第一部分

绪　　论

第 1 章
统计机器学习

计算机与网络的近期发展使得我们能够即刻访问大量信息，如文字、声音、图像与影像。此外，日志、消费记录和病历等广泛的个人数据也日复一日地累积起来。如此大量的数据被称为*大数据*(big data)。并且，存在一股通过从数据中抽取有用的知识来创造新价值与商机的增长趋势。这一过程通常被称为*数据挖掘*(data mining)。*机器学习*是用于抽取有用知识的关键技术。

1.1 学习的类型

依据可用数据的类型，机器学习能够分为*监督学习*(supervised learning)、*非监督学习*(unsupervised learning)和*强化学习*(reinforcement learning)。

监督学习可谓机器学习最基本的类型。它被视为一种学生学习的过程，即向导师提问并回答。在机器学习情境中，学生对应于计算机，导师对应于计算机的用户；计算机从问与答的成对样本中学习一种从问题到其答案的映射。监督学习的目标在于获得*泛化能力*(generalization ability)。其指的是一种能够为从未被学习过的问题猜出恰当答案的能力。因此，用户不必再将每件事情都教给计算机，而计算机仅通过学习一小部分知识就能够自动地应对未知情况。监督学习已经被成功地应用到广泛的真实问题中，如：手写字母识别、语音识别、图像识别、垃圾邮件过滤、信息检索、在线广告、推荐系统、脑电波分析、基因分析、股票价格预测、天气预报和天文数据分析。监督学习的问题也会被具体称为*回归*(regression)，当答案是一个实数值(如：温度)；*分类*(classification)，如果答案是一个分类值(如：“是”或“否”)以及*排序*(ranking)，如果答案是一个数列值(如：“好”、“中”或“劣”)。

非监督学习被认为是，导师不存在并且学生自学。在机器学习情境中，计算机通过互联网自动地收集数据并且尝试在没有用户任何指导下抽取有用的知识。因此，非监督学习比监督学习更加自动化，尽管其目标不一定指定清楚。非监督学习的典型任务包括*数据聚类*(data clustering)和*异常点检测*(outlier detection)。同时，这些非监督学习技术已经就广泛的真实问题取得巨大成功，如系统诊断、安全、事件检测和社交网络分析。非监督学习也通常被用作监督学习的预处理过程。

*强化学习*与监督学习类似，也是以使计算机获得对没有学习过的问题做出正确解答的泛化能力为目标，但是在学习过程中，不设置导师提示对错、告知最终答案的环节。相

反，导师评价(evaluate)学生的行为并给予其反馈。强化学习的目标是基于来自导师的反馈，使得学生提高其行为，从而最大化导师的评价。强化学习是一个人与机器人的行为的重要的模型。它已经被广泛地应用于诸多领域，如：自主机器人控制、电脑游戏和营销策略优化。在强化学习背后，监督学习与非监督学习的方法，如：回归、分类与聚类，都常被使用。

本书重点在于监督学习与非监督学习。对于强化学习，请参考文献[99，105]。

1.2　机器学习任务举例

在本节，会详尽介绍不同的监督学习与非监督学习任务。

1.2.1　监督学习

回归(regression)的目标是为了从其样本近似估计一个实数值函数(如图 1.1 所示)。让我们不妨定义输入为 d 维实数向量 $\boldsymbol{x}$，输出为实数标量 y，学习目标函数为 $y=f(\boldsymbol{x})$。这个学习目标函数 f 被认为是未知的，但是其输入输出样本对$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$是可以被观测的。在实践中，被观测输出值 y_i 可能被某些误差ϵ_i 所污染，即

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i$$

在这个设定中，x_i 对应学生向导师提出的一个问题；y_i 对应导师给学生的回答。噪声ϵ_i 或许对应导师的口误或者学生的误解。学习目标函数 f 对应导师的知识，这些知识使得他(她)可以回答任何问题。回归的目标在于让学生学习这个函数，学生由此也可以回答任何问题。泛化的水平能够通过真实函数 f 与其近似 $\hat{f}$ 的接近程度来测量。

