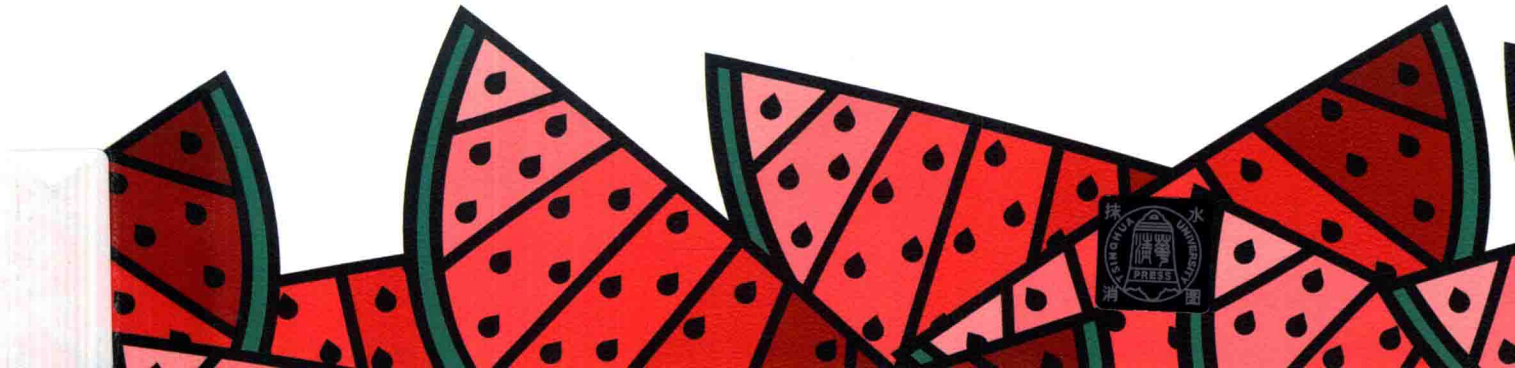
周志华 著

MACHINE LEARNING

机器学习

清华大学出版社



周志华 著

MACHINE LEARNING

机器学习

清华大学出版社
北京

内 容 简 介

机器学习是计算机科学的重要分支领域. 本书作为该领域的入门教材，在内容上尽可能涵盖机器学习基础知识的各方面. 全书共 16 章，大致分为 3 个部分：第 1 部分（第 1～3 章）介绍机器学习的基础知识；第 2 部分（第 4～10 章）讨论一些经典而常用的机器学习方法（决策树、神经网络、支持向量机、贝叶斯分类器、集成学习、聚类、降维与度量学习）；第 3 部分（第 11～16 章）为进阶知识，内容涉及特征选择与稀疏学习、计算学习理论、半监督学习、概率图模型、规则学习以及强化学习等. 每章都附有习题并介绍了相关阅读材料，以便有兴趣的读者进一步钻研探索.

本书可作为高等院校计算机、自动化及相关专业的本科生或研究生教材，也可供对机器学习感兴趣的研究人员和工程技术人员阅读参考.

图书在版编目(CIP)数据

机器学习/周志华著. --北京：清华大学出版社，2016 (2016.5 重印)
ISBN 978-7-302-42328-7

Ⅰ. ①机… Ⅱ. ①周… Ⅲ. ①机器学习 Ⅳ. ①TP181

中国版本图书馆 CIP 数据核字(2015)第 287090 号

责任编辑：薛　慧
封面设计：常雪影
责任校对：刘玉霞
责任印制：宋　林

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn, http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969, c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015, zhiliang@tup.tsinghua.edu.cn
印 装 者：北京亿浓世纪彩色印刷有限公司
经　　销：全国新华书店
开　　本：210mm×235mm　　**印　张**：27.75　　**字　数**：626 千字
版　　次：2016 年 1 月第 1 版　　**印　次**：2016 年 5 月第 8 次印刷
印　　数：53001～63000
定　　价：88.00 元

产品编号：064027-01

序　言

在人工智能界有一种说法, 认为机器学习是人工智能领域中最能够体现智能的一个分支. 从历史来看, 机器学习似乎也是人工智能中发展最快的分支之一. 在二十世纪八十年代的时候, 符号学习可能还是机器学习的主流, 而自从二十世纪九十年代以来, 就一直是统计机器学习的天下了. 不知道是否可以这样认为: 从主流为符号机器学习发展到主流为统计机器学习, 反映了机器学习从纯粹的理论研究和模型研究发展到以解决现实生活中实际问题为目的的应用研究, 这是科学研究的一种进步. 有关机器学习的专著国内出版的不是很多. 前两年有李航教授的《统计学习方法》出版, 以简要的方式介绍了一批重要和常用的机器学习方法. 此次周志华教授的鸿篇巨著《机器学习》则全面而详细地介绍了机器学习的各个分支, 既可作为教材, 又可作为自学用书和科研参考书.

