引　言

1.1　什么是机器学习

这是一个“大数据”时代。过去，只有公司才拥有数据。那时，有一些计算中心，数据在那里存储和处理。先是个人计算机的出现，而后是无线通信的广泛使用，使得我们都成了数据的生产者。每当我们购买一件商品、租借一部电影、访问一个网页、书写一个博客或在社交媒体上发帖子时，甚至当我们散步或开车闲逛时，我们都在产生数据。

我们每个人不仅是数据的生产者，而且也是数据的消费者。我们想要适合的产品和服务，希望我们的需要能被理解，我们的兴趣能被预测到。

以一家连锁超市为例，它通过遍布全国的数百家实体商店或通过网上的虚拟商店向数百万顾客销售数千种商品。每笔交易的细节，包括交易日期、顾客ID、购买的商品和数量、付款金额等都存储在计算机中。这意味每天都有大量的数据。连锁超市希望能够预测哪位顾客可能会购买哪种商品，以便能够使销售和利润最大化。类似地，每位顾客都希望找到最适合他们需要的商品。

这一任务并非显而易见。我们并不确切地知道哪些人比较倾向于购买这种口味的冰激凌，这位作家的下一本书是什么，也不知道谁喜欢看这部新电影、访问这座城市，或点击这一链接。顾客的行为随时间和地点而变化。但是，我们知道这不是完全随机的。人们去超市并不是随机购买商品。当他们买啤酒时，也会买薯片；夏天买冰激凌，而冬天为Glühwein[⊖]买香料。数据中存在确定的模式。

为了在计算机上解决问题，我们需要算法。算法是指令的序列，它把输入变换成输出。例如，我们可以为排序设计一个算法，输入是数的集合，而输出是它们的有序列表。对于相同的任务，可能存在不同的算法，而我们感兴趣的是找到需要的指令、内存最少，或者二者都最少的最有效算法。

然而，对于某些任务，我们没有算法。预测顾客的行为就是一个例子，另一个例子是区分垃圾邮件和正常邮件。我们知道输入是邮件文档，在最简单的情况下是一个字符文件。我们还知道输出应该是指出消息是否为垃圾邮件的“是”或“否”。但是我们不知道如何把这种输入变换成输出。所谓的垃圾邮件随时间而变，因人而异。

我们缺乏的是知识，作为补偿我们有数据。我们可以很容易地编辑数以千计的实例消息，其中一些我们知道是垃圾邮件，而我们要做的是希望从中“学习”垃圾邮件的结构。换言之，我们希望计算机(机器)自动地为这一任务提取算法。不需要学习如何将数排序，因为我们已经有这样的算法。但是，对于许多应用而言，我们确实没有算法，而是有实例数据。

我们也许不能够完全识别该过程，但是我们相信，我们能够构造一个好的并且有用的近似。尽管这样的近似还不可能解释一切，但其仍然可以解释数据的某些部分。我们相

⊖ Glühwein是一种温热、有点甜味儿、加香料的葡萄酒。圣诞节期间，在欧洲很受欢迎。——译者注

信，尽管识别整个过程也许是不可能的，但是我们仍然能够发现某些模式或规律。这正是机器学习的定位。这些模式可以帮助我们理解该过程，或者我们可以使用这些模式进行预测：假定将来(至少是不远的将来)情况不会与收集样本数据时有很大的不同，则未来的预测也将有望是正确的。

机器学习方法在大型数据库中的应用称为*数据挖掘*(data mining)。类似的情况如大量的金属氧化物以及原料从矿山中开采出来，处理后产生少量非常珍贵的物质。类似地，在数据挖掘中，需要处理大量的数据以构建有使用价值的简单模型，例如具有高准确率的预
2 测模型。数据挖掘的应用领域非常广泛：除零售业以外，在金融业，银行分析历史数据，构建用于信用分析、诈骗检测、股票市场等方面的应用模型；在制造业，学习模型可以用于优化、控制以及故障检测等；在医学领域，学习程序可以用于医疗诊断等；在电信领域，通话模式的分析可用于网络优化和提高服务质量；在科学研究领域，比如物理学、天文学以及生物学的大量数据只有使用计算机才可能得到足够快的分析。万维网是巨大的，并且在不断增长，因此在万维网上检索相关信息不可能依靠人工完成。

然而，机器学习不仅仅是数据库方面的问题，它也是人工智能的组成部分。为了智能化，处于变化环境中的系统必须具备学习能力。如果系统能够学习并且适应这些变化，那么系统的设计者就不必预见所有的情况并为它们提供解决方案了。

机器学习还可以帮助我们解决视觉、语音识别以及机器人方面的许多问题。以人脸识别问题为例。我们做这件事毫不费力。即使姿势、光线、发型等不同，我们每天还是可以通过观察真实的面孔或照片来认出家人和朋友。但是我们做这件事是无意识的，而且无法解释我们是如何做的。因为我们不能够解释我们所具备的这种技能，所以我们也就不可能编写相应的计算机程序。但是我们知道，脸部图像并非只是像素点的随机组合；人脸是有结构的、对称的。脸上有眼睛、鼻子和嘴巴，并且它们都位于脸的特定部位。每个人的脸都有各自的眼睛、鼻子和嘴巴的特定组合模式。通过分析一个人的脸部图像的多个样本，学习程序可以捕捉到那个人特有的模式，然后在所给的图像中检测这种模式，从而进行辨认。这就是*模式识别*(pattern recognition)的一个例子。

