第 1 章绪论

1.1引言

寻找数据中的模式，是一个基本问题，此问题拥有悠久的历史。例如，16世纪丹麦天文学家第谷·布拉赫(Tycho Brahe，1546-1601)所进行的大量天文观测为开普勒发现行星运动的经验法则奠定了基础，这也为古典力学的发展提供了一个跳板。同样地，原子光谱规律的发现在20世纪早期量子物理的发展和验证阶段扮演了关键性角色。在模式识别领域，常常依据经验规则并利用计算机算法去发现数据中的规律，比如分类不同类别中的数据。

考虑识别手写数字的例子，如图1-1所示。每个数字相当于一个28×28像素的图像，因此可以将数字图像表示为由784个数字组成的矢量 x。我们的目标是建立一个机器，将这样一个向量 x 作为输入，并将产生的辨别数字0,…,9作为输出。由于手写风格的广泛变化性，让计算机去识别手写数字并不是一个简单的问题。可以使用基于笔画形状的手工处理规则或启发式方法区分数字，但是这种方法会导致规则扩散和规则不适用等情况，并总是给出不理想的识别结果。

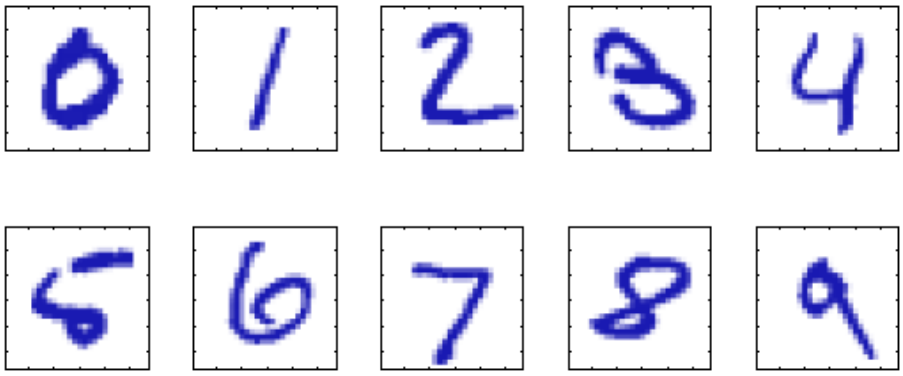


图1-1 美国邮政编码中的手写数字

较好的结果可以通过机器学习方法获得。利用一个大小为N个数字的集合$\{\mathrm{x}_1,\mathrm{x}_2,\cdots,\mathrm{x}_N\}$，称为**训练样本（training samples）**来自适应地调整模型的参数。这些在训练集中的数字的分类结果被认为是预先知道的，可以逐个地检查它们所属的类别并手工标定它们。我们可以使用目标向量t来表示分类，这种表示表达

了与相关数字的同一性。利用适当的向量来表示类别的方法将会在后面讨论。注意这里一个目标向量t对应一个数字图像x。

运行机器学习算法的结果可以看作是一个函数$y(x)$，此函数以一个新的数字图像x作输入，以一个输出向量y作输出。这些函数$y(x)$的精确形式通过**训练（training）阶段**决定，也称为**学习（learning）阶段**，这一阶段是在训练数据的基础之上进行学习或训练。一旦模型训练完成，它就可以用来辨别新的图像，这些新的图像称为**测试集（testing data）**。正确分类新图像（测试集合）的能力，称为**泛化（generation）能力**。在实际的应用中，训练数据可能仅仅由所有可能的输入向量中的一小部分组成，因此模型良好的泛化能力是模式识别的一个中心目标。

在大部分实际应用中，原始的输入变量通常经过**预处理（preprocessing）**转换到新的变量空间，以使得模式识别问题更加容易解决。例如，在手写数字识别的问题中，这些数字图像均被转化和放缩到一个固定大小的尺寸。由于所有数字的位置和尺寸都是相同的，这大大降低了每个数字类里的变异性，使得它们更容易被随后的模式识别算法区分。这一预处理阶段有时也称为特征提取。值得注意的是，与训练数据一样，新的测试数据必须使用相同步骤来进行预处理。

预处理也可以用来加速计算。具体来说，如果目标任务是在高分辨率视频流中进行人脸检测，计算机必须每秒钟处理大量的像素点，在一个复杂的模式识别算法中进行这些操作计算上几乎是不可能的。取而代之，目标任务是去寻找能够快速计算的有用特征，这些特征保留了有用、有辨识度的信息，能够区分人脸与非人脸区域，并将这些特征作为模式识别程序的输入。例如，一幅图像在一个矩形子区域上的灰度平均值能被高效地计算，可以证明此类特征在快速人脸检测中具有高效的性能。由于特征点的数目小于原始像素点的数量，这种预处理过程也可以视为是一种**降维（dimensionality reduction）**的过程。在预处理过程中必须小心，因为此过程经常丢弃信息，如果丢弃的信息对于问题很重要，解决方案的整体精度将受到影响。

在训练数据中，输入向量具有相应的目标向量的例子被称为**监督学习（supervised learning）**问题。如手写数字识别的例子。若问题的目的是将每个输入向量分配给有限数目的离散值，则此类问题被称为**分类（classification）**问

题。如果所需要的输出包含一个或多个连续变量，那么这个任务称为**回归（regression）**问题。一个回归问题的例子是预测化学过程的产出，这里输入包括反应物的浓度、温度和压力。

在另一类模式识别问题中，训练数据的输入向量是没有任何相应目标值的一组输入向量 x，这称为**无监督学习（unsupervised learning）**。无监督学习的目标可能是去发现在一组数据中相似的组分，被称为**聚类（clustering）**，或是确定在分布空间内部的数据分布形式，称为**密度估计（density estimation）**，或者是去建立数据的**可视化（visualization）**，目的是从高维空间降低到二维或三维空间。

最后，**强化学习（reinforcement learning）**技术指的是在给定情况下以最大化回报为目标，如何采取合适的行动的问题。与监督学习不同，这里的学习算法并不是给定例子的最优输出，而是通过不断的试验去发现它们。有代表性的是，一系列的学习算法通过状态和行为与它所在的环境进行交互。在许多环境中，当前行为不仅影响直接反馈而且对于随后所有时间内的反馈都有影响。举个例子来说，通过使用适当的强化学习技术，一个神经网络能学习玩高标准的西洋双陆棋游戏。这里的网络连同掷骰子的结果一起作为输入，产生一个有力的移动作为输出。这样，神经网络就可以和自己的一个拷贝进行百万次的对战，完成训练过程。其中一个主要的挑战是西洋双陆棋游戏可能出现几十种移动步骤，只有等到比赛结束才有反馈（胜利或失败）。尽管一些移动会产生好的效果而另一些移动的效果一般，导致反馈的结果必须恰当地归结于所有产生这一结果的移动操作上。增强学习的一个普遍特点是在探测（**exploration**）和探索（**exploitation**）之间进行权衡。在探测阶段，系统尝试新的行为，来测试它们的有效性。在探索阶段，系统利用已知的产生好结果的行为。在两者之间过于偏向其中任何一个均会产生不好的结果。

增强学习依然是机器学习研究中一个活跃的领域。然而由于时间关系，本书仅能做简要介绍。

本书的章节书序如下：第一章，绪论；第二章，贝叶斯统计决策理论；第三章，概率密度函数估计；第四章，线性回归分析；第五章，线性判别函数；第六章，其他分类方法；第七章，无监督学习和聚类；第八章，神经网络；第九章，

核方法和支撑向量机；第十章，特征选择与提取。

1.2模式识别的基本概念

1.2.1 模式和模式识别

在人们的日常生活中，模式识别是普遍存在和经常进行的过程。首先来看一个简单的例子。考虑鱼的品种分类问题，假设有一个鱼类加工厂，希望能将传送带上的鱼的品种的分类过程自动进行。首先要拍摄若干**样本**（鲑鱼和鲈鱼）的图像，从图像上看，这两种鱼确实存在一些差异，比如长度、光泽度、宽度以及嘴的位置等等。

我们将以上差异性的要素称为**特征**。在得到一些鱼类样本的图片后，需要选择一些可用的特征进行分类。假设有人告诉我们“鲈鱼一般比鲑鱼长”，这就提供了一种可尝试的分类特征——长度。我们可以仅仅通过看一条鱼的长度l是否超过某个临界值l^*来判别鱼的种类。为确定恰当的l^*，必须先获得不同类别的鱼的若干样本（称为“设计样本”或“**训练样本**”），进行长度测量并检查结果。假设我们已经完成上述工作，并将长度直方图绘于图 1-2。

