

第 1 章

绪论

本章先给出机器学习的定义；然后介绍机器学习的两种主要类型：监督学习和无监督学习；其次给出机器学习在日常生活及不同专业中的应用；最后简单介绍网上公开的部分机器学习数据库。

1.1 机器学习的定义

机器学习 (machine learning) 研究计算机如何模拟或实现人类的学习行为，以获取新的知识或技能，使之不断改善自身的性能。它是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。在某种程度上，机器学习的工作方式与人类学习类似。例如，如果向孩子显示带有特定对象的图像（如小汽车、小动物等），孩子可以学习识别和区分它们。机器学习以相同的方式工作：通过数据输入和某些命令，计算机（机器人）被告知特定对象是小汽车而另一个对象不是小汽车，通过反复训练，计算机能够“学习”识别某些对象（如小汽车等）并最终能够清楚地区分“小汽车”和“非小汽车”。

1.2 机器学习的种类

机器学习的种类有多种，主要有监督学习 (supervised learning)、无监督学习 (unsupervised learning) 和强化学习 (reinforcement learning)。

1.2.1 监督学习

监督学习有预先标记好的样本数据（通常称为训练样本）。如有 N 个训练样本

$$D: \mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^n\}_{i=1}^N, \mathbf{Y} = \{y_i \in \mathbb{R}\}_{i=1}^N \quad (1-1)$$

其中 $\mathbf{X}$ 代表输入， $\mathbf{Y}$ 代表相应的输出或标签。每个输入变量有 n 个特征，那么每个输入变

量是 n 维向量 $\mathbf{x}_i = [x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)}]^T$, 称为特征向量。

计算机基于这些给定的输入和输出被训练和进行学习, 并最终建立起输入和输出之间的关系。监督学习的目的是利用建立的输入和输出的关系, 对没在训练样本中的新样本预测出相应的输出或标签 y 。如果每一个特征向量对应的标签是连续的实数 $y \in \mathbb{R}$, 这种学习称为回归问题。如果每一个特征向量对应的标签是离散的, $y \in \mathbf{L}$, $\mathbf{L} = \{l_1, l_2, \dots, l_k\}$, k 是大于 2 的整数, 这种学习称为分类问题, 即监督学习主要用于解决回归问题和分类问题。

(1) 回归问题

例如, 我们有 30 个样本数据, 输入特征变量 x 是每周学生花在某一课程的小时数, y 是对应的输出变量, 是该学生的分数, 如表 1-1 所示。

表 1-1 30 个样本的小时数和分数

序号	小时数	分数	序号	小时数	分数	序号	小时数	分数
1	15	68	11	36	93	21	26	74
2	32	88	12	23	79	22	34	87
3	38	92	13	20	75	23	11	55
4	28	82	14	29	80	24	31	85
5	14	58	15	17	59	25	22	74
6	24	72	16	40	96	26	21	73
7	25	74	17	12	57	27	37	94
8	33	85	18	16	69	28	35	88
9	10	52	19	27	76	29	19	70
10	30	82	20	39	95	30	18	71

回归问题就是根据 30 个训练样本数据, 建立输出变量学生分数 y 与每周学生花在课程的小时数 x 之间的函数关系, 如公式 (1-2) 所示。

$$y = f(x) \quad (1-2)$$

然后利用建立的这个关系, 可以预测新的输入对应的分数。

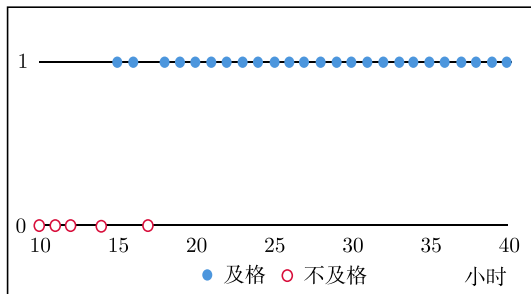
(2) 分类问题

分类问题分为二分类问题和多分类问题。

① 二分类问题

在表 1-1 中, 学生成绩可以分为两类: 及格和不及格, 如图 1-1 所示。

二分类问题就是根据 30 个训练样本数据, 建立输出 y (y 值只有两个离散值 0 和 1, 0 代表不及格; 1 代表及格) 与每周学生花在课程的小时数 x 之间的关系。根据建立的关系, 可以预测相应新的输入 (每周学生花在课程上的小时数) 对应的学生是否及格。 y 值是 0 代表不及格, 1 代表及格。

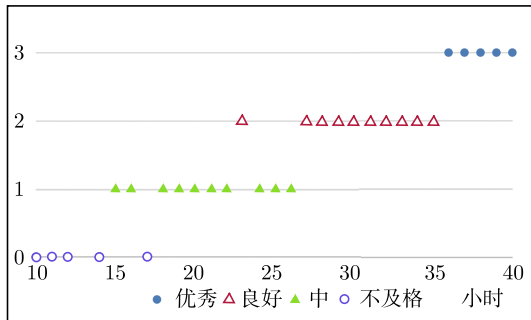


扫码看彩图

图 1-1 二分类示意图

② 多分类问题

在表 1-1 中，学生成绩可以分为 4 类：不及格、中、良好、优秀，如图 1-2 所示。



扫码看彩图

图 1-2 多分类示意图

多分类问题就是根据 30 个训练样本数据，建立输出 y (y 值只有 4 个离散值 0、1、2、3；0 代表不及格，1 代表中，2 代表良好，3 代表优秀) 与每周学生花在课程上的小时数 x 之间的关系。根据建立的关系，可以预测新的输入（每周学生花在课程的小时数）对应学生分数的等级，如果 y 值是 0 代表不及格；1 代表中；2 代表良好；3 代表优秀。

可见回归预测的输出是连续值，分类的预测输出是离散值（类的标签）。

1.2.2 无监督学习

无监督学习是从没有标签的样本数据中学习模型。有 N 个训练样本的无标签训练集可用符号表示为 $D: \mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^n\}_{i=1}^N$ 。