机器学习使用实例数据或过去的经验训练计算机来优化某种性能标准。我们有依赖于某些参数的模型，而学习就是执行计算机程序，利用训练数据或以往经验来优化该模型的参数。模型可以是*预测性的*(predictive)，用于未来的预测；或者是*描述性的*(descriptive)，用于从数据中获取知识；也可以二者兼备。

机器学习在构建数学模型时利用了统计学理论，因为其核心任务就是由样本推理。计
3 算机科学的角色是双重的：第一，在训练时，我们需要求解优化问题以及存储和处理通常所面对的海量数据的高效算法。第二，一旦学习得到了一个模型，它的表示和用于推理的算法解也必须是高效的。在特定的应用中，学习或推理算法的效率，即它的空间复杂度和时间复杂度，可能与其预测的准确率同样重要。

现在，让我们更详细地讨论一些应用领域的例子，以便进一步深入了解机器学习的类型和用途。

1.2 机器学习的应用实例

1.2.1 学习关联性

在零售业，例如超市连锁店，机器学习的一个应用是*购物篮分析*(basket analysis)。它的任务是发现顾客所购商品之间的关联性：如果购买商品 X 的人通常也购买商品 Y，而

一位顾客购买了商品 X 却未购买商品 Y，则他就是商品 Y 的潜在顾客。一旦我们发现这类顾客，我们就能针对他们实施交叉销售策略。

为了发现关联规则(association rule)，我们对学习形如 $P(Y|X)$ 的条件概率感兴趣，其中 X 是我们知道的顾客已经购买的商品或商品集，Y 表示在条件 X 下可能购买的商品。

假定考察已有的数据，计算得到 $P(\text{chips}|\text{beer})=0.7$，那么我们就可以定义规则：

购买啤酒(beer)的顾客中有70%的人也买了薯片(chip)

我们也许想要区分不同的顾客。针对这个问题，我们需要估计 $P(Y|X, D)$，其中 D 是顾客的一组属性，如性别、年龄、婚姻状况等，这里假定我们已经得到了这些属性信息。如果考虑书店而不是超市销售问题，商品就可能是书或作者等。对于 Web 门户网站入口问题，商品对应于到 Web 网页的链接，而我们可以估计用户可能点击的链接，并利用这些信息预先下载这些网页，以便取得更快的网页访问速度。

4

1.2.2 分类

信贷是金融机构(例如银行)借出的一笔钱，需要连本带息偿还，通常分期偿还。对银行来说，重要的是能够提前预测贷款风险。这种风险是客户不履行义务和不全额还款的可能性。既要确保银行获利，又要确保不会因提供超出客户财力的贷款而给客户带来不便。

在资信评分(credit scoring)(Hand 1998)中，银行计算在给定信贷额度和客户信息情况下的风险。客户信息包括我们已经获取的数据以及与计算客户财力相关的数据，即收入、存款、担保、职业、年龄、以往经济记录等。银行有以往贷款的记录，包括客户数据以及贷款是否偿还。通过这类特定的申请数据，可以推断出表示客户属性及其风险关联性的一般规则。也就是说，机器学习系统用一个模型来拟合过去的数据，以便能够对新的申请计算风险，从而决定接受或拒绝该项申请。

这是分类(classification)问题的一个例子，这里有两个类：低风险客户和高风险客户。客户信息作为分类器的输入(input)，分类器的任务是将输入指派到其中的一个类。

利用以往数据进行训练后，学习得到的规则可能具有如下形式：

IF income $>\theta_1$ AND savings $>\theta_2$

THEN low-risk ELSE high-risk

其中 θ_1 和 θ_2 是合适的值(参见图 1-1)。这是判别式(discriminant)的一个例子，判别式是将不同类的样本分开的函数。

有了这样的规则，主要用途就是预测(prediction)：一旦我们拥有拟合以往数据的规则，如果未来与过去类似，那么我们就能够对新的实例做出正确的预测。如果给定一个具有特定收入(income)和存款(savings)的新申请，则我们就可以容易地判断出它是低风险(low-risk)还是高风险(high-risk)。