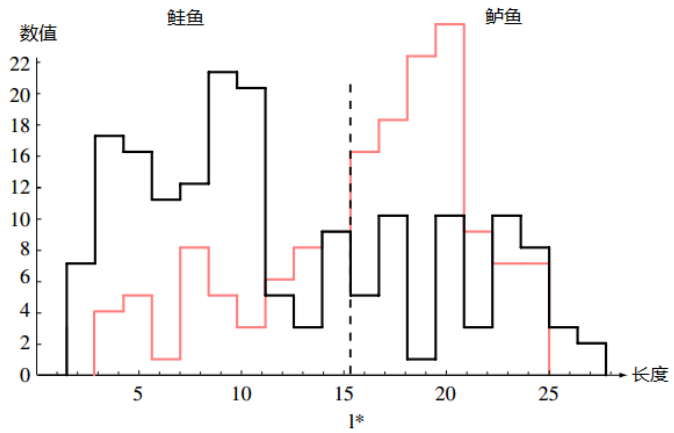


图 1-2 两种鱼的长度特征直方图

图 1-2 验证了在平均意义上，鲈鱼比鲑鱼要长的结论。然而，单一的特征判据是不足以分类的。也就是说，无论怎样确定临界值l^*，都无法仅凭长度就把两种鱼分开。

继续尝试其他的特征，我们发现，两种鱼的光泽度也存在很大差异。当小心地消除外界照明的影响后，最终可以获得如图 1-3 所示的光泽度直方图。这个结果比较令人满意，因为两种鱼的分离性更好。

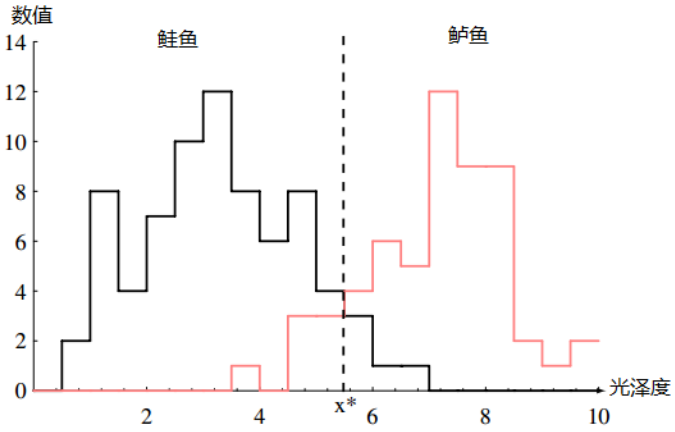


图 1-3 两种鱼的光泽度特征直方图（不存在单一的阈值能将两种鱼无歧义地分开）

此外，在寻求其他可用于分类的特征时，我们还发现，鲈鱼通常比鲑鱼要宽。这样就有了两个特征——光泽度和宽度。因此选择光泽度 x_1 和宽度 x_2 作为两个用于分类的特征，这两个特征组成一个二维特征向量，或二维特征空间中的一个点 $\mathbf{x}=\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$。

现在，我们要把特征空间分为两个区域，使得落在其中一个区域的数据点（鱼）被分类为鲈鱼，而落在另外一个区域的数据点被分类为鲑鱼。假定已经对样本特征向量作了测量，并绘制了散布图（如图 1-4 所示）。这个图显示出可以根据如下的准则来区分两种鱼：如果特征向量落在**判别边界**（decision boundary）的上方，则是鲈鱼，否则是鲑鱼。图中的斜线是分类判决的分界线。很明显，这里的总体误差要比图 1-3 中小，但仍存在一些错误。

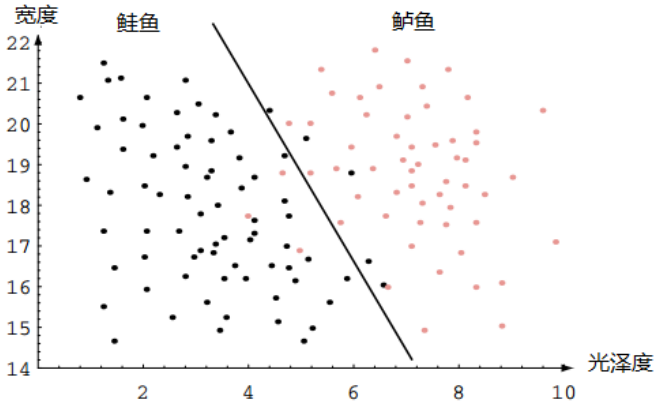


图 1-4 两种鱼的光泽度特征和宽度特征散布图

如果分类的判决模型非常复杂，分界面也十分复杂，不再是一条直线，所有的训练样本可以被完美地正确分类（如图 1-5 所示）。虽然如此，这样的结果依然不令人满意。这是因为设计分类器的中心目标是能够对新样本（训练过程中未知的某条鱼）做出正确的反应，这就是**推广能力**的概念。图 1-5 所示的过分复杂的分界面过分“调谐”到某些特定的训练样本上去了，而不是类别的共同特征，或者说是待分类的全部鲈鱼（或鲑鱼）的总体模型。

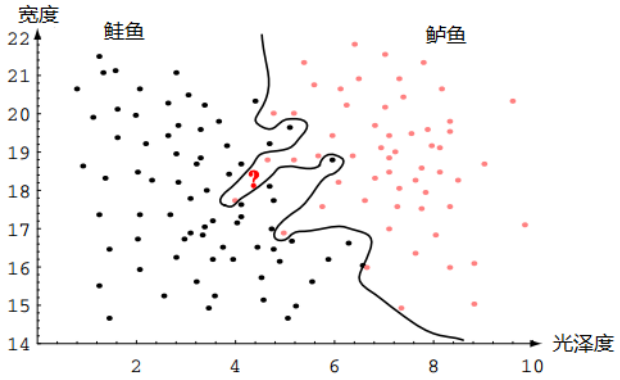


图 1-5 过分复杂的模型将导致复杂的判决曲线

所以，我们宁可去寻求某种“简化”分类器的方案。其背后的信念是，分类器所需的模型或判别边界并不需要像图 1-5 那样复杂。图 1-6 中所标出的判决曲线是对训练样本分类性能和分界面复杂度的一个最优折衷，因而对新模式的分类性能也很好。

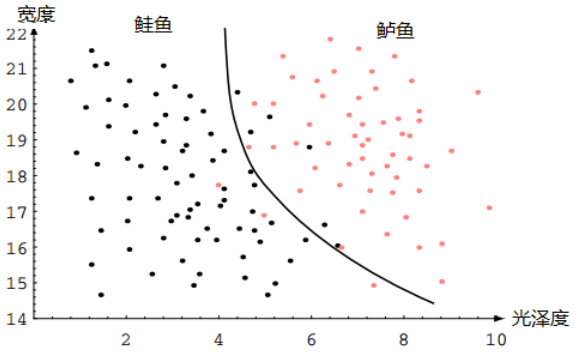


图 1-6 折衷的判决曲线

以上我们给出了对不同品种的鱼进行分类的问题。此外，人们可以根据飞机的飞行高度、速度、形状、结构判断飞机的种类；根据人的身高、面容和体型，判断是张三还是李四……诸如此类。上述判断过程，都是模式识别的具体过程。

因此，我们可以得出以下的结论。**模式**是取自世界有限部分的单一样本的被测量值的综合。**模式识别**就是试图确定一个样本的类型的属性，也就是把某一个样本归属于多个类型中的某一个类型。

模式识别的主要研究内容是通过机器的自动识别。这就需要把人的知识和经验交给机器，为机器制定一些规则和方法，使之具有综合分析、自动分类的判断能力，以便机器能够完成自动识别的任务。

模式识别技术已广泛应用于：人工智能、计算机工程、机器人、神经生物学、医学、侦探学、高能物理、考古、地址勘探、宇航、武器等领域。

1.2.2 模式空间、特征空间和类型空间

一般来说，模式识别必须经历由模式空间经过特征空间到类型空间的过程。为了说明这些概念，首先解释一下“物理上可以察觉到的世界”。在客观世界里存在一些物体和事件，它们都可被适当的和足够多的函数来描述，也就是说它们在物理上是可以被测量的，他们的可测数据的集合就称为物理上觉察到的世界。显然，这些可测数据，或者说这个世界的维数是无限多的。