翻阅书稿的过程引起了一些自己的思考, 平时由于和机器学习界的朋友接触多了, 经常获得一些道听途说的信息以及专家们对机器学习现状及其发展前途的评论. 在此过程中, 难免会产生一些自己的疑问. 我借此机会把它写下来放在这里, 算是一种“外行求教机器学习”.

问题一: 在人工智能发展早期, 机器学习的技术内涵几乎全部是符号学习. 可是从二十世纪九十年代开始, 统计机器学习犹如一匹黑马横空出世, 迅速压倒并取代了符号学习的地位. 人们可能会问: 在满目的统计学习期刊和会议文章面前, 符号学习是否被彻底忽略了? 它还能成为机器学习的研究对象吗? 它是否将继续在统计学习的阴影里生活并苟延残喘? 对这个问题有三种可能的答案: 一是告诉符号学习: “你就是该退出历史舞台, 认命吧!” 二是告诉统计学习: “你的一言堂应该关门了!” 单纯的统计学习已经走到了尽头, 再想往前走就要把统计学习和符号学习结合起来. 三是事物发展总会有“三十年河东, 三十年河西”的现象, 符号学习还有“翻身”的日子. 第一种观点我没有听人明说过, 但是我想恐怕有可能已经被许多人默认了. 第二种观点我曾听王珏教授多次说过. 他并不认为统计学习会衰退, 而只是认为机器学习已经到了一个转折点, 从今往后, 统计学习应该和知识的利用相结合, 这是一种“螺旋式上升, 进入更高级的形式”, 否则, 统计学习可能会停留于现状而止步不前. 王珏教授还认为: 进入转折点的标志就是 Koller 等的《概率图模型》一书的出版. 至于第三种观点, 恰好我收到老朋友, 美国人工智能资深学者、俄亥俄大学 Chandrasekaran 教授的来信, 他正好谈起符号智能被统计智能“打压”的现象, 并且正好表达了河东河西的观点. 我请求他允许我把这段话引进正在撰写的序言中, 他爽快地同意了, 仅仅修改了几处私人通信的口吻. 全文如下: “最近几年, 人工智能在很大程度上集中于统计学和大数据. 我同意由于计算能力的大幅提高, 这些技术曾经取得过某些令人印象深刻的成果. 但是我们完全有理由相信, 虽然这些技术还会继续改进、提高, 总有一天这个领域(指 AI)会对它们说再见, 并转向更加基本的认知科学研究. 尽管钟摆的摆回去还需要一段时间, 我

相信定有必要把统计技术和对认知结构的深刻理解结合起来.”看来, Chandrasekaran 教授也并不认为若干年以后 AI 真会回到河西, 他的意见和王珏教授的意见基本一致, 但不仅限于机器学习, 而是涉及整个人工智能领域. 只是王珏教授强调知识, 而 Chandrasekaran 教授强调更加基本的“认知”.

问题二: 王珏教授认为统计机器学习不会“一路顺风”的判据是: 统计机器学习算法都是基于样本数据独立同分布的假设. 但是自然界现象千变万化, 王珏教授认为“哪有那么多独立同分布?”这就引来了下一个问题: “独立同分布”条件对于机器学习来讲真是必需的吗? 独立同分布的不存在一定是一个不可逾越的障碍吗? 无独立同分布条件下的机器学习也许只是一个难题, 而不是不可解问题. 我有一个“胡思乱想”, 认为前些时候出现的“迁移学习”也许会对这个问题的解决带来一线曙光. 尽管现在的迁移学习还要求迁移双方具备“独立同分布”条件, 但是不同分布之间的迁移学习, 同分布和异分布之间的迁移学习也许迟早会出现?

问题三: 近年来出现了一些新的动向, 例如“深度学习”、“无终止学习”等等, 社会上给予了特别关注, 尤其是深度学习. 但它们真的代表了机器学习的新的方向吗? 包括本书作者周志华教授在内的一些学者认为: 深度学习掀起的热潮也许大过它本身真正的贡献, 在理论和技术上并没有太多的创新, 只不过是由于硬件技术的革命, 计算机的速度大大提高了, 使得人们有可能采用原来复杂度很高的算法, 从而得到比过去更精细的结果. 当然这对于推动机器学习应用于实践有很大意义. 但我们不禁要斗胆问一句: 深度学习是否又要取代统计学习了? 事实上, 确有专家已经感受到来自深度学习的压力, 指出统计学习正在被深度学习所打压, 正如我们早就看到的符号学习被统计学习所打压. 不过我觉得这种打压还远没有强大到像统计学习打压符号学习的程度. 这一是因为深度学习的“理论创新”还不明显; 二是因为目前的深度学习主要适合于神经网络, 在各种机器学习方法百花盛开的今天, 它的应用范围还有限, 还不能直接说是连接主义方法的回归; 三是因为统计学习仍然在机器学习中被有效地普遍采用, “得道多助”, 想抛弃它不容易.