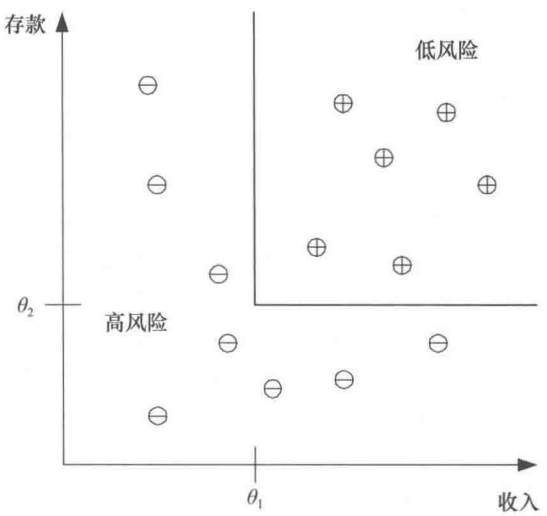

图 1-1　训练数据集例子，其中每个圆圈对应于一个数据实例，输入值在对应的坐标上，符号指示类别。为了简单起见，输入只包括客户的收入和存款两种属性，两个类分别为低风险("+")和高风险("−")。图中还显示了分隔两类样本的判别式的例子

在某些情况下，我们可能不是希望做0/1(低风险/高风险)类型的判断，而是希望计算一个概率值$P(Y|X)$，其中X是顾客属性，Y是0或1，分别表示低风险和高风险。从这个角度来看，我们可以将分类看作学习从X到Y的关联性。于是，给定$X=x$，如果有$P(Y=1|X=x)=0.8$，则我们就说该客户为高风险的可能性有80%，或者等价地说，该客户为低风险的可能性有20%。然后，我们可以根据可能的收益和损失来决定接受还是拒绝这笔贷款业务。

机器学习在模式识别(pattern recognition)方面有很多应用。其中之一是光学字符识别(Optical Character Recognition，OCR)，即从字符图像识别字符编码。这是多类问题的一个例子，类与我们想要识别的字符一样多。特别有趣的是手写体字符的识别问题。人们有不同的书写风格，字体有大有小，倾斜角度不同，还有用钢笔或用铅笔之别，所以同一个字符可能会有许多种可能的图像。尽管书写是人类的发明创造，但是还没有像人类读者一样准确的系统。我们没有字符“A”的形式化描述，涵盖所有“A”而不涵盖任何非“A”。没有这种形式化描述，我们就要从书写者那里取样，从这些实例中学习关于“A”的定义。然而，尽管我们不知道是什么因素使得一个图像被识别为“A”，但是我们确信所有这些不同的“A”的图像都具有某些共同的特征，这正是我们希望从实例中提取的。我们知道，字符图像不只是随机点的集合。它是笔画的集合，并且是有规律的，通过学习程序我们能够捕获这些规律。

5～6

阅读文本时，我们能够利用的一个因素是人类语言的冗余性。词是字符的序列，并且相继的符号不是独立的，而是被语言的词所约束。这有好处，即便有一个符号不能识别，我们仍可以读出词t?e[⊖]。根据语言的语法和语义，这种上下文的依赖性还可能出现在词和句子之间等较高的层次上。目前存在用于学习序列和对这种依赖性建模的机器学习算法。

对于人脸识别(face recognition)，输入是人脸图像，而类是需要识别的人，并且学习程序应当学习人脸图像与身份之间的关联性。这个问题比光学字符识别更困难，因为人脸会有更多的类，输入图像也更大一些，并且人脸是三维的，不同的姿势和光线等都会导致图像的显著变化。另外，对于特定人脸的输入也会出现问题，比如说眼镜可能会把眼睛和眉毛遮住，胡子可能会把下巴盖住等。

在医学诊断(medical diagnosis)中，输入是关于患者的信息，而类是疾病。输入包括患者的年龄、性别、既往病史、目前症状等。当然，患者可能还没有做过某些检查，因此这些输入将会缺失。检查需要时间，还可能要花很多钱，而且也许还会给患者带来不便。因此，除非我们确信检查将提供有价值的信息，否则我们不对患者进行检查。在医学诊断的情况下，错误的诊断结果可能会导致错误的治疗或根本不治疗。在不能确信诊断结果的情况下，分类器最好还是放弃判定，而等待医学专家来决断。

在语音识别(speech recognition)中，输入是语音，类是可以读出的词汇。这里要学习的是从语音信号到某种语言的词汇的关联性。由于年龄、性别或口音方面的差异，不同的人对于相同词汇的读音不同，这使得语音识别相当困难。语音识别的另一个特点是其输入信号是时态的(temporal)，词汇作为音素的序列实时读出，而且有些词汇的读音会比其他词汇长一些。

7

语音信息的作用有限，并且与光学字符识别一样，在语音识别中，“语言模型”的集成是至关重要的，而且提供语言模型的最好方法仍然是从实例数据的大型语料库中学习。机

⊖ 这里，“?”表示不能识别的符号。——译者注

器学习在自然语言处理(natural language processing)方面的应用与日俱增。垃圾邮件过滤就是一种应用，那里垃圾邮件的制造者为一方，过滤者为另一方，它一直都在寻找越来越精巧的方法来超越对方。大型文档汇总是另一个有趣的例子；还有一个例子是分析博客或社交网站上的帖子，以便提取"流行"主题或决定做什么广告。也许最吸引人的是机器翻译(machine translation)。经历了数十年手工编写翻译规则的研究之后，最近人们认识到最有希望的方法是提供大量两种语言文本的实例对，让程序自动地揣摩把一种语言映射到另一种语言的规则。