在物理上可以觉察到的世界中，所选择出的某些物体和事件称为**样本**。对样本进行观测得到观测数据，每个样本观测数据的综合都构成模式，所有样本的观测数据则构成**模式空间**。模式空间的维数与选择的样本和测量方法有关，也与特

定应用有关。维数很大，但是，是一个有限值。在模式空间里，每个模式样本都是一个点，点的位置由该模式在各维上的测量值确定。由物理上可觉察到的世界到模式空间所经历的过程称为**模式采集**。

模式空间的维数虽然是有限的，但还是非常多，其中一些并不反映样本的实质，机器在做出判断之前要对模式空间里的各坐标元素进行综合分析，以获取最能揭示样本属性的观测量作为主要特征，这些主要特征就构成**特征空间**。显然，特征空间的维数大大压缩了，远小于模式空间。特征空间中每个坐标都是样本的主要特征，简称特征。每个样本在特征空间里也是一个点，点的位置由该样本的各特征值来确定。从模式空间到特征空间所需要的综合分析，往往包含适当的变换和选择，这个过程称为**特征提取和特征选择**。

由某些知识和经验可以确定分类准则，叫做**判决规则**。根据适当的判决规则，可以把特征空间里的样本区分成不同的类型，从而把特征空间塑造成了**类型空间**。不同类型之间的分界面称为**决策面**。类型空间的维数与类型的数目相等，一般小于特征空间的维数。由特征空间到类型空间所需要的操作就是分类判决。

从物理上可以觉察到的世界，通过模式空间、特征空间到类型空间，经历了模式采集、特征提取/选择、以及分类决策等过程，这就是一个完整的模式识别过程。为完成上述过程，还需要对机器进行训练，使机器具有识别的能力。模式识别过程的图形表示如图 1-7 所示。

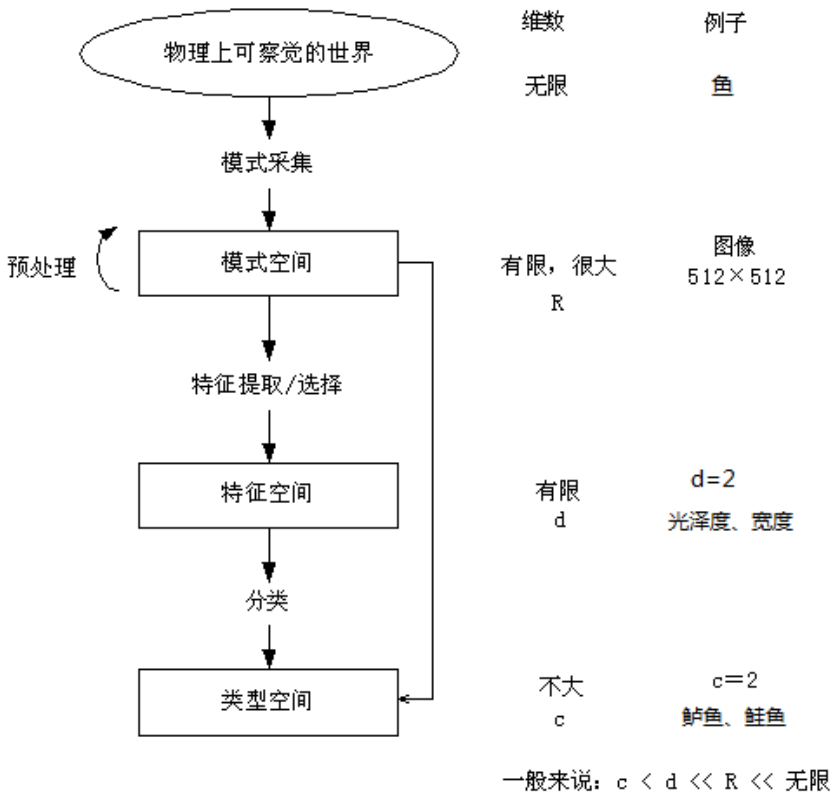


图 1-7 模式识别过程

1.2.3 预处理

模式空间里，针对具体的研究对象，往往需要进行适当的预处理。预处理的功能包括：清除或减少模式采集中的噪声及其它干扰，提高信噪比；消除或减少数据图像的模糊及几何失真，提高清晰度；转变模式的结构，以便后续处理（如非线性模式转为线性模式）。

预处理的方法包括滤波、变换、编码、标准化等。为了便于计算机处理，往往需要将模拟量转化为数字量，也就是进行 A/D 转换。在此过程中必须考虑两个问题，即采样间隔与量化等级。采样间隔（采样频率）表示单位时间内（秒），要求多少个采样值。量化级表示每个采样值要有多少个量化级，才能满足要求。

预处理的内容很多，这里不做详细的论述。有关内容可以在《数字信号处理》、《数字图像处理》等相关文献中找到更加详细的介绍。

1.2.4 特征提取/选择

本节对特征提取/选择的必要性和原则作简要介绍。

一般的情况，人们对客观世界里的具体物体或事件进行模式采集时，总是尽可能多的采集测量数据，造成样本在模式空间里的维数很大。模式维数很大首先带来的问题是处理的困难，处理时间很长，费用很高，有时甚至直接用于分类是不可能的，即所谓“维数灾难”。另外，在过多的数据坐标中，有些对刻画事物的本质贡献不大，甚至很小。因此，特征提取/选择十分必要。

特征提取/选择的目的目标，就是要压缩模式的维数，使之便于处理，减少消耗。特征提取往往以在分类中使用的某种判决规则为准则，所提取的特征使在某种准则下的分类错误最小。为此，必须考虑特征之间的统计关系，选用适当的正交变换，才能提取最有效的特征。在该准则下，选择对分类贡献较大的特征，删除贡献甚微的特征。

1.2.5 分类

分类的目标是：把特征空间划分成类型空间，把未知类别属性的样本确定为类型空间的某一个类型，以及在给定条件下，可以否定样本属于某种类型。分类的难易程度取决于两个因素，其一是来自同一个类别的不同个体之间的特征值的波动。其二是属于不同类别的样本的特征值之间的差异。

实际分类过程中，对于预先给定的条件，分类中出现错误是不可避免的。因此，分类过程只能以某种错误率来完成。显然，错误率越小越好。但是，分类错误率又受很多条件的制约：分类方法、分类器设计、选用的样本及提取的特征等。因此，分类错误率不能任意小。此外，分类错误率的分析、计算也很困难，只有在较简单的情况下才能有解析的解。分类错误率是分类过程中的重要问题。

1.3模式识别系统

一个模式识别系统应该完成模式采集、特征提取/选择、分类等功能。系统方框图 1-8 所示：

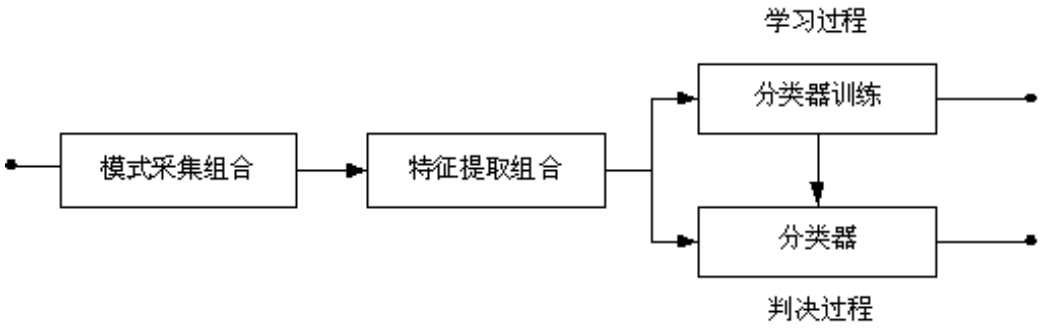


图 1-8 模式识别系统框图

1. 模式采集组合

模式采集组合完成模式的采集。根据处理对象的不同，可以选用不同的传感器、测量装置、图像录取输入装置等。采集之后，还要进行滤波、消除模糊、减少噪声、纠正几何失真等预处理操作。

2. 特征提取组合

特征提取组合实现由模式空间到特征空间的转变，有效压缩维数。一般来说，特征提取组合应该是在一定分类准则下的最佳或次佳变换器，或是实现某特征选择算法的装置。

3. 分类器

分类器要实现对未知类别属性样本的分类判决。为了设计分类器，首先要确定对分类错误率的要求，选用适当的判决规则。但是为了使分类器能有效地进行分类判决，还必须对它进行训练。也就是，分类器首先要进行学习。