与监督学习相比，无监督学习定义不太明确。无监督学习主要用于聚类处理，即根据样本数据的特征或者属性对输入的数据进行分组。下面给出无监督学习的三个例子。

例子 1：如图 1-3 所示，无监督学习使计算机对不同颜色的气球，根据“颜色”特征，将相同颜色的气球放在一起，即聚为 2 类，聚类结果如图 1-3(b) 所示。

例子 2：一群毕业生，其中包含了本科生、硕士研究生和博士研究生三类。无监督学习使计算机把相同类型的毕业生放在一起，即聚为 3 类。

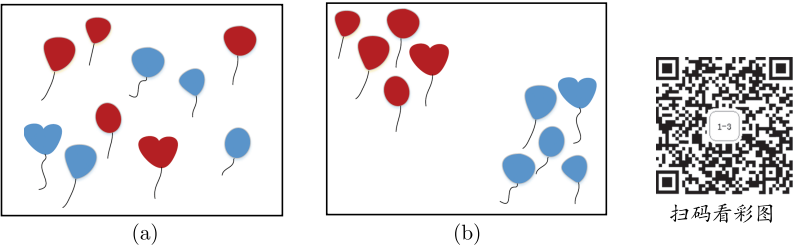
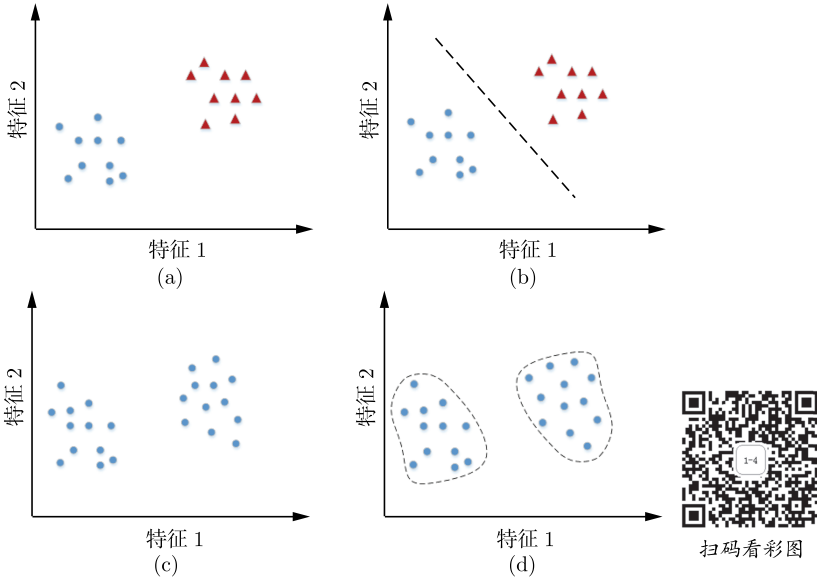


图 1-3 聚类例子

例子 3: 报纸中有许多新闻报道, 无监督学习使计算机把相同类型的新闻内容放在一起, 如按照新闻内容可以分为经济报道、运动报道和娱乐报道 3 类, 即聚为 3 类。

图 1-4 显示了监督学习和无监督学习的区别, 其中 (a) 和 (b) 显示的是监督学习, 即 (a) 显示两种标签数据集, 经过监督学习, 两类的划分边界可用 (b) 中的虚线表示; (c) 和 (d) 显示的是无监督学习, 即 (c) 显示无标签数据集, 经过无监督学习, 可得到如 (d) 所示的聚类结果。



(a)(b) 监督学习, (c)(d) 无监督学习 (聚类)

图 1-4 监督学习和无监督学习的区别

1.2.3 强化学习

强化学习用于解决决策问题(通常是一系列决策), 例如机器人感知和移动、自动棋手和自动驾驶。强化学习不在本书学习范畴内, 读者可以参考文献 [1] 以了解更多相关内容。

除了这三种类型之外, 第四种类型的机器学习是半监督学习。半监督学习最近引起了越来越多的关注。它定义在有监督学习和无监督学习之间, 包含有标记和无标记的数据, 并从

中共同学习知识。

本书仅限于研究两种最基本机器学习：监督学习和无监督学习。

1.3 机器学习的应用

毫无疑问，机器学习可以帮助人们更有创造性和更有效地工作，人们可以通过机器学习将相当复杂或单调的工作委派给计算机完成。当前，机器学习已经渗透到人们生活、科学研究和工程实践的各个方面，可以说，当今我们被基于机器学习的技术所包围。

1.3.1 在生活中的应用

机器学习已经广泛应用到我们日常生活的方方面面，给我们的生活带来了极大的便利，让我们的生活变得更加美好。以下是机器学习在我们生活中的应用。

(1) 垃圾邮件过滤

电子邮件已经是人们信息交流的最主要方式。机器学习算法（通常采用决策树归纳法）通过分析大量垃圾邮件当中的词汇来学会判断一封邮件是否为垃圾邮件，以此来实现对垃圾邮件的筛选，为我们营造一个干净有效的邮箱环境，保护我们免受垃圾邮件的骚扰。

(2) 银行卡在线交易检测

在互联网时代，网银在线支付因为具有较高的安全性、交易的实时性和交易记录的真实性变得越来越流行。银行为提高银行卡在线交易，采用机器学习算法，通过分析以往的欺诈交易，能够学会判断一笔交易是否有诈骗的可能，从而提醒信用卡用户谨慎操作。

(3) 网络安全检测

机器学习已经广泛地应用于网络安全检测中。如机器学习算法可以用于恶意域名的检测。人们根据恶意域名与正常域名在字符组成、生成方法、解析过程等方面的差异，设计域名的字符统计特征、相似度特征、解析特征，并结合机器学习算法提出了基于字符及解析特征的恶意域名检测方法。该方法检测准确率高，能够保障网络安全。此外随着网络的普及、网页的数量飞速增长，混杂其中的恶意网页占据的比例也呈上升趋势。恶意网页的检测一直是网络安全领域的研究重点和难点，传统的恶意网页检测模型在新形势下的表现不尽如人意。机器学习算法在恶意网页领域的应用是突破传统恶意网页检测局限的一种途径。

(4) 产品推荐

网上购物因具有可以在家“逛商店”，订货不受时间、地点的限制，同时还能获得较大量的商品信息等优点，广受大众百姓的喜爱。比如你在网上买了一个产品，那么你可能已经注意到，购物网站机器学习根据你在网站的行为，如过去购买的商品、收藏或添加到购物车中的商品等，向你推荐了一些与你的偏好相匹配的其他商品。

(5) 搜索引擎结果优化

浏览器中的搜索引擎使用机器学习算法使计算机学会根据相关度排列网络信息，从而为用户提供优质的搜索结果。不仅如此，浏览页面中的广告、购物网站的物品推荐也是根据用户点击情况的有意安排。比如你每次执行搜索时，后端的算法都会监视你对结果的响应。如果你打开顶部的链接并在网页上停留很长时间，搜索引擎会假定它显示的结果与查询相符。类似地，如果你到达搜索结果的第二页或第三页，但没有打开任何结果，搜索引擎会估计所提供的结果不符合要求。这样，在后端工作的算法可以改善搜索结果。

(6) 视频监控

与文本文档和其他媒体文件（如音频、图像）相比，视频文件包含更多信息。因此，从视频中提取有用的信息，即自动视频监控系统成为一个研究热点。机器学习已经广泛地应用于视频监控，比如视频监控系统通过机器学习能够跟踪和识别人们的异常行为，如长时间不动、绊倒或在长凳上打盹等。因此，该系统可以向人工服务人员发出警报，从而帮助及时发现危险。此外通过机器学习算法训练网络，实现对航拍视频中的烟火自动识别，对预防和扑救森林火灾、减少损失有着重要意义。

(7) 通勤时的预测

我们在出行时常常使用导航系统。当我们使用导航系统时，我们当前的位置和速度被保存在一个中央服务器上，然后中央服务器使用这些数据构建当前交通的地图，机器学习根据日常经验估计可能出现拥堵的区域。我们使用在线交通网络预订出租车时，系统会应用机器学习方法预估车价。

(8) 其他方面

还有些机器学习应用极大地方便了我们的生活，如文字识别广泛应用于邮政编码、统计报表、财务报表、银行票据等；在智能图书馆系统中，机器学习算法结合用户画像对用户和资源实现聚类，然后在知识组织体系内进行分配，提供动态的、个性化的精准服务；基于语音识别的各类智能家居使我们的生活更加便利；基于机器学习的自然语言处理技术构成了强大的翻译软件，帮助我们突破语言的屏障；等等。