生物测定学(biometrics)使用人的生理和行为特征来识别或认证人的身份，它需要集成来自不同形态的输入。生理特征的例子有面部图像、指纹、虹膜和手掌；行为特征的例子有签字的力度、嗓音、步态和击键。与通常的鉴别过程(照片、印刷签名或口令)相反，会有许多不同的(不相关的)输入，伪造(欺骗)更困难，并且系统更准确，有望不会对用户太不方便。机器学习既用于对这些不同形态构建不同的识别器，也考虑这些不同数据源的可靠性，用于组合它们的决策，以便得到接受或拒绝的总体决断。

从数据中学习规则也为知识抽取(knowledge extraction)提供了可能性。规则是一种解释数据的简单模型，而观察该模型我们就能得到潜在数据处理的解释。例如，一旦我们学习得到区分低风险客户和高风险客户的判别式，我们就拥有了关于低风险客户特性的知识。然后，我们就能够利用这些信息，通过广告等方式，更有效地争取那些潜在的低风险客户。机器学习还可以进行压缩(compression)。用规则拟合数据，我们得到比数据更简 8
单的解释，需要的存储空间更少，处理所需要的计算更少。例如，一旦掌握了加法规则，就不必记忆每对可能数的和是多少。

机器学习的另一种用途是离群点检测(outlier detection)，即发现那些不遵守规则和例外的实例。基本思想是，典型的实例具有一些可以简单陈述的特征，而不具备这些特征的实例都是非典型的。在这种情况下，我们感兴趣的是找到一个尽可能简单并且覆盖尽可能多的典型实例的规则。落在外面的实例都是例外，它们可能是提示我们需要注意的异常(如诈骗)，也可能是新颖的、先前未曾见过但又合理的情况。因此，离群点检测又称为新颖性检测(novelty detection)。

1.2.3　回归

假设我们想要一个能够预测二手车价格的系统。该系统的输入是我们认为会影响车价的属性信息：品牌、车龄、发动机排量、里程以及其他信息。输出是车的价格。这种输出为数值的问题是回归(regression)问题。

设 X 表示车的属性，Y 表示车的价格。调查以往的交易情况，我们能够收集训练数据，而机器学习程序用一个函数拟合这些数据来学习 X 的函数 Y。图 1-2 给出了一个例子，其中对于 w 和 w_0 的合适值，拟合函数具有以下形式：

$$y=wx+w_0$$

回归和分类均为监督学习(supervised learning)问题，其中给定输入 X 和输出 Y，任务是学习从输入到输出的映射。机器学习的方法是，先假定依赖于一组参数的模型：

$$y=g(x|\theta)$$

其中，$g(\cdot)$是模型，而 θ 是模型的参数。对于回归，Y 是数值；对于分类，Y 是类编码(如 0/1)。$g(\cdot)$为回归函数，或者(对于分类)是将不同类的实例分开的判别式函数。机器学习程序优化参数 θ，使近似误差最小，也就是说，我们的估计要尽可能地接近训练集

中给定的正确值。例如，图 1-2 所示的模型是线性的，w 和 w_0 是为最佳拟合训练数据优化后的参数。在线性模型限制过强的情况下，我们可以利用二次函数

$$y=w_2x^2+w_1x+w_0$$

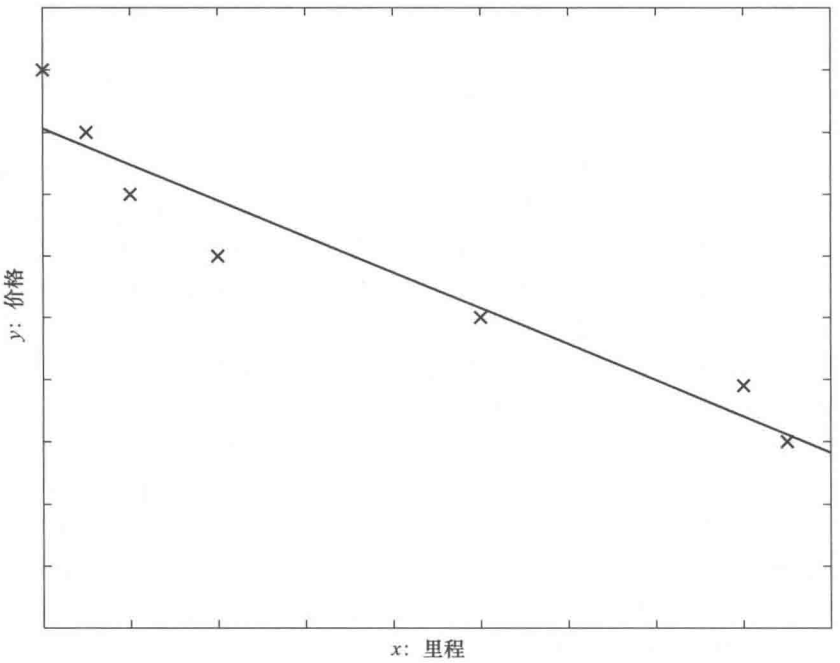

图 1-2 二手车的训练数据及其拟合函数。为简单起见，这里采用线性模型，输入属性也只有里程

或更高阶的多项式，或其他非线性函数，为最佳拟合优化它们的参数。

回归的另一个例子是移动机器人导航，例如，自动汽车导航，其中输出是每次转动车轮的角度，使汽车前进而不会撞到障碍物或偏离车道。在这种情况下，输入由汽车上的传感器(如视频相机、GPS 等)提供。训练数据可以通过监视和记录驾驶员的动作来收集。