4. 分类器的训练

分类器的训练/学习也是模式识别中的一个重要概念。由于我们研究分类器的自动识别，那么对分类器进行训练，使它具有自动识别的能力就尤为重要。大家都知道，小孩认字是一个反复学习的过程，那么机器要掌握某种判决规则，学习过程必不可少。前面讲过的医生诊病的例子，如果要让机器代替医生来给患者诊断，就必须把医生的知识和经验教给机器，并且输入一些病例，对机器进行训练。这种训练的过程就是机器学习的过程。这个过程往往需要反复多次，不断纠正错误，最后才能使机器自动诊断的错误率不超过给定的值。

经过特征提取/选择进入学习过程的样本常常被称为训练样本，其属性预先知道或者不知道。分类判决规则常常是样本各特征的函数，训练过程就是要确定函数的所有权因子。这个过程是一个输入、修正、再输入、再修正，不断反复的

过程，直到分类错误率不大于给定值为止。分类器完成训练之后，根据已经确定的判决规则，对未知类别属性的样本进行分类。此时，分类器就具有自动识别的能力。

通过前面的介绍，我们已经知道，模式识别就是要将模式进行正确分类。分类器训练/学习目的是确定判决规则，使之具有自动分类识别能力。在统计模式识别中，特征空间中的“类条件概率密度函数”是各种分类方法的基础，此时，分类器训练/学习就是最终完全确定类条件概率密度函数（类概率密度）。类概率密度的估计有两种方法，即参数估计法和非参数估计法。

1) 参数估计法

已知类概率密度函数或能从样本估计出类概率密度函数形式，但其中有未知参数，训练就是得到未知参数值。如：已知正态分布，但均值、协方差未知，通过训练求得这些值，进一步得到概率密度函数。参数估计主要有两种方法：Bayes估计法和最大似然估计法。

2) 非参数估计

假若预先不知道类概率密度函数的形式，就不能使用参数估计的方法，而只能借助于非参数估计法。非参数估计的方法也很多，并且随着研究工作的推进在不断发展，目前常用的方法：Parzen 窗法、Kn－近邻法、正交函数逼近法等。

1.4机器学习的主要方法

机器学习是近 20 多年兴起的一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。机器学习理论主要是设计和分析一些让计算机可以“自动学习”的算法。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。因为学习算法中涉及了大量的统计学理论，机器学习与统计推断学联系尤为密切，也被称为统计学习理论。算法设计方面，机器学习理论关注可以实现的，行之有效的学习算法。很多推论问题属于无程序可循难度，所以部分的机器学习研究是开发容易处理的近似算法。机器学习有下面几种定义： “机器学习是一门人工智能的科学，该领域的主要研究对象是人工智能，特别是如何在经验学习中改善具体算法的性能”；“机器学习是对能通过经验自动改进的计算机算法的研究”；“机器学习是用数据或以往的经

验，以此优化计算机程序的性能标准”。一种经常引用的英文定义是：A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

机器学习已经有了十分广泛的应用，例如：数据挖掘、计算机视觉、自然语言处理、生物特征识别、搜索引擎、医学诊断、检测信用卡欺诈、证券市场分析、DNA 序列测序、语音和手写识别、战略游戏和机器人运用。机器学习可以分成下面几种类别：监督学习、无监督学习、半监督学习、集成学习和增强学习。

1.4.1 监督学习

监督学习从给定的训练数据集中学习出一个函数，当新的数据到来时，可以根据这个函数预测结果。监督学习的训练集要求包括输入和输出，也可以说是特征和目标。训练集中的目标是由人标注的。常见的监督学习算法包括回归分析和统计分类。

监督式学习（Supervised learning）可以由训练资料中学到或建立一个模式（learning model），并依此模式推测新的实例。训练资料是由输入物件（通常是向量）和预期输出所组成。函数的输出可以是一个连续的值（称为回归分析），或是预测一个分类标签（称作分类）。

一个监督式学习者的任务在观察完一些训练范例（输入和预期输出）后，去预测这个函数对任何可能出现的输入的值的输出。要达到此目的，学习者必须以"合理"（见归纳偏向）的方式从现有的资料中一般化到未观察到的情况。在人类和动物感知中，则通常被称为概念学习（concept learning）。

监督式学习有两种形态的模型。最一般的，监督式学习产生一个全域模型，会将输入物件对应到预期输出。而另一种，则是一个区域模型。（如案例推论及最近邻居法）。为了解决一个给定的监督式学习的问题（手写辨识），必须考虑以下步骤：

1、决定训练资料的范例的形态。首先，工程师应决定要使用哪种资料为范例。譬如，可能是一个手写字符，或一整个手写的词汇，或一行手写文字。

2、搜集训练资料。这些资料须要具有真实世界的特征。所以，可以由人类

专家或（机器或传感器的）测量中得到输入物件和其相对应输出。

3、决定学习函数的输入特征的表示法。学习函数的准确度与输入的物件如何表示是有很大的关联度。传统上，输入的物件会被转成一个特征向量，包含了许多关于描述物件的特征。因为维数灾难的关系，特征的个数不宜太多，但也要足够大，才能准确的预测输出。

4、决定要学习的函数和其对应的学习算法所使用的数据结构。譬如，工程师可能选择人工神经网络和决策树。

5、完成设计。工程师接着在搜集到的资料上运行学习算法。可以借由将资料跑在资料的子集（称为验证集）或交叉验证（cross-validation）上来调整学习算法的参数。参数调整后，算法可以运行在不同于训练集的测试集上。

另外对于监督式学习所使用的词汇则是分类。现有的各式的分类器，各自都有强项或弱项。分类器的表现很大程度上地跟要被分类的资料特性有关。并没有某一单一分类器可以在所有给定的问题上都表现最好，这被称为‘天下没有白吃的午餐’理论。各式的经验法则被用来比较分类器的表现及寻找会决定分类器表现的资料特性。决定适合某一问题的分类器仍旧是一项艺术，而非科学。

目前最广泛被使用的分类器有人工神经网络、支持向量机、最近邻居法、高斯混合模型、朴素贝叶斯方法、决策树和径向基函数分类。

1.4.2 无监督学习

无监督学习与监督学习相比，训练集没有人为标注的结果。常见的无监督学习算法有聚类。比如 k-means 算法。

无监督式学习对原始资料进行分类，以便了解资料内部结构，有 clustering、density estimation、visualization 三种形式。无监督式学习在学习时并不知道其分类结果是否正确，亦不知道何种学习是正确的。其特点是仅提供输入样本，而它会自动从这些样本中找出其潜在类别规则。当学习完毕并经测试后，也可以将之应用到新的案例上。

1.4.3 半监督学习

半监督学习介于监督学习与无监督学习之间。

在机器学习领域中，传统的学习方法有两种：监督学习和无监督学习。半监督学习（Semi-supervised Learning）是模式识别和机器学习领域研究的重点问题，是监督学习与无监督学习相结合的一种学习方法。它主要考虑如何利用少量的标注样本和大量的未标注样本进行训练和分类的问题。半监督学习对于减少标注代价，提高学习机器性能具有非常重大的实际意义。

半监督学习的主要算法有五类：基于概率的算法；在现有监督算法基础上作修改的方法；直接依赖于聚类假设的方法；基于多试图的方法；基于图的方法。

1.4.4 集成学习

集成学习（Ensemble Learning）的思路是在对新的实例进行分类时，把若干个单个分类器集成起来，通过对多个分类器的分类结果进行某种组合来决定最终的分类，以取得比单个分类器更好的性能。如果把单个分类器比作一个决策者，那么集成学习的方法就相当于多个决策者共同进行一项决策。图 1-9 表示了利用神经网络进行集成学习的基本思想。图中的集成分类器包括了 N 个单一的人工神经网络分类器，对于同样的输入，N 个人工神经网络分别给出各自的输出 $(O_1, O_2, \cdots, O_n)$，然后这些输出通过整合以后得到集成分类器整体的输出作为最终分类结果。