1.3.2 在不同学科中应用

目前几乎所有学科（工科、理科、金融和文法等）的研究方向都与机器学习相关，机器学习已经像高等数学一样成为从事不同学科研究的有力工具。下面简单介绍机器学习在一些学科中的应用。

在信息与通信领域，目前主要包括机器学习对于图像、音频、视频等方面的信息处理以及无线通信方面的应用。基于机器学习的边缘检测算法，在目标轮廓分割、轮廓识别等工作中取得了比传统方法更好的效果。机器学习能更快地进行图像分类、修复、压缩编码等。在音频方面，目前主流的方式便是运用机器学习方法对声音进行分类识别。在视频方面，机器学习算法主要应用在短视频的处理，对视频信息的内容与视频发布初期的信息进行提取，采

用分类或者回归的算法进行流行度预测。而越来越多的深度学习神经网络应用到大视频的处理中。机器学习在无线通信中的应用主要体现在调制模式识别、信道编译码、信道估计与信号检测、MIMO（多输入多输出）检测以及图像信号处理等方面。随着近年来人工智能的发展以及 5G 乃至 6G 无线通信的到来，无线通信网络必将智能化，机器学习与人工智能技术在无线通信领域的应用也将更加广泛。

机器学习在电力系统的复杂问题处理中得到了广泛的应用，其主要体现在三个方面，即电力系统的数据预测、电力系统运营管理以及系统的运行安全维护。电力系统的数据预测主要涵盖负荷预测、分布式能源发电预测、系统运行电压预测。在电力系统的运营管理方面，人工神经网络等算法被应用到电压控制当中，形成了较为完整的智能电网控制系统。而机器学习在电力系统运行安全维护方面主要是体现在系统的故障诊断和保护装置的设计与应用之中。此外，机器学习在电气工程领域的应用还有了许多细小分支，包括其在综合能源系统中的应用、在电力系统继电保护中的应用、在发电功率预测中的应用、在电力系统稳定安全评估中的应用等。因此机器学习在电气工程、电力系统中的大小各方面都扮演着重要的角色。

在集成电路工程与微电子学领域，神经网络算法在多种微波器件和电路建模中得到应用。神经网络算法提高了模型精度，节省了建模时间。在基于侧信道信息的硬件木马检测方法中，机器学习算法有着非常广泛的应用，例如遗传算法、最邻近算法、支持向量机算法等都发挥了重大作用。随着半导体制造工艺的发展以及电路规模的爆炸式的增长，传统方法不再适用，而机器学习在集成电路优化、故障检测、芯片测试、集成电路设计、热点检测技术等方面提供了很大的帮助。随着集成电路技术的发展，集成电路的复杂程度不断增大，从设计到制造，再到封装测试以及后续的芯片使用过程中，有大量数据需要进行合理分析和有效处理，机器学习可以更加高效、迅速地解决其中一些问题。

机器学习在建筑与土木工程领域内的应用近年来势头迅猛，较为广泛使用的算法包括专家系统、支持向量机、人工神经网络等相关算法，主要用于结构分析、设计选型、建造、结构安全、损伤评估以及工程管理与运营维护等领域。例如，利用机器学习的分类算法实现对施工现场工人的智能识别，识别在特定工况下的工人是否有危险行为；训练图像分类和识别模型，实现实时判断结构类型和检测其主要构件；根据对目标的识别定位以及轨迹追踪捕捉到工人与施工构件之间的空间关系；利用逻辑回归根据位置高度等特征值输出应力应变状态，以判断高耸建筑物等结构设计安全与否。除此之外，机器学习在结构健康监测领域发挥着重要角色，针对结构智能检测与诊断系统，可以借助机器学习算法对大量的监测数据进行处理和分析，如混凝土裂缝的检测识别。机器学习在建筑与土木工程中已发挥着不可替代的作用，从各方面保障了应有的工程效率和安全性能。

在机械工程专业和仪器科学与技术专业，机器学习被应用于精密加工中有限元仿真分析和热误差补偿与处理等方面，通过机器学习技术可以使加工更加高效。在轴承故障诊断中，机器学习算法可以对滚动轴承故障的模式进行分类，并对不同运行时刻的轴承退化程度

做评估。在焊接领域，机器学习算法在焊接路径规划、焊缝缺陷识别、焊接性能预测、工艺参数优化等诸多方面有广泛的应用。传感器和机器学习的结合也有诸多应用，例如自动驾驶汽车、利用大数据-人工智能技术提高铁路探伤仪器的伤损检出率、让机器进行学习并使其具备自动检索功能、采用循环神经网络对卫星故障进行检测等。机器学习还应用于仪器故障检测、构造软测量仪表和进行具体测量上。目前，仪器仪表正从自动化向智能化方向发展，今后机器学习的应用将会更加广泛。

在医学方面，机器学习也有着重要的角色。主要应用于在人脑和计算机或者其他机电设备之间建立直接信息交流途径的脑-机接口技术以及基于医学图像的各种疾病的预防、诊断和治疗。在脑-机接口技术中，利用机器学习技术，将脑部传来的并经过预处理的信号模式产生控制信号，实现脑机之间的交互。随着医疗技术手段的提高，各种超声成像、病理切片等成像技术的发展，基于医学图像处理的机器学习方法不断兴起。机器学习可应用到对病理图片的病变区域的分割、提取以及病变类型的诊断分类中。目前在肺结节、脑部、心脏、眼部视网膜、各类癌症等领域有着良好的发展前景。特别是在癌症的预防、诊断和治疗中，机器学习发挥着巨大作用。作为人类死亡的重要原因之一，癌症的高效诊疗具有深远的意义，而深度学习网络作为机器学习学科的分支正在进一步简化诊断步骤和提高诊断准确率。医学图像识别可以提高识别准确率，减轻医生工作量，降低医疗成本，节省医疗资源。

在海洋技术领域，深度学习可以应用于海浪高度的预测，使其识别精度有显著的提升。传统的海洋数据处理分析往往受主管因素影响，而深度学习以数据为驱动，通过多层学习提取数据中有用的信息，能够提高数据处理效率和精度，深度学习逐渐应用于海洋大数据的分类识别、高分辨率重构、现象预测等方面，为海洋大数据的智能分析挖掘带来新的契机。在船舶与海洋工程领域，机器学习多用于航行器控制、船舶图像识别、航线决策等方面。在海洋预报领域，机器学习算法在业务化预报和数据同化中取得了不错的效果。随着水下导航定位技术的不断发展，基于当前水下定位与导航的传统手段和优缺点，引入多种机器学习算法可以优化各种定位与导航技术的结果，消除误差，提高精度。水声信道特性频率依赖有限的带宽，漫长的时变多径传播和严重的多普勒效应，对水下通信构成巨大的挑战，利用机器学习及其分支，可以较好地将各种算法应用到调制及其解调等具体模式中。与传统的方法相比，机器学习算法在可靠性及有效性方面都有显著的提高。机器学习的发展为海洋技术未来的发展开启了新的方向。