我们可以想象回归的其他应用，这里我们试图优化一个函数[⊖]。假设我们想要制造一个焙炒咖啡的机器。该机器有多个影响咖啡品质的输入：温度、时间、咖啡豆种类等。我们针对不同的输入配置进行大量试验，并估量咖啡的品质。例如，根据消费者的满意度测量咖啡的品质。为找到最优配置，我们拟合一个联系这些输入和咖啡品质的回归模型，并在当前模型的最优样本附近选择一些新的点，以便寻找更好的配置。我们抽取这些点，检测咖啡的品质，将它们加入训练数据，并拟合新的模型。这通常称为*响应面设计*(response surface design)。

有时，我们希望能够学习一个相对位置，而不是估计一个绝对数值。例如，在电影*推荐系统*(recommendation system)中，我们希望产生一张表，按照用户的喜欢程度将电影排序。根据电影的体裁、演员等属性，并使用用户对他们所看过电影的评级，我们希望能够学习一个*排名*(ranking)函数，然后可以使用它选择新电影。

9 ~ 10

1.2.4 非监督学习

在监督学习中，我们的目标是学习从输入到输出的映射关系，其中输出的正确值已经

⊖ 感谢 Michael Jordon 提供这个例子。

由指导者提供。然而，在非监督学习中却没有这样的指导者，只有输入数据。我们的目标是发现输入数据中的规律。输入空间存在着某种结构，使得特定的模式比其他模式更常出现，而我们希望知道哪些经常发生，哪些不经常发生。在统计学中，这称为密度估计(density estimation)。

密度估计的一种方法是聚类(clustering)，其目标是发现输入数据的簇或分组。对于拥有老客户数据的公司，客户数据包括客户的个人统计信息及其以前与公司的交易，而公司也许想知道其客户的分布，搞清楚什么类型的客户会频繁出现。这种情况下，聚类模型会将属性相似的客户分派到相同的分组，为公司提供其客户的自然分组，这称作客户划分 11
(customer segmentation)。一旦找出了这样的分组，公司也许会做出一些决策，比如对不同分组的客户提供特别的服务和产品等，这称作客户关系管理(customer relationship management)。这样的分组也可以用于识别“离群点”，即那些不同于其他客户的客户，这可能意味新的市场商机，公司可以进一步开发。

聚类的一个有趣的应用是图像压缩(image compression)。在这种情况下，输入实例是由 RGB 值表示的图像像素。聚类程序将颜色近似的像素分到相同的分组，而这样的分组对应于图像中频繁出现的颜色。如果图像中只有少数几种颜色，并且属于同一分组的像素用一种颜色(例如，颜色的平均值)进行编码，则图像被量化。假设像素是 24 位，表示 1600 万种颜色，但是如果只有 64 种主色调，那么对于每个像素，只需要 6 位而不是 24 位。例如，如果景象在图像的不同部分有多种不同的蓝色色调，并且采用它们的平均值来表示所有这些蓝色，那么就丢失了图像的细节，但是赢得了图像的存储空间和传送时间。在理想情况下，人们希望通过分析重复的图像模式(如纹理、对象等)来识别更高层次的规律性。这为更高层次、更简单、更有用地描述景象提供了可能，并且实现了比像素级更好的压缩。如果我们扫描了文档页，则我们得到的不是随机的有/无像素，而是一些字符的位图。这样的数据是有结构的，并且我们利用这些冗余信息，找出数据的较短描述：“A”的 16×16 的位图占 32 字节，其 ASCII 码只占 1 字节。

在文档聚类(document clustering)中，目标是把相似的文档分组。例如，新闻报道可以进一步划分为政治、体育、时尚、艺术等子组。通常，文档用词袋(bag of words)表示，即预先定义 N 个词的词典，并且每个文档都是一个 N 维二元向量，如果第 i 个词出现在该文档中，则其第 i 个分量取 1。删除后缀“-s”和“-ing”等，以避免重复，并且不用“of”、“and”等不包含什么信息的词。然后，文档根据它们包含的相同词的个数分组。当然，如何选取词典是至关重要的。