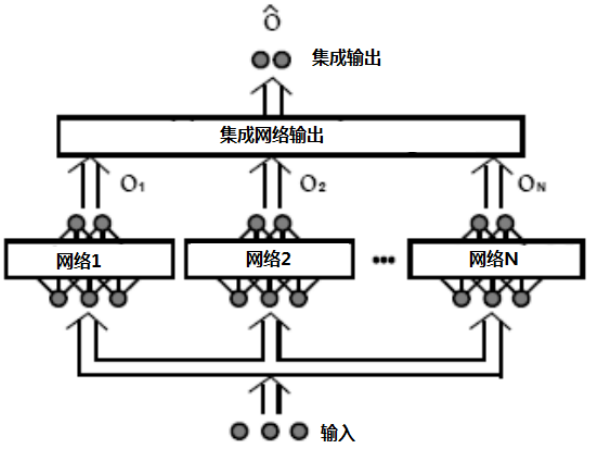


图 1-9 人工神经网络集成示意图

Thomas G. Dietterich 指出：集成学习的有效性可以归为统计上、计算上和表示上三个方面的原因。

a) 统计上的原因：对于一般的学习任务，往往要搜索的假设空间十分巨大，但是能够用于训练分类器的训练集中实例个数不足够用来精确地学习到目标假设，学习的结果便可能是一系列满足训练集的假设，而学习算法之能够选择这些假设的其中之一作为学习到的分类器进行输出。然而通过机器学习的过拟合问题我们看到，能够满足训练集的假设不一定在实际应用中有同样好的表现，这样学习算法选择哪个假设进行输出的时候就面临着一定的风险，把多个假设集成起来能够降低这种风险（这可以理解为通过集成使得各个假设和目标假设之间的误差得到一定程度的抵消）。

b) 计算上的原因：已经证明了在人工神经网络学习和决策树学习中，学习到最好的人工神经网络或者是决策树是一个 NP-hard 问题，其他的分类器模型也面临着类似的计算复杂度的问题。这使得我们只能用某些启发式的方法来降低寻找目标假设的复杂度，但这样的结果是找到的假设不一定是最优的。通过把多个假设集成起来能够使得最终的结果更加接近实际的目标函数值。

c) 表示上的原因：由于假设空间是人为规定的，在大多数机器学习的应用场合中实际目标假设并不在假设空间之中，如果假设空间在某种集成运算下不封闭，那么我们通过把假设空间中的一系列假设集成起来就有可能表示出不在假设空间中的目标假设。

1.4.5 增强学习

增强学习通过观察来学习做成如何的动作。每个动作都会对环境有所影响，学习对象根据观察到的周围环境的反馈来做出判断。

增强学习（Q-learning）要解决的是这样的问题：一个能感知环境的自治 agent，怎样通过学习选择能达到其目标的最优动作。这个很具有普遍性的问题应用于学习控制移动机器人，在工厂中学习最优操作工序以及学习棋类对弈等。当 agent 在其环境中做出每个动作时，施教者会提供奖励或惩罚信息，以表示结果状态的正确与否。例如，在训练 agent 进行棋类对弈时，施教者可在游戏胜利时给出正回报，而在游戏失败时给出负回报，其他时候为零回报。agent 的任务就是从这个非直接的，有延迟的回报中学习，以便后续的动作产生最大的累积效应。

1.5随机变量及分布

概率论是模式识别与机器学习的基础。故本小节简要回顾一下概率论的基本概念，严格的定义及推导请参考相关书籍。

一般地，如果 A 为某个随机事件，则一定可以通过如下示性函数使它与数值发生联系：

$$X=\begin{cases}1, & \text{如果A发生}\\ 0, & \text{如果A不发生}\end{cases} \tag{1-1}$$

试验结果能用一个数 X 来表示，这个数 X 是随着试验的结果不同而变化的，也即它是样本点的一个函数，这种量以后称为**随机变量**。

从随机现象的可能出现的结果来看，随机变量至少有两种不同的类型。一种是试验结果 X 所可能取的值为有限个或至多可列个，我们能把其可能的结果一一列举出来，这种类型的随机变量称为**离散型随机变量**。

与离散型随机变量不同，一些随机现象所出现的试验结果 X 不止取可列个值，例如测量误差、分子运动速度、候车时的等待时间、降水量、风速、洪峰值等等，这时用来描述试验结果的随机变量还是样本点的函数，但是这随机变量能取某个区间 $[c,d]$ 或 $(-\infty,+\infty)$ 的一切值。此时，这种随机变量为**连续型随机变量**。

正如对随机事件一样，我们所关心的不仅是试验会出现什么结果，更重要的是要知道这些结果将以怎样的概率出现，也即对随机变量，我们不但要知道它取什么值，而且要知道它取这些数值的概率。

1.5.1 分布函数与参数

称

$$F(x)=P\{X(\omega)<x\},-\infty<x<+\infty \tag{1-2}$$

为随机变量 X 的**分布函数**。其中，ω 表示样本点。为书写方便，通常把“随机变量 $X(\omega)$ 服从分布函数 $F(x)$”简记为 $X\sim \mathrm{F(x)}$。

对于**离散型随机变量**，设 $\{x\}_i$ 为离散型随机变量 X 的所有可能取值，而 $P(\mathrm{x}_i)$

是 X 取 x_i 的概率。即

$$P(X=\mathrm{x}_i)=\mathrm{p}(\mathrm{x}_i),\mathrm{i}=1,2,3,\cdots \tag{1-3}$$

$\{\mathrm{p}(\mathrm{x}_i),\mathrm{i}=1,2,3\cdots\}$ 称为随机变量 X 的**概率分布**，它应该满足下面关系：

$$\begin{aligned} &p(\mathrm{x}_i)\geq 0, i=1,2,3\cdots \\ &\sum_{i=1}^{\infty} p(\mathrm{x}_i)=1 \end{aligned} \tag{1-4}$$

有了概率分布，可以通过式(1-5)可以求得分布函数

$$F(\mathrm{x})=\mathrm{P}\{X<\mathrm{x}\}=\sum_{x_k<x} p(\mathrm{x}_k) \tag{1-5}$$

对于**连续型随机变量**，这种随机变量 X 可取某个区间 $[\mathrm{c},\mathrm{d}]$ 或 $(-\infty,+\infty)$ 中的一切值，而且其分布函数 $F(\mathrm{x})$ 是绝对连续函数，即存在可积函数 $p(\mathrm{x})$，使

$$F(\mathrm{x})=\int_{-\infty}^{x} p(\mathrm{y})\,\mathrm{dy} \tag{1-6}$$

其中，$p(\mathrm{y})$ 称为 X 的**概率密度函数**。显然，$p(\mathrm{x})=\mathrm{F}'(\mathrm{x})$，且 $p(\mathrm{x})$ 满足:

$$\begin{cases} p(\mathrm{x})\geq 0 \\ \int_{-\infty}^{+\infty} p(\mathrm{x})\,\mathrm{dx}=1 \end{cases} \tag{1-7}$$

- 常见的离散型随机变量的分布有：

1. **伯努利分布：**

在一次试验中，事件 A 出现的概率为 p，不出现的概率为 $q=1-p$，若以 β 记事件 A 出现的次数，则 β 仅取 0,1 两个值，相应的概率分布如式(1-8)所示。

$$b_k=p\{\beta=\mathrm{k}\}=\mathrm{p}^k\,q^{i-k},k=0,1 \tag{1-8}$$

这个分布称为伯努利分布，也称两点分布。图 1-10 给出了 p=0.6 时伯努利分布的概率密度函数的图形：

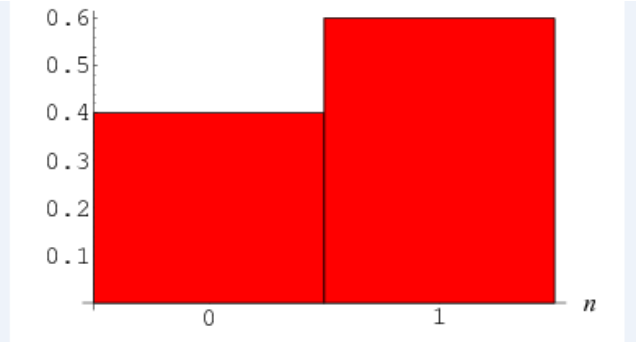


图 1-10 p=0.6 时的伯努利分布概率密度函数

2. 二项分布：

在 n 重伯努利试验中，若以 μ 记事件 A 出现的次数，则它是一个随机变量，μ 可能取的值为 0,1,2,…,n,其对应的概率由二项分布给出：

$$b(\mathrm{k};\mathrm{n},\mathrm{p})=\mathrm{p}\{\mu=\mathrm{k}\}=\binom{n}{k}p^{k}q^{n-k},k=0,1,2,...,\mathrm{n} \tag{1-9}$$