另一个方面，以监督学习方式，分类则是一个模式识别(pattern recognition)问题(如图 1.2 所示)。不妨用 d 维向量 $\hat{\boldsymbol{x}}$ 来定义输入模式和由标量 $y\in\{1, \cdots, c\}$定义其类别，其中 c 定义类别数目。为了训练分类器，输入-输出对样本$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$以与回归一样的方式提供。倘若真实分类规则被定义为 $y=f(\boldsymbol{x})$，那么分类也能够被视为函数逼近问题。然而，在回归情境下，一个本质的不同在于在 y 中不存在接近程度这一概念 y：$y=2$ 比 $y=1$更接近 $y=3$，但是“y 和 y‘是否相同’”是分类唯一关心的。

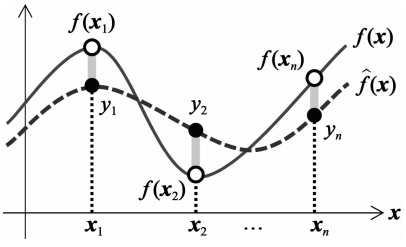

图 1.1　回归

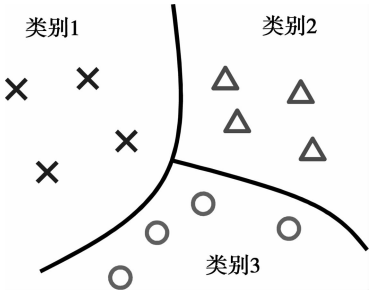

图 1.2　分类

在监督学习中，排序(ranking)问题旨在学习样本 $\boldsymbol{x}$ 的次序 y。因为次序有顺序，如：$1<2<3$，排序更像是回归而非分类。为此，排序问题也可以被视为有序回归(ordinal regression)。然而，不同于回归，提出输出值 y 是没有必要被预测的，但也仅仅是其相对

值是需要的。例如，假设三个实例的“值”是 1、2 和 3。那么，既然在排序问题中只有顺序关系 1<2<3 是重要的，预测值，如同 2<4<9 也是一个完美答案。

1.2.2 非监督学习

聚类是分类的一个非监督的副本(如图 1.3 所示)。其目标在于在没有任何监督 $\{(y_i)\}_{i=1}^n$ 5 下，将输入样本 $\{\boldsymbol{x}_i\}_{i=1}^n$ 归类到簇 1，2，…，c 中。通常地，类似样本被认为属于同一个簇；不同样本被认为属于不同的簇。因此，如何测量样本间的相似度成为了聚类的关键问题。

离群点检测也被称为异常检测(anomaly detection)。其目标在于从给定数据集 $\{\boldsymbol{x}_i\}_{i=1}^n$ 中找出不正常的样本。以如聚类同样的方式，在离群点检测中，样本间相似度的定义扮演着中枢角色，因为那些不同于其他的样本的点通常被认为是离群点(如图 1.4 所示)。

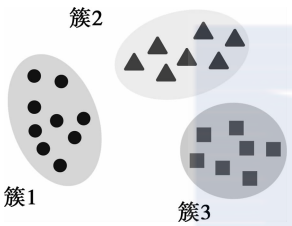

图 1.3 聚类

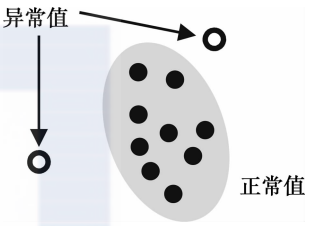

图 1.4 离群点检测

变化检测(change detection)也被称为新奇检测(novelty detection)。其目标在于判断新的给定点集 $\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ 是否具有原数据集 $\{\boldsymbol{x}_i\}_{i=1}^n$ 同样的属性。在变化检测中需要有样本间相似度时，类似于离群点检测中样本间的相似度就被使用了。当 $n'=1$(也就是，只有一个单个点被提供在变化检测中)，变化检测问题可以被缩减为离群点检测。

1.2.3 进一步的主题

除了监督学习与非监督学习，机器学习中有许多有用的技术。

输入-输出对的样本 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ 被用于训练监督学习；相反，只有输入 $\{\boldsymbol{x}_i\}_{i=1}^n$ 的样本被用于非监督学习。在许多监督学习技术中，收集仅有输入的样本 $\{\boldsymbol{x}_i\}_{i=1}^n$ 或许容易，但是获取 $\{\boldsymbol{x}_i\}_{i=1}^n$ 的输出样本 $\{y_i\}_{i=1}^n$ 是费力的。在这样的情况下，输出样本或许只能收集到 m 个($\ll n$ 个输入样本)；余下 $n-m$ 个样本只是输入样本。半监督学习(semisupervised learning)旨在从输入-输出对样本 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ 和仅输入样本 $\{\boldsymbol{x}_i\}_{i=m+1}^n$ 中学习。经典地，半监督学习方法提取诸如来自仅输入样本 $\{\boldsymbol{x}_i\}_{i=m+1}^n$ 的簇结构的分布信息，并使用这些信息提升针对输入-输出对样本 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ 6 的监督学习。