在化学化工领域中，机器学习的理论和方法已被广泛应用于解决工程应用和科学领域的复杂问题，例如基于机器学习的中药化学成分作用靶点的识别研究、机器学习算法在药物专利分类中的应用研究、Boosting 算法在化学数据挖掘中的应用等。在传统化学研究的基础上，利用机器学习算法可以帮助我们解决一些难处理的问题，同时也推进了整个化学领域的进步。同时，机器学习广泛应用于药物晶体组成和性能等方面的预测，包括结晶度预测、结晶溶剂选择、多晶形成、共晶形成、理化性质预测等。通过对大量的已知实验数据在相对短的时间内进行数据分析、建模和训练，来预测并辅助研究固态组成和晶体性质。

对于光学而言,机器学习的一个应用方向是在光学陀螺温度误差建模与补偿中的应用。对于中精度陀螺,机理上消除温度影响成本高、难度大,温度误差建模与补偿是解决问题的理想方法,常见的建模方法有多项式模型、马尔科夫链模型、神经网络模型、支持向量机等。光学陀螺温度误差的补偿如果能利用机器学习的算法恰当解决,可以节省大量成本,拥有广阔的应用前景。

在建筑学科,机器学习算法被广泛应用于科学选址以及节能环保设计中。随着城市的快速发展,原有的格局逐渐被打破,不同区域、不同密度的人口对城市的发展带来一定的影响,利用人口密度、风险评估等大数据的优势并借助数据挖掘等机器学习算法,挖掘出未来人口流动的方向,这可用于指导对建筑物的科学选址。此外,通过采用神经网络等机器学习算法,可对室内的灯光、空调和通风系统进行更节能更环保的设计。

在经济学科,机器学习算法被应用于预测金融市场的未来走势以及交易的风险,比如股票、公司的销售额、客户满意度的预测,以及帮助金融公司及时发现不久的将来是否存在客户迁移的威胁等。常用的机器学习算法包括特征提取、数据降维、判别分析以及人工神经网络模型等。基于机器学习方法的金融公司不仅能够提高客户满意度,还能降低成本,较全面地评估客户的愿望和需求,提高客户忠诚度。因此机器学习算法可谓是未来金融领域的“制胜法宝”,其凭借可靠性与有效性为金融领域未来的发展增添了活力。

1.4 网上公开的部分机器学习数据库

网址 <https://guides.library.cmu.edu/machine-learning/datasets> 介绍了以下几种常用的部分机器学习数据库。

- **UCI Machine Learning Repository**

机器学习领域用于对机器学习算法进行验证分析的公开数据库。这是网络上最早的数据源之一,并且已被全世界的学生、学者和研究人员广泛用作机器学习数据集的主要来源。如果你想要寻找有趣的数据集,这个站点值得优先访问,尽管这些数据集是用户提供的,具有不同程度的清洁度。你可以直接从该数据库下载数据,无须注册。

- **WordNet**

一个大型的英语词汇数据库。名词(nouns)、动词(verbs)、形容词(adjectives)和副词(adverbs)被分为多组同义词(synsets),每组表达不同的概念。同义词集之间通过概念语义和词汇关系相互关联。该数据库已经成为计算语言学和自然语言处理的有用工具,并且可以免费下载。

- **ImageNet**

用于计算机视觉目标检测研究的大型实例图像数据库。ImageNet 中包含 2 万多个对象类别,例如气球、鸟等。超过 1400 万个的高清图像被手动标注,以指示图片中的物体。在至少一百万个图像中,ImageNet 还标注了图片中主要物体的定位边框。自 2010 年以来,ImageNet

项目每年举办一次计算机视觉比赛，即 ImageNet 大规模视觉识别挑战赛 (ILSVRC)，软件程序竞相正确分类和检测物体和场景。2012 年卷积神经网络在解决 ImageNet 挑战方面取得了巨大的突破，被广泛认为是深度学习革命的开始。

- **MS COCO**

MS COCO 的全称是 Microsoft Common Objects in Context，其前身是微软于 2014 年出资标注的 Microsoft COCO 数据集。与 ImageNet 竞赛一样，MS COCO 竞赛被视为计算机视觉领域最受关注和最权威的比赛之一。COCO 数据集是一个大型的、丰富的物体检测、分割和字幕数据集，包括 91 个类别，超过 30 万张图像和超过 200 万个标注。

- **MNIST**

MNIST (Mixed National Institute of Standards and Technology database)，是一个非常简单的机器学习视觉数据集，由几万张 28×28 像素的手写数字图片集组成，只包含图片的灰度值信息，用于图像分类。

- **Kaggle**

这是一个主要为开发商和数据科学家提供举办机器学习竞赛、托管数据库、编写和分享代码的平台，内有各种有趣的数据集。

- **StateOfTheArt.ai**

一个完全由社区驱动的网站，提供任务、数据集、指标和结果。

- **Papers With Code**

这是一个免费开放的提供机器学习论文及其公开代码的资源网站。

- **NLP-Progress**

这是一个用于跟踪自然语言处理 (Natural Language Processing, NLP) 发展进度的数据库，包括了数据集和 NLP 任务的最新技术。

- **Berkeley DeepDrive BDD100k**

这是目前最大的自动驾驶 AI 数据集。该数据集有超过 10 万个在一天中不同时段以及在不同天气条件下拍摄的共 1100 多个小时的驾驶体验的视频。这些带注释的图像来自美国纽约和旧金山地区。

参 考 文 献

Kaelbling L P, Littman M L, Moore A W. Reinforcement Learning: A Survey[J]. J Artificial Intelligence Research, 1996, 4(1): 237-285.

习 题

在“1.3 机器学习的应用”中，请找出：哪些是分类问题？哪些是回归问题？哪些是聚类问题？

第一部分

机器学习中的数学基础

- 第 2 章 线性代数
- 第 3 章 概率统计与信息论
- 第 4 章 最优化方法
- 第 5 章 张量分析

第 2 章

线性代数

线性代数作为数学的一个分支，广泛应用于科学和工程中。特别是其中矩阵内容是描述和求解线性方程组等问题的最基本和有用的数学工具。掌握好线性代数对于理解和从事机器学习算法相关工作是很重要的。因此在开始学习机器学习的内容之前，我们先集中讨论一些有关线性代数的内容。

本章为线性代数内容的基本介绍。若详细了解此部分内容可详细参考关于线性代数^[1]的书籍。若学习过此部分内容，可直接跳过本章节或者结合其与机器学习^[2]有关书籍进行学习。

2.1 标量、向量、矩阵和张量

标量 (scalar) 是一个单独的数，一般采用小写的斜体字母表示，例如标量 a 。

向量 (vector) 是有序排列的一列数，通常采用粗斜体的小写字母表示，如向量 $\mathbf{x}$ 。向量中的元素通过带下角标的斜体字母表示。例如，向量 $\mathbf{x}$ 的第一个元素是 x_1 ，第二个元素是 x_2 ，等等。我们通常利用方括号将向量中的元素包围成一个纵列，具体表示为

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad (2-1)$$

矩阵 (matrix) 是以方阵或长方阵形式有序排列的一组数，我们用粗斜体的大写字母表示矩阵，比如矩阵 $\mathbf{A}$ 。矩阵 $\mathbf{A}$ 中第 i 行第 j 列的元素表示为 $A_{i,j}$ 。 $\mathbf{A} \in \mathbb{R}^{m \times n}$ 表示矩阵 $\mathbf{A}$ 是一个 m 行 n 列的实数矩阵，具体表示为

$$A_{m \times n} = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m,1} & A_{m,2} & \cdots & A_{m,n} \end{bmatrix} \quad (2-2)$$