简记为 $\mu\sim B(\mathrm{n},\mathrm{p})$。图 1-11 给出了 n 分别取 7、8， $p=\frac{1}{3}$ 时二项分布的概率密度函数的图形：

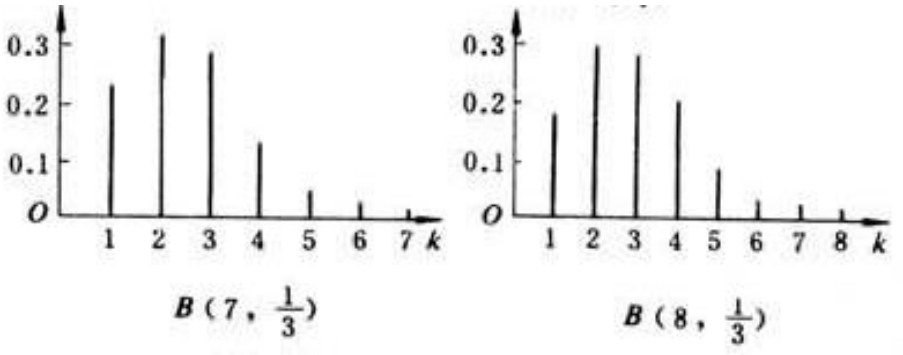


图 1-11 两组二项分布的概率密度函数图

3. 泊松分布：

若随机变量 X 可取一切非负整数值，且

$$p\{X=\mathrm{k}\}=\frac{\lambda^{k}}{k!}e^{-\lambda},k=0,1,2,\cdots \tag{1-10}$$

其中 $\lambda>0$，则称 X 服从泊松分布。简记为 $X\sim p(\lambda)$。图 1-12 给出了 λ 分别取 1、4、10 时泊松分布的概率密度函数的图形：

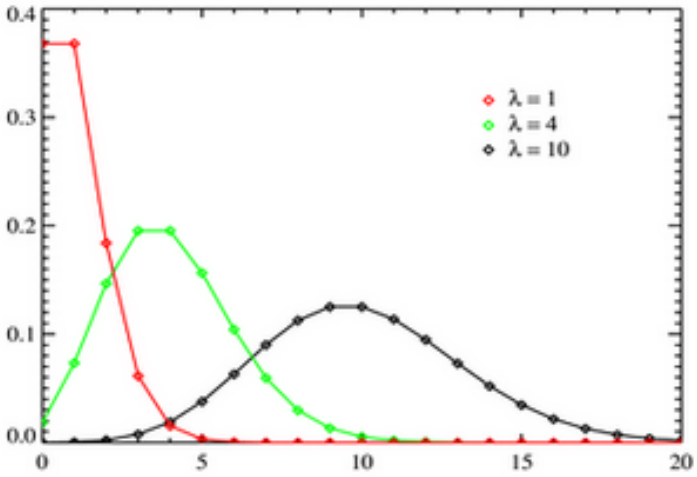


图 1-12 泊松分布的概率密度函数图

- 常见的连续型随机变量及其分布：

1. 均匀分布：

若 a,b 为有限数，由下列密度函数定义的分布称为[a,b]上的均匀分布：

$$p(\mathrm{x})=\begin{cases}\dfrac{1}{b-a}, & a\le x\le b\\ 0, & x<a\text{或}x>b\end{cases} \tag{1-11}$$

若随机变量 X 服从[a,b]上的均匀分布，则 X 在[a,b]中取值落在某一区域内的概率与这个区间的测度成正比。粗略地讲就是， X 取[a,b]中的任一点的可能性一样。当然也可以反过来看，均匀分布正是把这种直观的讲法严格化。图 1-13 给出了均匀分布的概率密度函数的示意图：

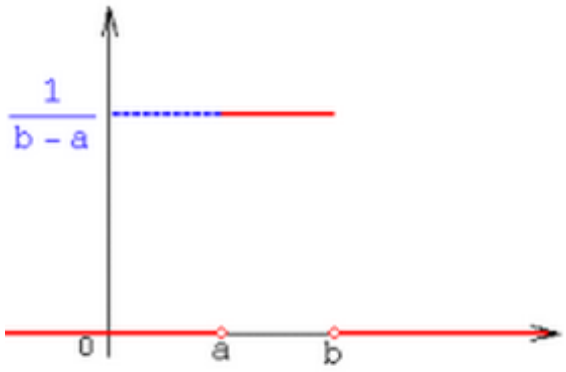


图 1-13 均匀分布的概率密度函数图

2. 正态分布

正态分布，也称高斯分布，是我们最常见的分布。下一小节将着重介绍。

1.5.2 正态分布及其性质

1．单变量正态分布

单变量正态分布的定义如式(1-12)所示。

$$\rho(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2] \tag{1-12}$$

其中：μ为随机变量 x 的期望，也就是平均值；σ^2为 x 的方差，σ为均方差，又称为标准差。μ与σ^2的定义分别由式(1-13)以及式(1-14)给出：

$$\mu=E(x)=\int_{-\infty}^{\infty}x\cdot\rho(x)dx \tag{1-13}$$

$$\sigma^2=\int_{-\infty}^{\infty}(x-\mu)^2\cdot\rho(x)dx \tag{1-14}$$

单变量正态分布的概率密度函数一般图形如图 1-14 所示：

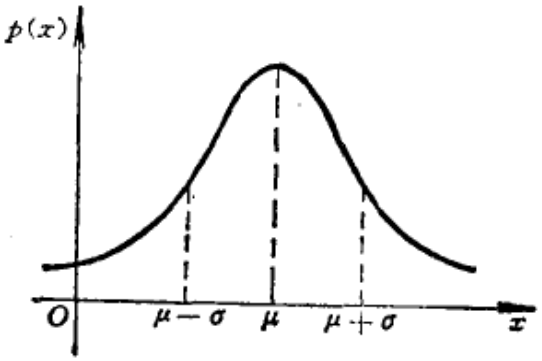


图 1-14 一维概率密度函数图

$\rho(x)$具有以下性质：

$$\begin{cases}\rho(x)\geq 0,(-\infty<x<\infty)\\ \int_{-\infty}^{\infty}\rho(x)dx=1\end{cases} \tag{1-15}$$

从$\rho(x)$的图形上可以看出，只要有两个参数μ和σ^2就可以完全确定其曲线。为了简单，常记$\rho(x)$为$N(\mu,\sigma^2)$。若从服从正态分布的总体中随机抽取样本 x，约有 95%的样本落在$(\mu-2\sigma,\mu+2\sigma)$中。样本的分散程度可以用$\sigma$来表示，$\sigma$越大分散程度越大。

2．多元正态分布

多元正态分布的定义如下：

$$\rho(\mathbf{x})=\frac{1}{(2\pi)^{d/2}|\sum|^{1/2}}\exp[-\frac{1}{2}(\mathbf{x}-\mu)^T\sum{}^{-1}(\mathbf{x}-\mu)] \tag{1-16}$$

其中：$\mathbf{x}=[x_1,x_2,\cdots\cdots,x_d]^T$ 为 d 维随机向量，对于 d 维随机向量 x，它的均值向量μ是 d 维的。也就是：$\mu=[\mu_1,\mu_2,\cdots\cdots,\mu_d]^T$ 为 d 维均值向量。

$\sum$是$d\times d$维协方差矩阵，$\sum^{-1}$是$\sum$的逆矩阵，$|\sum|$为$\sum$的行列式。协方差矩阵$\sum$是对称的，其中有$d\times(d+1)/2$个独立元素。由于$\rho(\mathbf{x})$可由μ和$\sum$完全确定，所以实际上$\rho(\mathbf{x})$可由$d+d\times(d+1)/2$个独立元素来确定。$(\mathbf{x}-\mu)^T$是$(\mathbf{x}-\mu)$的转置，且：$\mu=E\{\mathbf{x}\}$，$\sum=E\{(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\}$。

μ、$\sum$分别是向量 x 和矩阵$(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T$的期望。具体说：$x_i$是$\mathbf{x}$的第 i 个分量，$\mu_i$是$\mu$的第 i 个分量，$\sigma_{ij}{}^2$是$\sum$的第 i、j 个元素。

$$\mu_i=E[x_i]=\int x_i\rho(x)dx=\int_{-\infty}^{\infty}x_i\rho(x_i)dx_i \tag{1-17}$$

其中$\rho(x_i)$为边缘分布，$\rho(x_i)$的定义形如(1-18)所示：

$$\rho(x_i)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\rho(x)dx_1dx_2\cdots dx_d \tag{1-18}$$