在弱学习算法表现仅微弱优于随机猜想的情况下，集成学习(ensemble learning)旨在通过联合如上弱学习算法来构建强学习算法。最受欢迎的方法之一是通过弱学习算法的投票表决来补充彼此的弱点。

标准学习算法考虑列数据 $\boldsymbol{x}$。然而，如果数据具有二维结构(如：图像)，直接从矩阵数据中学习比向量化矩阵数据更加有效。矩阵学习(matrix learning)或张量学习(tensor learning)针对高维数据处理。

当数据样本一个一个顺序地提供时，更新学习结果以包含新数据相比于从零开始重新

学习会更自然。在线学习(online learning)旨在有效地处理这样的给定序列数据。

在解决学习任务时，从其他相似学习任务中迁移知识是有帮助的。这样的范式被称为迁移学习(transfer learning)或邻域适应(domain adaptation)。如果多相关学习任务需要被解决，同步比起分别解决它们或许更加有效。这个想法被称为多任务学习(multitask learning)。其可以被视作迁移学习的双向变异。

从高维数据中学习是具有挑战性的。其通常被称作维度灾难(the curse of dimensionality)。降维(dimensionality reduction)的目的是从高维数据样本$\{\boldsymbol{x}_i\}_{i=1}^n$抽取本质信息并获取它们的低维表达$\{\boldsymbol{z}_i\}_{i=1}^n$(图 1.5)。在线性降维中，低维表达$\{\boldsymbol{z}_i\}_{i=1}^n$可以利用一个大矩阵(a fat matrix)$\boldsymbol{T}$通过$\boldsymbol{z}_i=\boldsymbol{T}\boldsymbol{x}_i$获得。监督降维尝试找到低维表达$\{\boldsymbol{z}_i\}_{i=1}^n$作为预处理，以便随后的监督学习任务能够轻松地被解决。另一方面，非监督降维尝试找到低维表达$\{\boldsymbol{z}_i\}_{i=1}^n$使得原始数据的某些结构被保持，例如，可视化目的。度量学习(metric learning)类似于降维，但是它更加强调高维空间而非降低数据的维度的度量的学习。

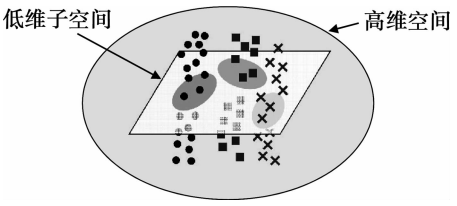

图 1.5　降维

1.3　本书结构

本书的主要内容包括以下四个部分。

第二部分介绍概率与统计的基本概念。这些将会在随后章节中广泛地使用。那些熟悉概率与统计或者想要学习机器学习的读者可以忽略第二部分。

基于概率与统计的基本概念，第三、四和五章介绍多种机器学习技术。这几部分相对独立。因此，读者可以依据个人兴趣从任何章节开始学习。

第三部分旨在统计模式识别的生成方法。生成方法的基本思想是数据的概率分布被建模以执行模式识别。当数据生成机制的先验知识是可以获取的，生成方法就是高度有用的。多种基于频率论和贝叶斯框架的生成模型估计的分类和聚类技术将会被介绍。

第四部分集中在针对统计机器学习的判别方法。判别方法的基本思想是在不对数据生成概率分布建模的情况下，直接地解决目标机器学习任务，如：回归和分类。当不存在用于数据生成机制的先验知识的情况下，判别方法能够比生成方法更加有前景。除了统计与概率，优化理论(optimization theory)的知识也在判别方法中扮演着重要角色。这也会被提及。

第五部分致力于介绍机器学习中的高级问题，包括：集成学习(ensemble learning)、在线学习(online learning)、预测置信度(confidence of prediction)、半监督学习(semisupervised learning)、迁移学习(transfer learning)、多任务学习(multitask learning)、降维(dimensionality reduction)、聚类(clustering)、离群点检测(outlier detection)和变化检测(change detection)。

本书提供了针对第三、四和五章介绍方法的简洁 MATLAB 代码。读者可以及时测试算法并学习其数值行为。