机器学习方法还应用于生物信息学(bioinformatics)。在我们的基因组中，DNA 是“生命的蓝图”，也是碱基(即 A、G、C 和 T)序列。RNA 由 DNA 转录而来，而蛋白质由 12
RNA 转换而来。蛋白质就是生命体和生命体的产物。正如 DNA 是碱基序列，蛋白质则是氨基酸(由碱基定义)序列。计算机科学在分子生物学的应用领域之一就是比对(alignment)，即将一个序列与另一个匹配。这是一个困难的串匹配问题，因为序列可能相当长，有很多模板串要进行匹配，并且还可能会删除、插入和置换。聚类用于学习基序(motifs)，这是蛋白质结构中反复出现的氨基酸序列。基序之所以令人感兴趣，是因为它们可能对应于它们所表征的序列内部的结构或功能要素。比方说，如果氨基酸是字母，蛋白质是句子，那么基序就像单词，即具有特别意义、频繁出现在不同句子中的一串字母。

1.2.5　增强学习

在某些应用中，系统的输出是动作(action)的序列。在这种情况下，单个的动作并不

重要，重要的是策略(policy)，即达到目标的正确动作的序列。不存在中间状态中最好动作这种概念。如果一个动作是好的策略的组成部分，那么该动作就是好的。在这种情况下，机器学习程序就应当能够评估策略的好坏程度，并从以往好的动作序列中学习，以便能够产生策略。这种学习方法称为增强学习(reinforcement learning)算法。

游戏(game playing)是一个很好的例子。在游戏中，单个移动本身并不重要，正确的移动序列才是重要的。如果一个移动是一个好的游戏策略的一部分，则它就是好的。游戏是人工智能和机器学习的一个重要研究领域。这是因为游戏容易描述，但又很难玩好。像国际象棋这样的游戏，其规则只有少量的几条，但是它非常复杂，因为在每种状态下都有大量可行的移动，并且每局又都包含大量的移动。一旦有了能够学习如何玩好游戏的好算法，我们也可以将这些算法用在具有更显著经济效益的领域。

在某种环境下搜寻目标位置的机器人导航是增强学习的另一个应用领域。在任何时
13 候，机器人都能够朝着多个方向之一移动。经过多次试运行，机器人应当学到正确的动作序列，尽可能快地从某一初始状态到达目标状态，并且不会撞到任何障碍物。

使增强学习更困难的一个因素是系统具有不可靠和不完整的感知信息。例如，装备视频照相机的机器人就得不到完整的信息，因此该机器人总是处于部分可观测状态(partially observable state)，并且在决定其动作时应当将这种不确定性考虑在内。例如，机器人可能不知道它在房间的准确位置，而只知道其左边有一道墙。一个任务还可能需要多智能主体(multiple agents)的并行操作，这些智能主体将相互作用并协同操作，以便完成一个共同的目标。机器人足球是这种情况的例子之一。

1.3 注释

进化是形成我们的身体形状和我们内在本能的主要力量。我们还需要终生学习，以改变我们的行为。这有助于我们适应进化论还不能预测的环境变化。在合适的环境中，具有短暂寿命的生物体可能具备它们所有天生的行为能力，而上苍并未赋予我们应对在有限生命中可能遇见的所有状况的能力。但是，进化赋予我们大脑和学习机制，使得我们可以根据经验实现自我更新，从而适应各种环境。当我们在特定情境下学习最好的策略时，知识就存储在我们的大脑里。当情境再现时，当我们再认知("认知"意味认出)情境时，我们就能够回忆起合适的策略并采取相应的动作。

不过，学习有其局限性。就我们大脑的有限容量来说，也许有些东西我们永远都不可能学会，正像我们永远不可能"学会"长出第三只手臂或在脑袋后面长眼睛，即使它们是有用的我们也学不会。注意，与心理学、认知科学以及神经系统科学不同，机器学习的目标并不是理解人类和动物学习的过程，而是像任何工程领域一样，机器学习旨在构建有用的系统。

几乎所有的科学领域都在用模型拟合数据。科学家设计实验、进行观测并收集数据。然后，通过寻找解释所观测数据的简单模型，尝试抽取知识。该过程称为归纳(induc-
14 tion)，它是从一组特别的示例中提取通用规则的过程。

现在，这样的数据分析已经不能依赖人工完成了，原因有二：一是数据量巨大；二是能够做这种分析的人非常短缺且人工分析又很昂贵。因此，对于能够分析数据且自动从中提取信息的计算机模型，也就是说对于学习，人们的兴趣正在不断地增长。

在下面的章节中，我们要讨论的方法源于不同的科学领域。有时，相同的算法会在多个领域中沿着各自不同的历史轨迹被独立地发现。

在统计学中，从特殊观测到一般描述称为推断(inference)，而学习称为估计(estimation)。分类在统计学中称为判别式分析(discriminant analysis)(McLachlan 1992；Hastie，Tibshirani 和 Friedman 2001)。在计算机价格低廉且数量充足以前，统计学家只能处理小样本。作为数学家，统计学家主要使用能够精确分析的简单参数模型。在工程学中，分类称为模式识别(pattern recognition)，方法是非参数的，并且更大程度是凭借经验的(Duda，Hart 和 Stork 2001；Webb 1999)。