对于二维随机变量 X 和 Y 作为一个整体，其分布函数 F（x，y），而 X 和 Y 都是随机变量，各别也有分布函数 $F_X(x)$、$F_Y(y)$，分别称为二维随机变量（X，Y）关于 X 和 Y 的边缘分布函数。有：$F_X(x)=F(x,+\infty)$和$F_Y(y)=F(+\infty,y)$。

对于离散随机变量有：$F_X(x)=F(x,+\infty)=\sum_{x_i\le x}\sum_{j=1}^{\infty}p_{ij}$，从中得到 X 的分布律为：$P\{X=x_i\}=\sum_{j=1}^{\infty}p_{ij}$。同样，Y 的分布律为$P\{Y=y_j\}=\sum_{i=1}^{\infty}p_{ij}$。

对于连续型随机变量（X，Y），假定它的概率密度为$f(x,y)$，由：$F_X(x)=F(x,+\infty)=\int_{-\infty}^{x}[\int_{-\infty}^{+\infty}f(x,y)dy]dx$可以知道，X 的概率密度为：$f_X(x)=\int_{-\infty}^{+\infty}f(x,y)dy$，同样也可以求出 Y 的概率密度函数。而：

$$\begin{aligned}\sigma_{ij}^2 &= E[(x_i-\mu_i)(x_j-\mu_j)] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_i-\mu_i)(x_j-\mu_j)\cdot\rho(x_i,x_j)dx_i dx_j]\end{aligned} \tag{1-19}$$

协方差矩阵：

$$\Sigma=\begin{bmatrix}\sigma_{11}^2 & \sigma_{12}^2 & \cdots & \cdots & \sigma_{1d}^2\\ \sigma_{12}^2 & \sigma_{22}^2 & \cdots & \cdots & \sigma_{2d}^2\\ \vdots & & & & \vdots\\ \vdots & & & & \vdots\\ \sigma_{1d}^2 & \sigma_{2d}^2 & & & \sigma_{dd}^2\end{bmatrix} \tag{1-20}$$

是一个对称矩阵，只考虑Σ为正定矩阵的情况，也就是$|\Sigma|$所有的子式都大于0。

即$|\sigma_{11}^2|\geq 0$，$\begin{vmatrix}\sigma_{11}^2 & \sigma_{12}^2\\ \sigma_{12}^2 & \sigma_{22}^2\end{vmatrix}\geq 0$，……

同单变量正态分布一样，多元正态分布$\rho(\mathbf{x})$可以由μ和Σ完全确定，常记为$N(\mu,\Sigma)$。

3．多元正态分布的性质

（1）参数μ和Σ对分布的决定性

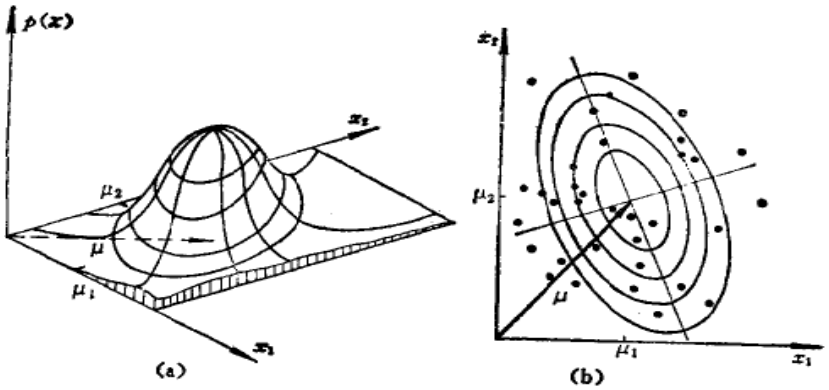
对于 d 维随机向量 x，它的均值向量μ也是 d 维的，协方差矩阵是对称的，其中有$d(d+1)/2$个独立元素。$\rho(\mathbf{x})$可由μ和Σ完全确定，实际上$\rho(\mathbf{x})$可由$d+d(d+1)/2$个独立元素决定。常记为：$\rho(\mathbf{x})\sim N(\mu,\Sigma)$。

（2）等密度点的轨迹为一超椭球面

由$\rho(\mathbf{x})$的定义可知，当右边指数项为常数时，密度$\rho(\mathbf{x})$的值不变，所以等密度点满足：

$$(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)=\text{常数} \tag{1-21}$$

可以证明，上式的解是一个超椭球面，其主轴方向取决于Σ的本征向量（特征向量），主轴的长度与相应的本征值成正比。如图 1-15 所示：



(a) $p(x)$ ($d=2$)；(b) 等密度点轨迹

图 1-15 等密度点的轨迹为一超椭球面

从上图可以看出，从正态分布总体中抽取的样本大部分落在由 μ 和 $\sum$ 所确定的一个区域里，这个区域的中心由均值向量 μ 决定，区域的大小由协方差矩阵决定。

在数理统计中，令：$\gamma^2=(\mathbf{x}-\mu)^T\sum^{-1}(\mathbf{x}-\mu)$，式中 γ 称为 x 到 μ 的马氏距离（Mahalanobis）距离。所以，等密度点轨迹是 x 到 μ 的马氏距离 γ 为常数的超椭球面。该超椭球面构成的球体的大小是样本对于均值向量的“离散度度量”。

体积：$v=v_d\cdot|\sum|^{\frac{1}{2}}\cdot\gamma^d$，其中，

$$v_d=\begin{cases}\dfrac{\pi^{\frac{d}{2}}}{(\frac{d}{2})!} & d\text{为偶数}\\ \dfrac{2^d\cdot\pi^{(d-1)/2}\cdot(\frac{d-1}{2})!}{d!} & d\text{为奇数}\end{cases}\tag{1-22}$$

如果 d 确定了，则 v_d 不变，v 与 $|\sum|^{\frac{1}{2}}$ 有关。也就是对于给定的维数 d，样本离散度随 $|\sum|^{\frac{1}{2}}$ 而变。

（3）不相关性等价于独立性

概率论中，两个随机变量 x_i 和 x_j 之间不相关，并不意味着它们一定独立。

如果 x_i 和 x_j 之间不相关，则 $x_i x_j$ 的数学期望有：

$$E(x_i x_j) = E(x_i) \cdot E(x_j) \tag{1-23}$$

如果 x_i 和 x_j 相互独立，则有：

$$P(x_i, x_j) = P(x_i) \cdot P(x_j) \tag{1-24}$$

独立性是比不相关更强的条件。不相关反映了 x_i 和 x_j 的总体性质。如果 x_i 和 x_j 相互独立，则它们之间一定不相关，反之则不成立。但是对服从正态分布的两个分量 x_i 和 x_j，若 x_i 与 x_j 互不相关，则它们之间一定独立。

证明：根据定义，x_i 和 x_j 的协方差 $\sigma_{ij}^2 = E[(x_i - \mu_i)(x_j - \mu_j)]$，又根据不相关定义 $E(x_i, x_j) = E(x_i) \cdot E(x_j)$ 有：

$$\sigma_{ij}^2 = E[(x_i - \mu_i)(x_j - \mu_j)] = E(x_i - \mu_i) \cdot E(x_j - \mu_j) \tag{1-25}$$

又：$\mu_i = E(x_i)$，$E[(x_i - \mu_i) = E(x_i) - E(\mu_i) = E(x_i) - \mu_i = 0$.所以有 $\sigma_{ij}^2 = 0$.