张量 (tensor) 是元素分布在若干维规则网格中的数组。这里我们用与矩阵一样的粗斜体大写字母表示张量, 如张量 $\mathbf{A}$ 。

2.2 矩阵的运算

2.2.1 基本运算

矩阵乘积 (matrix-product): 若 $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times p}$, 矩阵 $\mathbf{A}$ 的列数和矩阵 $\mathbf{B}$ 的行数相等。矩阵 $\mathbf{C}$ 为矩阵 $\mathbf{A}$ 和矩阵 $\mathbf{B}$ 的矩阵乘积, 那么 $\mathbf{C} \in \mathbb{R}^{m \times p}$, 矩阵乘法定义如下:

$$C_{i,j} = \sum_k A_{i,k} B_{k,j} \quad (2-3)$$

矩阵乘法表示为

$$\mathbf{C} = \mathbf{AB} \quad (2-4)$$

Hadamard 积 (hadamard-product): 两个形状相同的矩阵对应元素的乘积称为矩阵的 Hadamard 积。如果 $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{B} \in \mathbb{R}^{m \times n}$, 矩阵 $\mathbf{C}$ 为矩阵 $\mathbf{A}$ 和矩阵 $\mathbf{B}$ 的 Hadamard 积, 则 $\mathbf{C} \in \mathbb{R}^{m \times n}$, 记为 $\mathbf{C} = \mathbf{A} \odot \mathbf{B}$ 。

$$C_{i,j} = A_{i,j} B_{i,j} \quad (2-5)$$

Kronecker 积 (Kronecker-product): 矩阵的 Kronecker 积分为右 Kronecker 积和左 Kronecker 积。

右 Kronecker 积: 若 $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{B} \in \mathbb{R}^{p \times q}$, 矩阵 $\mathbf{C}$ 表示矩阵 $\mathbf{A}$ 和矩阵 $\mathbf{B}$ 的右 Kronecker 积, 则 $\mathbf{C} \in \mathbb{R}^{mp \times nq}$, 记为 $\mathbf{A} \otimes \mathbf{B}$ 。

具体地, 右 Kronecker 积的操作定义为:

$$\mathbf{A} \otimes \mathbf{B} = [\mathbf{A}_1 \mathbf{B}, \cdots, \mathbf{A}_n \mathbf{B}] = \begin{bmatrix} A_{1,1} \mathbf{B} & A_{1,2} \mathbf{B} & \cdots & A_{1,n} \mathbf{B} \\ A_{2,1} \mathbf{B} & A_{2,2} \mathbf{B} & \cdots & A_{2,n} \mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m,1} \mathbf{B} & A_{m,2} \mathbf{B} & \cdots & A_{m,n} \mathbf{B} \end{bmatrix} \quad (2-6)$$

同样方法可以定义左 Kronecker 积。如果用右 Kronecker 积的形式书写, 左 Kronecker 积可写为

$$[\mathbf{A} \otimes \mathbf{B}]_{\text{left}} = \mathbf{B} \otimes \mathbf{A}$$

通常情况下采用右 Kronecker 积。

逆矩阵 (inverse matrix): 方阵 $\mathbf{A}$ 的逆矩阵记作 $\mathbf{A}^{-1}$, 其定义为:

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n \quad (2-7)$$

2.2.2 矩阵的微分 (matrix differential)

矩阵微分^[3] 是计算标量、向量或者矩阵函数关于其向量或矩阵变元的偏导的有效工具。在机器学习算法学习过程中, 理论推导通常涉及矩阵的求导。但矩阵的求导本质上是多元变量的微积分问题。设 $\mathbf{y} \in \mathbb{R}^{m \times 1}$, $\mathbf{x} \in \mathbb{R}^{n \times 1}$, 因此 $\mathbf{y}$ 关于 $\mathbf{x}$ 的导数 $\frac{d\mathbf{y}}{d\mathbf{x}}$ 是一个 $m \times n$ 的矩阵。可以表示为

$$\begin{bmatrix} \frac{dy_1}{dx_1} & \frac{dy_1}{dx_2} & \cdots & \frac{dy_1}{dx_n} \\ \frac{dy_2}{dx_1} & \frac{dy_2}{dx_2} & \cdots & \frac{dy_2}{dx_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{dy_m}{dx_1} & \frac{dy_m}{dx_2} & \cdots & \frac{dy_m}{dx_n} \end{bmatrix} \quad (2-8)$$

该矩阵称为 Jacobian 矩阵, 并且将一阶矩阵微分简记为 $d\mathbf{X}$ 。

在机器学习算法中, 矩阵的求导原则遵循以下几种情况:

- 若 y 是一个标量, 则 Jacobian 矩阵为 $1 \times n$ 的矩阵。
- 若 x 是一个标量, 则 Jacobian 矩阵为 $m \times 1$ 的矩阵。
- 若 y 为标量, $\mathbf{X}$ 为 $m \times n$ 矩阵, 则 $\frac{dy}{d\mathbf{X}}$ 是 $m \times n$ 的一个矩阵。相当于 y 对 $\mathbf{X}$ 中的每一个元素进行求导。
- 若 x 为标量, $\mathbf{Y}$ 为 $m \times n$ 矩阵, 则 $\frac{d\mathbf{Y}}{dx}$ 是 $n \times m$ 的一个矩阵。相当于 $\mathbf{Y}$ 对 x 中的每一个元素进行求导, 然后再转置。

一阶实矩阵微分的性质:

- (1) 常数矩阵的微分矩阵为零矩阵, 即 $d\mathbf{X} = \mathbf{0}$ 。
- (2) 矩阵转置的微分等于矩阵微分的转置, 即 $d\mathbf{X}^T = (d\mathbf{X})^T$ 。
- (3) 线性性质, 即 $d(\alpha\mathbf{X} + \beta\mathbf{Y}) = \alpha d\mathbf{X} + \beta d\mathbf{Y}$ 。
- (4) 两个矩阵函数的和 (差) 的微分矩阵为 $d(\mathbf{X} \pm \mathbf{Y}) = d\mathbf{X} \pm d\mathbf{Y}$ 。
- (5) 常数矩阵与矩阵乘积的微分矩阵为 $d(\mathbf{A}\mathbf{X}\mathbf{B}) = \mathbf{A}(d\mathbf{X})\mathbf{B}$ 。
- (6) 矩阵函数 $\mathbf{U} = \mathbf{F}(\mathbf{X})$, $\mathbf{V} = \mathbf{G}(\mathbf{X})$, $\mathbf{W} = \mathbf{H}(\mathbf{X})$ 乘积的微分

$$d(\mathbf{UV}) = (d\mathbf{U})\mathbf{V} + \mathbf{U}(d\mathbf{V})$$

$$d(\mathbf{UVW}) = (d\mathbf{U})\mathbf{V}\mathbf{W} + \mathbf{U}(d\mathbf{V})\mathbf{W} + \mathbf{UV}(d\mathbf{W})$$

(7) 逆矩阵的微分矩阵为 $d(\mathbf{X}^{-1}) = -\mathbf{X}^{-1}(d\mathbf{X})\mathbf{X}^{-1}$ 。

