图），数据生成的真实分布（绿色曲线所示）的双峰结构被过度平滑。

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 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 C_k 的点有 K_k 个。那么 (2.246) 给出了各个类别的密度估计

$$p(\mathbf{x} | C_k) = \frac{K_k}{N_k V} \quad (2.253)$$

同样，无条件密度为

$$p(\mathbf{x}) = \frac{K}{NV} \quad (2.254)$$

而类别先验概率为

$$p(C_k) = \frac{N_k}{N} \quad (2.255)$$

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

$$p(C_k | \mathbf{x}) = \frac{p(\mathbf{x} | C_k) p(C_k)}{p(\mathbf{x})} = \frac{K_k}{K} \quad (2.256)$$

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。

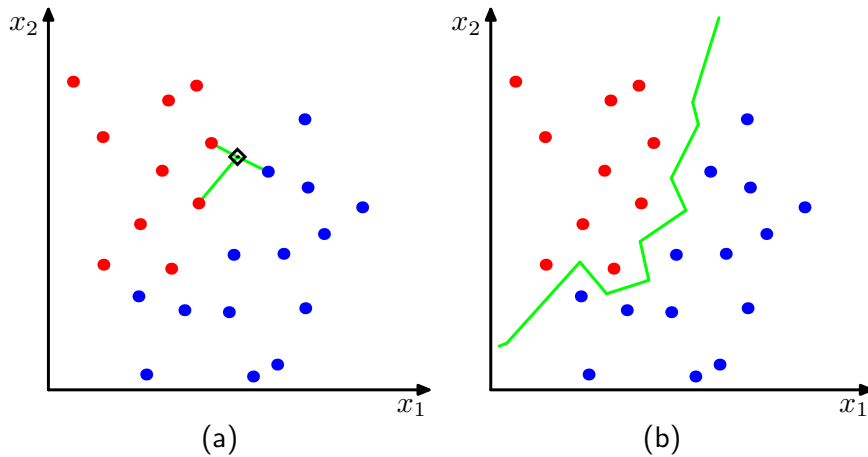


图 2.27 (a) 在 K 近邻分类器中，新数据点（黑色菱形所示）的类别由其最近的 K 个训练样本中占多数的类别决定，此处 $K=3$ 。(b) 在最近邻 ($K=1$) 分类方法中，得到的决策边界由一系列超平面组成，这些超平面是来自不同类别的成对样本之间的垂直平分面。

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 $N \rightarrow \infty$ 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。

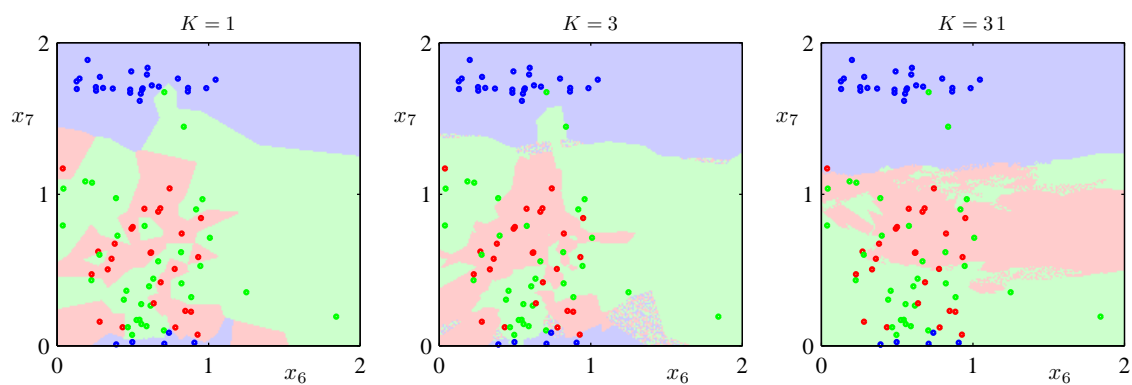


图 2.28 从油流数据集中抽取的 200 个数据点，其特征 x_6 与 x_7 的散点图如上所示。其中红色、绿色与蓝色点分别对应“层流 (laminar)”、“环状流 (annular)”和“均匀流 (homogeneous)”三类。图中还展示了 K 近邻算法在不同 K 取值下对输入空间的分类结果。