机器学习还与人工智能(artificial intelligence)有关(Russell 和 Norvig 1995)，因为智能系统应当能够适应其环境的变化。像视觉、语音和机器人这样的应用领域都是从样本数据中学习。在电子工程领域，信号处理(signal processing)的研究导致自适应计算机视觉和语音程序出现。其中，隐马尔科夫模型(Hidden Markov Model，HMM)的发展对于语音识别尤其重要。

20 世纪 80 年代后期，随着 VLSI 技术的发展和制造包含数千个处理器并行硬件的可能性出现，基于多处理单元的分布式计算理论的可行性使得人工神经网络(artificial neural network)研究领域获得重生(Bishop，1995)。随着时间的推进，人们认识到在神经网络研究领域中，大多数神经网络学习算法都具有统计学的基础(例如，多层感知器就是另一类的非参估计)，因此模拟人脑计算的说法开始逐渐淡出。

近年来，基于核的算法(如支持向量机)日趋流行。借助于使用核函数，支持向量机适用于各种应用，尤其适合生物信息学和自然语言处理方面的应用。如今，人们已经广泛认识到，对于学习而言，好的数据表示至关重要，而核函数是一种引进这种专家知识的好 15
方法。

另一种新方法是使用生成模型(generative model)，它通过一组隐藏因子的相互影响来解释观测数据。一般而言，图模型(graphical model)用来对这些因子和数据的相互影响进行可视化，而贝叶斯形式化机制(Bayesian formalism)使我们既可以定义隐藏因子和模型上的先验信息，又能推导模型的参数。

最近，随着存储和连接费用的降低，在因特网上使用非常大的数据库已经成为可能，再加上廉价的计算，已经使得在大量数据上运行学习算法成为可能。在过去的几十年中，人们一般相信，对于人工智能而言，我们需要新的范型、新的思维、新的计算模型或一些全新的算法。

考虑到机器学习最近在各领域的成功，也许可以说，我们需要的不是新算法，而是大量数据实例和在这些数据上运行算法的充足计算能力。例如，支持向量机源于势函数(potential function)、线性分类和基于最近邻的方法，这些都是 20 世纪 50 或 60 年代提出的，那时，我们只是没有适合这些算法的快速计算机或大型存储器，不能完全展示它们的潜力。可以推测，机器翻译甚至规划这样的任务都可以用这种相对简单的算法来解决，但需要在大量实例数据上训练或通过长时间试错运行。“深度学习”最近取得的成功支持了这种说法。智能看来不像源于某些稀奇古怪的公式，而是源于简单、直截了当的算法的耐心和近乎蛮力的使用。

数据挖掘(data mining)的命名来源于机器学习算法在商界海量数据上的应用(Witten 和 Frank 2011；Han 和 Kamber 2011)。在计算机科学领域中，数据挖掘也称为数据库中知识发现(Knowledge Discovery in Databases，KDD)。

在统计学、模式识别、神经网络、信号处理、控制、人工智能以及数据挖掘等不同领域中，研究工作遵循着各自的途径，并有其各自的侧重点。本书的目标是结合所有这些研

16 究重点，以便给出统一的处理问题方法和建议的解决方案。

1.4 相关资源

机器学习的最新研究成果发表在不同领域的会议和期刊上。机器学习专门的期刊有《Machine Learning》(机器学习)和《Journal of Machine Learning Research》(机器学习研究)。像《Neural Computation》(神经计算)、《Neural Networks》(神经网络)以及《IEEE Transactions on Neural Networksand Learning Systems》(IEEE 神经网络和学习系统汇刊)这样的期刊也发表了有关大量机器学习的论文。统计学方面的期刊，如《Annals of Statistics》(统计学年鉴)和《Journal of the American Statistical Association》(美国统计学会杂志)也会发表一些机器学习方面的文章，并且许多《IEEE Trousactions》，如《Pattern Analysis and Machine Intelligence》(IEEE 模式分析与机器智能汇刊)、《Systems，Man，and Cybernetics》(系统、人和控制论)、《Image Processing》(IEEE 图像处理汇刊)和《Signal Processing》(IEEE 信号处理汇刊)都有一些涉及机器学习的理论和它应用的有趣论文。

关于人工智能、模式识别和信号处理方面的期刊也包含机器学习方面的文章。以数据挖掘为主的期刊有《Data Mining and Knowledge Discovery》(数据挖掘与知识发现)、《IEEE Transactions on Knowledge and Data Engineering》(IEEE 知识与数据工程汇刊)以及《ACM Special Interest Group on Knowledge Discovery and Data Mining Explorations Journal》(ACM 知识发现和数据挖掘特别兴趣组期刊)。

关于机器学习方面的主要会议有"*Neural Information Processing Systems*(NIPS)"、"*Uncertainty in Artificial Intelligence*(UAI)"、"*International Conference on Machine Learning*(ICML)"、"*European Conference on Machine Learning*(ECML)"以及"*Computational Learning Theory*(COLT)"。模式识别、神经网络、人工智能、模糊逻辑和遗传算法方面的会议，以及关于计算机视觉、语音技术、机器人和数据挖掘等应用方面的会议，也会有针对机器学习的专题。