2.3 特殊矩阵

在实际应用中，经常会遇到元素之间存在某种特殊结构关系的矩阵，统称为特殊矩阵。了解这些矩阵的内部特殊结构，有助于灵活地使用这些矩阵，简化很多问题的表示和求解。

对角矩阵 (diagonal matrix): 除对角线外其余元素都为 0 的矩阵。主对角线元素可以为 0 或其他值。

对称矩阵 (symmetric matrix): 转置等于本身的矩阵。

$$\mathbf{A} = \mathbf{A}^T \quad (2-9)$$

例如，如果 $\mathbf{A}$ 是一个距离度量矩阵， $\mathbf{A}_{i,j}$ 表示点 i 到点 j 的距离， $\mathbf{A}_{j,i}$ 表示点 j 到点 i 的距离，则 $\mathbf{A}_{i,j} = \mathbf{A}_{j,i}$ 。

正交矩阵 (orthogonal matrix): 是指矩阵的逆等于其转置的矩阵，即

$$\mathbf{A}^T \mathbf{A} = \mathbf{A} \mathbf{A}^T = \mathbf{I} \quad (2-10)$$

上三角矩阵 (upper triangular matrix): 主对角线以下元素全为 0 的矩阵。当 $i > j$ 时， $\mathbf{A}_{ij} = 0$ 。其一般形式表示为

$$\begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\ 0 & A_{2,2} & \cdots & A_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{n,n} \end{bmatrix} \quad (2-11)$$

2.4 线性空间、线性相关和线性变换

数域: 设 $\mathbb{S}$ 是复数集 $\mathbb{C}$ 的一个非空子集， $\mathbb{S}$ 中含有非零的数。若对于 $\mathbb{S}$ 中任意两个数 a 和 b ，都有 $a+b \in \mathbb{S}$ ， $a-b \in \mathbb{S}$ ， $ab \in \mathbb{S}$ ，并且若 $b \neq 0$ ，有 $\frac{a}{b} \in \mathbb{S}$ ，则称 $\mathbb{S}$ 为一个数域。有理数集 $\mathbb{Q}$ 、实数集 $\mathbb{R}$ 和复数集 $\mathbb{C}$ 都是数域。

向量空间: 设 $\mathbb{V}$ 是一个非空向量集合， $\mathbb{S}$ 是一个数域。如果下列公理被满足，则称集合 $\mathbb{V}$ 是数域 $\mathbb{S}$ 上的一个向量空间。向量空间又称线性空间。

(1) 闭合性 (closure properties) 的公理

- 若 $\mathbf{x} \in \mathbb{V}$ 和 $\mathbf{y} \in \mathbb{V}$ ，则 $\mathbf{x} + \mathbf{y} \in \mathbb{V}$ ，即 $\mathbb{V}$ 在加法下是闭合的，简称加法的闭合性。
- 若 $a \in \mathbb{S}$ ， $\mathbf{y} \in \mathbb{V}$ ，则 $a\mathbf{y} \in \mathbb{V}$ ，即 $\mathbb{V}$ 在标量乘法下是闭合的，简称标量乘法的闭合性。

(2) 加法的公理

- $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{V}$, 称为加法交换律。
- $\mathbf{x} + (\mathbf{y} + \mathbf{w}) = (\mathbf{x} + \mathbf{y}) + \mathbf{w}$, $\forall \mathbf{x}, \mathbf{y}, \mathbf{w} \in \mathbb{V}$, 称为加法结合律。
- 在 $\mathbb{V}$ 中存在一个零向量, 使得 $\forall \mathbf{y} \in \mathbb{V}$, 恒有 $\mathbf{y} + \mathbf{0} = \mathbf{y}$ 。称为零向量的存在性。
- 给定一个向量 $\mathbf{y} \in \mathbb{V}$, 存在另一个向量 $\mathbf{y}' \in \mathbb{V}$ 使得 $\mathbf{y} + \mathbf{y}' = \mathbf{0}$, 称为负向量的存在性。

(3) 标量乘法的公理

- $a(b\mathbf{y}) = (ab)\mathbf{y}$, 对 $\forall \mathbf{y} \in \mathbb{V}$ 和 $\forall a, b \in \mathbb{S}$ 成立, 称为标量乘法结合律。
- $a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y}$, 对 $\forall \mathbf{x}, \mathbf{y} \in \mathbb{V}$ 和 $\forall a \in \mathbb{S}$ 成立, 称为标量乘法分配律。
- $(a + b)\mathbf{y} = a\mathbf{y} + b\mathbf{y}$, 对 $\forall \mathbf{y} \in \mathbb{V}$ 和 $\forall a, b \in \mathbb{S}$ 成立, 称为标量乘法的分配律。
- $1\mathbf{y} = \mathbf{y}$, 对 $\forall \mathbf{y} \in \mathbb{V}$ 成立, 称为标量乘法单位律。

假设 $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m$ 是向量空间 $\mathbb{V}$ 中的一组向量, 如果存在数域 $\mathbb{S}$ 中不全为零的数 $k_1, k_2, \dots, k_m$ 使得

$$k_1\mathbf{a}_1 + k_2\mathbf{a}_2 + \dots + k_m\mathbf{a}_m = \mathbf{0} \quad (2-12)$$

则称向量组 $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m$ 是线性相关的。若 $k_1, k_2, \dots, k_m$ 全为 0, 则称为线性无关。

令 $\mathbb{V}$ 和 $\mathbb{W}$ 为数域 $\mathbb{S}$ 上的向量空间, 并且 $T: \mathbb{V} \rightarrow \mathbb{W}$ 是一个映射, 若对 $\forall \mathbf{v}, \mathbf{w} \in \mathbb{V}$ 和 $\forall c \in \mathbb{S}$, 映射 T 满足叠加性

$$T(\mathbf{v} + \mathbf{w}) = T(\mathbf{v}) + T(\mathbf{w}) \quad (2-13)$$

和齐次性

$$T(c\mathbf{v}) = cT(\mathbf{v}) \quad (2-14)$$

则称 T 为线性映射或线性变换。

2.5 内积与范数

向量和矩阵的内积在计算范数时是非常有用的。而范数在机器学习中经常被用来衡量向量的大小。

2.5.1 向量和矩阵的内积

n 维向量 $\mathbf{x} = [x_1, x_2, \dots, x_n]$ 和向量 $\mathbf{y} = [y_1, y_2, \dots, y_n]$ 之间的内积 (inner-product) 表示为

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} = \sum_{i=1}^n x_i y_i \quad (2-15)$$

该内积通常情况下称为典范内积。

将向量的内积加以推广, 即可得到矩阵的内积。

形状为 $m \times n$ 的矩阵 $\mathbf{A}$ 和矩阵 $\mathbf{B}$ ，将这两个矩阵分别“拉长”为 $mn \times 1$ 的向量。

$$\mathbf{a} = \text{vec}(\mathbf{A}) = \begin{bmatrix} a_1 \\ \vdots \\ a_{mn} \end{bmatrix}, \quad \mathbf{b} = \text{vec}(\mathbf{B}) = \begin{bmatrix} b_1 \\ \vdots \\ b_{mn} \end{bmatrix}$$