协方差矩阵 $\sum = \begin{bmatrix} \sigma_{11}^2 & & \\ & \ddots & \\ & & \sigma_{dd}^2 \end{bmatrix}$ 成为对角阵。

可以计算出：$\sum^{-1} = \begin{bmatrix} \frac{1}{\sigma_{11}^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_{dd}^2} \end{bmatrix}$，$|\sum| = \prod_{i=1}^{d} \sigma_{ii}^2$，$|\sum|^{\frac{1}{2}} = \prod_{i=1}^{d} \sigma_{ii}$，故：

$$(x-\mu)T \cdot \sum^{-1}(x-\mu) = [x_1 - \mu_1, \cdots\cdots, x_d - \mu_d] \begin{bmatrix} \frac{1}{\sigma_{11}^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_{dd}^2} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ \vdots \\ x_d - \mu_d \end{bmatrix} = \sum_{i=1}^{d} \left(\frac{x_i - \mu_i}{\sigma_{ii}}\right)^2 \tag{1-26}$$

因此有：

$$\begin{aligned}\rho(\mathrm{x}) &= \frac{1}{(2\pi)^{d/2}|\sum|^{1/2}}\exp[-\frac{1}{2}(\mathrm{x}-\mu)^T\sum{}^{-1}(\mathrm{x}-\mu)] \\ &= \prod_{i=1}^{d}\frac{1}{\sqrt{2\pi}\sigma_{ii}}\cdot\exp\left\{-\frac{1}{2}(\frac{x_i-\mu_i}{\sigma_{ii}})^2\right\} = \prod_{i=1}^{d}\rho(x_i)\end{aligned} \tag{1-27}$$

根据独立性的定义：正态分布随机向量的各分量间互不相关性与相互独立等价。

（4）边缘分布与条件分布的等价性

不难证明正态随机向量的边缘分布与条件分布仍服从正态分布。从（3）证明得出的结论 $\rho(x)$ 表达式，如果 x 用 x_1 表示，有：

$$\rho(x_1) = \frac{1}{\sqrt{2\pi}\sigma_{ii}}\cdot\exp(-\frac{1}{2}(\frac{x_1-\mu_1}{\sigma_{11}})^2) \tag{1-28}$$

也就是说，边缘分布 $\rho(x_1)$ 服从均值为 μ_1，方差为 σ_{11}^2 的正态分布：$\rho(x_1)\sim N(\mu_1,\sigma_{11}^2)$.同理，$\rho(x_2)\sim N(\mu_2,\sigma_{22}^2)$.

另外，给定 x_1 的条件下 x_2 的分布：

$$\rho(x_2\mid x_1) = \frac{\rho(x_1,x_2)}{\rho(x_1)} \tag{1-29}$$

$$\rho(x_1,x_2) = \frac{1}{2\pi|\sum|^{1/2}}\exp\left\{-\frac{1}{2|\sum|}\left[\sigma_{22}^2(x_1-\mu_1)^2+\sigma_{11}^2(x_2-\mu_2)^2-\sigma_{12}^2(x_1-\mu_1)(x_2-\mu_2)\right]\right\} \tag{1-30}$$

则 $\rho(x_2\mid x_1)$ 服从正态分布，同理 $\rho(x_1\mid x_2)$ 也服从正态分布。

（5）线性变换的正态性

对于多元随机向量的线性变换，仍为多元正态分布的随机向量。就是：x 服从正态分布 $\rho(\mathbf{x})\sim N(\mu,\sum)$，对 x 作线性变换 $\mathbf{y}=A\mathbf{x}$，其中 A 为线性变换矩阵，且 $|A|\neq 0$，则 y 服从正态分布：$\rho(\mathbf{y})\sim N(A\mu,A\sum A^T)$

（6）线性组合的正态性

若 x 为多元正态随机向量，则线性组合 $\mathbf{y}=\mathbf{a}^T\mathbf{x}$ 是一维的正态随机变量：

$$\rho(\mathbf{y})\sim N(\mathbf{a}^T\mu,\mathbf{a}^T\sum\mathbf{a}) \tag{1-31}$$

其中，a 与 x 同维。

1.5.3 混合分布模型

上面提到的分布都相对简单，而实际中，很多数据并不完美地适应某个分布。此时，我们常用多个简单模型的线性组合来刻画数据。高斯混合模型使我们最常使用的模型，下面我们着重介绍混合高斯模型的基本概念，它的参数求解方法常用 EM 算法，我们将在第七章作介绍。

1、 单高斯分布模型（GSM）

多维变量 X 服从高斯分布时，它的概率密度函数 PDF 为：

$$N(\mathrm{x};\mu,\Sigma)=\frac{1}{\sqrt{2\pi|\Sigma|}}\exp[-\frac{1}{2}(\mathrm{x}-\mu)^{T}\Sigma^{-1}(\mathrm{x}-\mu)] \tag{1-32}$$

其中 x 是维度为 d 的列向量，u 是模型期望，Σ 是模型方差。在实际应用中 u 通常用样本均值来代替，Σ 通常用样本方差来代替。很容易判断一个样本 x 是否属于该类别。因为每个类别都有自己的 u 和 Σ，把 x 代入上式，当概率大于一定阈值时我们就认为 x 属于此类，即能用此高斯分布刻画。

从几何上讲，单高斯分布模型在二维空间应该近似于椭圆，在三维空间上近似于椭球。遗憾的是在很多分类问题中，属于同一类别的样本点并不满足“椭圆”分布的特性。这就引入了高斯混合模型。如图 1-16 所示，（a）中的数据显然不成“椭圆”形状，因此不能用单一的高斯模型去刻画，而像（b）所示，可以将数据分割为 3 个部分，每个部分都近似成“椭圆”形状，可以用高斯模型刻画，因此整个数据可以用高斯混合模型来刻画。

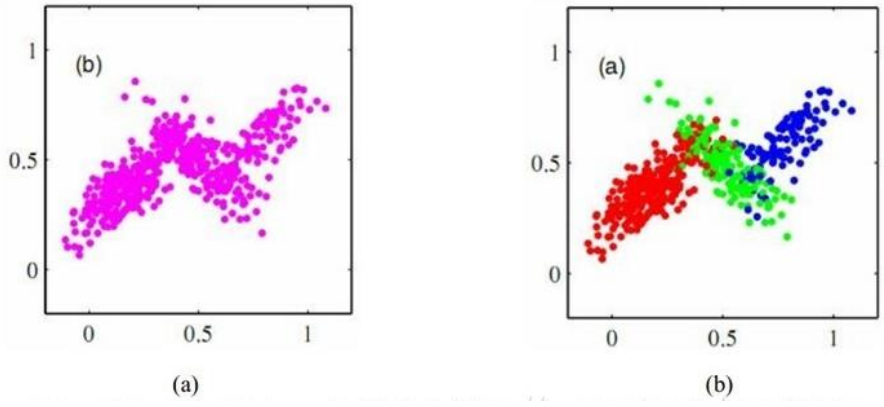


图 1-16 高斯混合模型

2、 高斯混合模型（GMM）

GMM 认为数据是从几个 GSM 中生成出来的，即

$$p(\mathrm{x})=\sum_{k=1}^{K}\pi_k N(\mathrm{x}\mid\mu_k,\Sigma_k) \tag{1-33}$$

其中，K 需要事先确定好，就像 K-means 中的 K 一样。π_k 是权值因子。其中的任意一个高斯分布 $N(\mathrm{x}\mid\mu_k,\Sigma_k)$ 称为这个模型的一个组分（component）。这里有个问题，为什么我们要假设数据是由若干个高斯分布组合而成的，而不假设是其他分布呢？实际上不管是什么分布，只要 K 取得足够大，这个 Mixture Model 就会变得足够复杂，就可以用来逼近任意连续的概率密度分布。因为高斯函数具有良好的计算性能，所以 GMM 被广泛地应用。

1.6 习题

1、举例说明模式和模式识别的概念。

2、试论述完整模式识别过程的主要阶段和操作（从在客观世界里采集模式样本，到把模式样本区分成不同的类型）。

3、为完成一次肝病的分析研究，对 100 名肝病患者化验了肝功，共取得 10 个原始数据。之后经过分析综合，对每个患者得到碘反应和转胺酶等 5 个主要数据，最后根据其中的 3 个数据把患者区分为甲肝和乙肝各 50 名。针对这一过程，说明什么是模式样本？共抽取了几个样本？这些样本被区分成几个类型？每个类型含几个样本？模式空间、特征空间和类型空间的维数各是多少？

4、试说明模式识别系统的组成，以及训练过程和判决过程的作用和关系。

5、结合实例谈谈你对机器学习的认识。并给出几种机器学习的主要方法，以及各种方法的特点。

6、试写出正态分布中的类概率密度函数表达式，并且给出它的边缘概率密度函

数；假设随机向量各个分量是彼此独立的，给出类概率密度函数和它的边缘密度函数之间的关系（类型数目为 c）。

7、证明正态随机向量的线性变换 $y = Ax$ 仍是正态分布的，其中 A 是非奇异的线性变换矩阵；给出 y 的均值向量和协方差矩阵与 x 的均值向量和协方差矩阵之间的关系。

8、假设两变量 x, z 相互独立，试证它们和的均值和方差满足：

$$\mathrm{E}[x+z] = \mathrm{E}[x] + \mathrm{E}[\mathrm{z}]$$

$$\mathrm{var}[x+z] = \mathrm{var}[x] + \mathrm{var}[z]$$