网站 http://www.ics.uci.edu/~mlearn/MLRepository.html 上的 UCI Repasitory 包含大量数据集，致力于机器学习的研究者经常把它们作为性能评价基准。另一个资源是网站 http://lib.stat.cmu.edu 上的 Statlib。此外，还有一些针对特定应用的数据库，例如，针对计算生物学、人脸识别、语音识别等。

新的、更大的数据集不断地添加到这些库中。但是，有些研究者仍然相信这些库的范围有限，不能反映实际数据的全部特征，因此在这些库中的数据集上的准确性并不说明问
17 题。甚至可以说，当反复使用固定库中的数据集并量身打造新算法时，我们正在产生针对这些数据集的一组新的"UCI 算法"。这就像仅通过解决一组实例问题来学习一门课程的学生。正如我们将在后面的章节中所看到的，不同的算法在不同的任务上会好一些，因此最好是针对一种应用，为该应用抽取一个或一些大型数据集，并针对特定的任务，在这些数据集上进行算法比较。

机器学习研究者近期的大多数文章都可以从因特网上找到，大部分作者还在网站上提供了他们的程序和数据。机器学习会议和暑期班上的辅导讲座也多半可以获取。还有一些实现各种机器学习算法的免费工具箱和软件包，其中 http://www.cs.waikato.ac.nz/ml/weak/上的 Weka 特别值得关注。

1.5 习题

1. 设想你有两种选择：可以扫描并传送图像；或者先使用光学字符阅读器(OCR)，然后再传送相应的文本文件。用对比方式讨论这两种方法的优缺点。在什么时候一种方法比另一种方法更可取？
2. 假定我们正在构建一个 OCR，并且对于每一字符，我们都存储该字符的位图作为与逐个像素读取的字符进行匹配的模板。请解释什么时候这样的系统会失败。为什么条码阅读器目前仍在使用？

   解：在这种系统中，每个字符只能有一个模板，并且不能识别来自多种字体的字符。存在 OCR-A 和 OCR-B 这样的标准字体(通常在我们购买的资料包装上看到的字体)，它们与 OCR 软件一起使用(这些字体的字符被稍加改变，以便使得它们之间的相似性最小)。条码阅读器仍然在使用，因为与阅读任意字体、字号和样式的字符相比，它仍然更好(更便宜、更可靠、更可用)。
3. 假定我们的既定目标是构建识别垃圾邮件的系统。请问是垃圾邮件中的什么特征使我们能够确认它为垃圾邮件？计算机如何通过语法分析来发现垃圾邮件？如果发现了垃圾邮件，你希望计算机如何处理它：自动删除？转到另一个文件夹？还是仅仅在屏幕上标亮显示？

   解：通常，基于文本的垃圾邮件过滤器检查邮件中是否有某些词或符号。像“机会”(opportunity)、“伟哥”(viagra)、“美元”(dollar)这样的词，以及像“$”和“!”这样的字符提高了邮件是垃圾邮件的概率。这些概率从用户先前已经标记为垃圾邮件的过去邮件样例的训练集中学习。在后面的章节中，我们会看到许多这样的算法。 18

   垃圾邮件过滤器没有 100%的可靠性，可能在分类时出错。如果有一个垃圾邮件没有被过滤掉，那么不太好，但是总比把好邮件当作垃圾邮件过滤掉好。稍后我们将讨论如何考虑这种假正和假负的相对代价。因此，不应该自动删除系统认为是垃圾邮件的信息，而是应该把它们放在一旁，使得如果用户愿意的话用户可以看到它们，特别是在使用垃圾邮件过滤器的早期阶段，系统训练尚不充分时尤其如此。垃圾邮件过滤可能是机器学习的最好应用领域之一，学习系统可以自动地适应垃圾邮件信息产生方式的变化。
4. 假设给定的任务是制造自动出租车，请定义约束。输入是什么？输出是什么？如何与乘客沟通？需要与其他的自动出租车沟通，即需要某种语言吗？
5. 在购物篮分析中，我们希望找出产品 X 和 Y 二者之间的依赖关系。对于给定的顾客交易数据库，如何能够发现这些数据之间的依赖关系？如何将依赖关系发现算法推广到多于两个的产品之间？
6. 在你的日报中，为政治、体育和艺术类各找出 5 个新闻报道样例。阅读这些报道，找出每类报道频繁使用的词，这些词可能帮助我们区别不同的类别。例如，政治方面的新闻报道多半会包含“政府”、“经济衰退”、“国会”等词，而在艺术类的新闻报道中可能包括“专辑”、“油画”或“剧院”等词。还有一些词(如“目标”)是模棱两可的。
7. 如果人脸图像是 100×100 的图像，按行写出，则它是一个 10 000 维向量。如果我们把图像向右移动一个像素，则将得到 10 000 维空间中的一个很不相同的向量。如何构造一个对于这种扰动具有鲁棒性人脸识别器？

   解：通常，人脸识别系统都有一个用于输入标准化的预处理阶段，在识别之前，