$\text{vec}(\mathbf{A})$ 称为矩阵 $\mathbf{A}$ 的向量化。

矩阵的内积记为 $\langle \mathbf{A}, \mathbf{B} \rangle$ ，定义为两个“拉长向量” $\mathbf{a}$ 和 $\mathbf{b}$ 之间的内积。

$$\langle \mathbf{A}, \mathbf{B} \rangle = \langle \text{vec}(\mathbf{A}), \text{vec}(\mathbf{B}) \rangle = \sum_{i=1}^{mn} a_i b_i \quad (2-16)$$

2.5.2 向量和矩阵的范数

范数是将向量映射到非负值的函数。直观上来说，向量 $\mathbf{x}$ 的范数衡量从原点到点 $\mathbf{x}$ 的距离。范数作为特殊的函数，满足以下三条性质。

- $f(\mathbf{x}) = 0 \Rightarrow \mathbf{x} = \mathbf{0}$
- $f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$
- $\forall \alpha \in \mathbb{R}, f(\alpha \mathbf{x}) = |\alpha| f(\mathbf{x})$

在机器学习中，我们经常使用范数 (norm) 衡量向量大小。向量的 L^p 范数定义如下：

$$\|\mathbf{x}\|_p = \left(\sum_i |x_i|^p \right)^{\frac{1}{p}} \quad (2-17)$$

其中 $p \in \mathbb{R}, p \geq 1$ 。

将 $m \times n$ 的矩阵排列成一个 $mn \times 1$ 的向量，采用向量的范数的定义即可得到矩阵的范数。由于这类范数是使用矩阵的元素表示的，故称“元素范数” (“entrywise”-norm)。因此矩阵 $\mathbf{A}$ 的 p 范数表示为

$$\|\mathbf{A}\|_p = \left(\sum_{i=1}^m \sum_{j=1}^n |A_{ij}|^p \right)^{1/p} \quad (2-18)$$

以下是三种典型的元素形式 p 范数：

(1) L_1 范数 ($p = 1$)

$$\|\mathbf{A}\|_1 = \left(\sum_{i=1}^m \sum_{j=1}^n |A_{ij}| \right) \quad (2-19)$$

(2) Frobenius 范数 ($p = 2$)

$$\|\mathbf{A}\|_F = \left(\sum_{i=1}^m \sum_{j=1}^n |A_{ij}|^2 \right)^{1/2} \quad (2-20)$$

(3) 最大范数 max-norm ($p = \infty$)

$$\|\mathbf{A}\|_{\infty} = \max_{i=1, \dots, m; j=1, \dots, n} \{|A_{ij}|\} \quad (2-21)$$

令 $\boldsymbol{\sigma} = [\sigma_1, \dots, \sigma_k]^T$, $k = \min\{m, n\}$ 表示矩阵 $\mathbf{A}$ 的全部奇异值组成的向量, 则称

$$\|\mathbf{A}\|_p = \|\boldsymbol{\sigma}\|_p = \left(\sum_{i=1}^{\min\{m, n\}} \sigma_i^p \right)^{1/p} \quad (2-22)$$

是矩阵 $\mathbf{A}$ 的 Schatten p 范数。

最常用的 Schatten 范数是以下三种情况:

(1) 核范数 ($p = 1$)

$$\|\mathbf{A}\|_1 = \|\boldsymbol{\sigma}\|_* = \left(\sum_{i=1}^{\min\{m, n\}} \sigma_i \right) \quad (2-23)$$

(2) Schatten 范数与 Frobenius 范数 ($p = 2$)

$$\|\boldsymbol{\sigma}\|_2 = \|\boldsymbol{\sigma}\|_F = \left(\sum_{i=1}^m \sum_{j=1}^n |A_{ij}|^2 \right)^{1/2} \quad (2-24)$$

(3) Schatten 范数与谱范数相同 ($p = \infty$)

$$\|\boldsymbol{\sigma}\|_{\infty} = \sigma_{\max}(\mathbf{A}) \quad (2-25)$$

2.6 矩阵分解

矩阵分解是矩阵理论中非常重要的内容。矩阵分解方法繁多, 本节不能详尽描述, 只对机器学习中常用的两种矩阵分解方法简要介绍。

2.6.1 特征分解

对于一个整数, 我们通常可以将其分解为质因数的乘积。例如 $12 = 2 \times 2 \times 3$ 。从该等式中我们可以发现, 12 的倍数可以被 3 整除。本节我们将整数的分解推广至矩阵的分解。

矩阵的特征分解 (eigendecomposition) 就是将矩阵分解为一组特征向量和特征值。方阵 $\mathbf{A}$ 的特征向量 (eigenvector) 是指与矩阵 $\mathbf{A}$ 相乘后相当于对该向量进行缩放的非零向量 $\mathbf{v}$:

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \quad (2-26)$$

λ 被称为这个特征向量对应的特征值 (eigenvalue)。

假设矩阵 $\mathbf{A}$ 有 n 个线性无关的特征向量 $\{\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(n)}\}$, 对应的特征值为 $\{\lambda_1, \dots, \lambda_n\}$ 。我们令每一个特征向量为 一列, 将这 n 个特征向量拼成一个矩阵: $\mathbf{V} = [\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(n)}]$ 。我们将 n 个特征值连接成一个向量 $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_n]^T$ 。因此矩阵 $\mathbf{A}$ 的特征分解可以表示为

$$\mathbf{A} = \mathbf{V} \text{diag}(\boldsymbol{\lambda}) \mathbf{V}^{-1} \quad (2-27)$$

并不是每一个矩阵都可以进行特征分解, 因此本书中考虑实对称矩阵。实对称矩阵的特征分解可以表示为

$$\mathbf{A} = \mathbf{Q} \boldsymbol{\Lambda} \mathbf{Q}^T \quad (2-28)$$

其中 $\mathbf{Q}$ 是 $\mathbf{A}$ 的特征向量组成的正交矩阵, $\boldsymbol{\Lambda}$ 是对角矩阵。

2.6.2 奇异值分解

除了特征分解外, 还有一种分解矩阵的方法, 被称为奇异值分解 (singular value decomposition, SVD), 将矩阵分解为 **奇异向量** (singular vector) 和 **奇异值** (singular value)。每个实数矩阵都可以进行奇异值分解, 但不一定可以特征分解。

方阵 $\mathbf{A}$ 的特征分解可以表示为

$$\mathbf{A} = \mathbf{V} \text{diag}(\boldsymbol{\lambda}) \mathbf{V}^{-1} \quad (2-29)$$

其中, $\mathbf{V}$ 表示 $\mathbf{A}$ 的特征向量构成的矩阵, $\boldsymbol{\lambda}$ 表示 $\mathbf{A}$ 的特征值构成的向量。

奇异值分解是类似的, 将矩阵 $\mathbf{A}$ 分解成三个矩阵的乘积:

$$\mathbf{A} = \mathbf{U} \mathbf{D} \mathbf{V}^T \quad (2-30)$$

若 $\mathbf{A}$ 的形状为 $m \times n$, 那么 $\mathbf{U}$, $\mathbf{D}$, $\mathbf{V}$ 的形状分别是 $m \times m$, $m \times n$, $n \times n$ 。其中矩阵 $\mathbf{U}$ 和 $\mathbf{V}$ 都为正交矩阵, 而矩阵 $\mathbf{D}$ 为对角矩阵。注意, 矩阵 $\mathbf{D}$ 不一定是方阵。

对角矩阵 $\mathbf{D}$ 对角线上的元素被称为矩阵 $\mathbf{A}$ 的奇异值。矩阵 $\mathbf{U}$ 的列向量被称为左奇异向量 (left singular vector), 矩阵 $\mathbf{V}$ 的列向量被称右奇异值向量 (right singular vector)。

2.7 Moore-Penrose 伪逆

Moore-Penrose 伪逆 (Moore-Penrose pseudoinverse) 是方阵的逆定义在非方阵中的推广。对于下面的线性方程组, 我们希望通过矩阵 $\mathbf{A}$ 的左逆矩阵 $\mathbf{B}$ 来求解

$$\mathbf{A} \mathbf{x} = \mathbf{y} \quad (2-31)$$

等式两边左乘矩阵 $\mathbf{B}$ 后, 有

$$\mathbf{x} = \mathbf{B} \mathbf{y} \quad (2-32)$$

取决于实际问题的形式, 矩阵 $\mathbf{A}$ 到 $\mathbf{B}$ 的映射可能并不唯一。

设 $\mathbf{A} \in \mathbb{R}^{m \times n}$, 若 $m > n$, 则上述方程组可能没有解。若 $m < n$, 则矩阵 $\mathbf{A}$ 到 $\mathbf{B}$ 的可能有多种映射。

Moore-Penrose 伪逆^[4] 使我们在这类问题上取得了一定的进展。矩阵 $\mathbf{A}$ 的伪逆定义为:

$$\mathbf{A}^\dagger = \lim_{\alpha \rightarrow 0} (\mathbf{A}^\mathrm{T} \mathbf{A} + \alpha \mathbf{I})^{-1} \mathbf{A}^\mathrm{T} \quad (2-33)$$

Moore-Penrose 伪逆的实际计算使用的是下面的公式:

$$\mathbf{A}^\dagger = \mathbf{V} \mathbf{D}^+ \mathbf{U}^\mathrm{T} \quad (2-34)$$

其中, 矩阵 $\mathbf{U}$, $\mathbf{D}$ 和 $\mathbf{V}$ 是矩阵 $\mathbf{A}$ 奇异值分解后得到的矩阵。 $\mathbf{D}^+$ 是对角矩阵 $\mathbf{D}$ 中非零元素取倒数之后再转置得到的。

2.8 MATLAB 函数和示例

2.8.1 矩阵求逆

设矩阵 $\mathbf{A}$, 其逆为 $\mathbf{Y} = \text{inv}(\mathbf{A})$ 。例如: 矩阵 $\mathbf{A}$ 表示为

$$\begin{bmatrix} 1 & 0 & 2 \\ -1 & 5 & 0 \\ 0 & 3 & -9 \end{bmatrix} \quad (2-35)$$

则矩阵 $\mathbf{Y}$ 可计算为

$$\begin{bmatrix} 0.8824 & -0.1176 & 0.1961 \\ 0.1765 & 0.1765 & 0.0392 \\ 0.0588 & 0.0588 & -0.0980 \end{bmatrix} \quad (2-36)$$

并且满足 $\mathbf{X} * \mathbf{Y} = \mathbf{I}$ 。

2.8.2 矩阵的范数

设矩阵 $\mathbf{A}$, 则 $\text{norm}(\mathbf{A}, 2)$ 表示矩阵 $\mathbf{A}$ 的 2 范数; $\text{norm}(\mathbf{A}, \text{INF})$ 表示矩阵 $\mathbf{A}$ 的无穷范数。例如, 矩阵 $\mathbf{A}$ 表示为

$$\begin{bmatrix} 2 & 0 & 1 \\ -1 & 1 & 0 \\ -3 & 3 & 0 \end{bmatrix} \quad (2-37)$$

则 $\text{norm}(\mathbf{X}, 2) = 4.7234$ 。

2.8.3 矩阵求特征值

设矩阵 $\mathbf{A}$, 则 $\text{eig}(\mathbf{A})$ 返回值为一个列向量, 表示矩阵 $\mathbf{A}$ 的全部特征值。 $[\mathbf{V}, \mathbf{D}] = \text{eig}(\mathbf{A})$

返回值 $\mathbf{D}$ 为一个对角矩阵，为矩阵 $\mathbf{A}$ 的全部特征值，矩阵 $\mathbf{D}$ 表示返回矩阵 $\mathbf{A}$ 特征值对应的特征向量。

例如：矩阵 $\mathbf{A}$ 表示为 $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{bmatrix}$ ，则其特征向量、特征值分别为

$$\mathbf{V} = \begin{bmatrix} -0.2235 & -0.8658 & 0.2783 \\ -0.5039 & 0.0857 & -0.8318 \\ -0.8343 & 0.4929 & 0.4802 \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} 16.7075 & 0 & 0 \\ 0 & -0.9057 & 0 \\ 0 & 0 & 0.1982 \end{bmatrix}$$

2.8.4 矩阵的奇异值 (SVD) 分解

设矩阵 $\mathbf{X}$ ，则奇异值分解为： $[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \text{svd}(\mathbf{A})$ ，满足 $\mathbf{A} = \mathbf{U} * \mathbf{S} * \mathbf{V}'$ ，例如矩阵 $\mathbf{A}$ 表示为

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{bmatrix} \quad (2-38)$$

则其 SVD 分解为

$$\mathbf{U} = \begin{bmatrix} -0.1525 & -0.8226 \\ -0.3499 & -0.4214 \\ -0.5474 & -0.0201 \\ -0.7448 & 0.3812 \end{bmatrix} \quad (2-39)$$

$$\mathbf{S} = \begin{bmatrix} 14.2691 & 0 \\ 0 & 0.6268 \end{bmatrix} \quad (2-40)$$

$$\mathbf{V} = \begin{bmatrix} -0.6414 & -0.7672 \\ 0.7672 & -0.6414 \end{bmatrix} \quad (2-41)$$

参 考 文 献

- [1] Golub G H, Van Loan C F. Matrix computations[M]. JHU Press, 2013.
- [2] Goodfellow I, Bengio Y, Courville A. Deep Learning[M]. The MIT Press, 2016.
- [3] 张贤达. 矩阵分析与应用 [M]. 北京: 清华大学出版社, 2013.
- [4] 吕红力. 基于稀疏和低秩表示的 OCT 图像去噪算法研究 [D]. 山东大学, 2018.

习 题

1. 令矩阵 $\mathbf{A}$ 是一个 3×4 的矩阵, 证明其为列线性相关。
2. 根据右 Kronecker 积得到左 Kronecker 积的计算式。
3. 设 $\mathbf{A}$ 为一对称矩阵, 并且矩阵 $\mathbf{M}$ 是 $\mathbf{A}$ 的 Moore-Penrose 逆矩阵。证明: 矩阵 $\mathbf{M}^2$ 是 $\mathbf{A}^2$ 的 Moore-Penrose 逆矩阵。
4. 证明 Kronecker 积的 Moore-Penrose 逆矩阵的关系为 $(\mathbf{A} \otimes \mathbf{B})^+ = \mathbf{A}^+ \otimes \mathbf{B}^+$ 。