问题四: 机器学习研究出现以来,我们看到的主要是从符号方法到统计方法的演变, 用到的数学主要是概率统计. 但是, 数学之大, 就像大海. 难道只有统计方法适合于在机器学习方面应用吗? 当然, 我们也看到了一些其他数学分支在机器学习上的应用的好例子, 例如微分几何在流形学习上的应用, 微分方程在归纳学习上的应用. 但如果和统计方法相比, 它们都只能算是配角. 还有的数学分支如代数可能应用得更广, 但在机器学习中代数一般是作为基础工具来使用, 例如矩阵理论和特征值理论. 又如微分方程求解最终往往归结为代数问题求解. 它们可以算是幕后英雄: “出头露面的是概率和统计, 埋头苦干的是代数和逻辑”. 是否可以想象以数学方法为主角, 以统计方法为配角的机器学习理论呢? 在这方面, 流形学习已经“有点意思”了, 而彭实戈院士的倒排随机微分方程理论之预测金融走势, 也许是用高深数学推动新的机器学习模式的更好例子. 但是从宏观的角度看, 数学理论的介入程度还远远不够. 这里指的主要是深刻的、现代的数学理论, 我们期待着有更多数学家的参与, 开辟机器学习的新模式、新理论、新方向.

问题五: 上一个问题的延续: 符号机器学习时代主要以离散方法处理问题, 统计机器学习时代主要以连续方法处理问题. 这两种方法之间应该没有一条鸿沟. 流形学习中李群、李代数方法的引入给我们以很好的启示. 从微分流形到李群, 再从李群到李代数, 就是一个沟通连续和离散的过程. 然而, 现有的方法在数学上并不完美. 浏览流形学习的文献可知, 许多论文直接把任意数据集看成微分流形, 从而就认定测地线的存在并讨论起降维来了. 这样的例子也许不是个别的, 足可说明数学家介入机器学习研究之必要.

问题六: 大数据时代的出现, 有没有给机器学习带来本质性的影响? 理论上讲, 似乎“大数据”给统计机器学习提供了更多的机遇, 因为海量的数据更加需要统计、抽样的方法. 业界人士估计, 大数据的出现将使人工智能的作用更加突出. 有人把大数据处理分成三个阶段: 收集、分析和预测. 收集和分析的工作相对来说已经做得相当好了, 现在关注的焦点是要有科学的预测, 机器学习技术在这里不可或缺. 这一点大概毋庸置疑. 然而, 同样是使用统计、抽样方法, 同样是收集、分析和预测, 大数据时代使用这类方法和以前使用这类方法有什么本质的不同吗? 量变到质变是辩证法的一个普遍规律. 那么, 从前大数据时代到大数据时代, 数理统计方法有没有发生本质的变化? 反映到它们在机器学习上的应用有无本质变化? 大数据时代正在呼唤什么样的机器学习方法的产生? 哪些机器学习方法又是由于大数据研究的驱动而产生的呢?

以上这些话也许说得远了, 我们还是回到本书上来. 本书的作者周志华教授在机器学习的许多领域都有出色的贡献, 是中国机器学习研究的领军人物之一, 在国际学术界有着很高的声誉. 他在机器学习的一些重要领域, 例如集成学习、半监督学习、多示例和多标记学习等方面都做出了在国际上有重要影响的工作, 其中一些可以认为是中国学者在国际上的代表性贡献. 除了自身的学术研究以外, 他在推动中国的机器学习发展方面也做了许多工作. 例如他和不久前刚过世的王珏教授从 2002 年开始, 组织了系列化的“机器学习及其应用”研讨会. 初在复旦, 后移至南大举行, 越办越兴旺, 从单一的专家报告发展到专家报告、学生论坛和张贴论文三种方式同时举行, 参会者从数十人发展到数百人, 活动搞得有声有色, 如火如荼. 最近更是把研讨会推向全国高校轮流举行. 他和王珏教授紧密合作, 南北呼应, 人称“南周北王”. 王珏教授的离去使我们深感悲伤. 令我们欣慰的是国内不但有周志华教授这样的机器学习领军人物, 而且比周教授更年轻的许多机器学习青年才俊也成长起来了. 中国的机器学习大有希望.

陆汝钤

中国科学院数学与系统科学研究院

2015 年 8 月于北京

前　言

这是一本面向中文读者的机器学习教科书, 为了使尽可能多的读者通过本书对机器学习有所了解, 作者试图尽可能少地使用数学知识. 然而, 少量的概率、统计、代数、优化、逻辑知识似乎不可避免. 因此, 本书更适合大学三年级以上的理工科本科生和研究生, 以及具有类似背景的对机器学习感兴趣的人士. 为方便读者, 本书附录给出了一些相关数学基础知识简介.

全书共 16 章, 大体上可分为 3 个部分：第 1 部分包括第 1~3 章, 介绍机器学习基础知识; 第 2 部分包括第 4~10 章, 介绍一些经典而常用的机器学习方法; 第 3 部分包括第 11~16 章, 介绍一些进阶知识. 前 3 章之外的后续各章均相对独立, 读者可根据自己的兴趣和时间情况选择使用. 根据课时情况, 一个学期的本科生课程可考虑讲授前 9 章或前 10 章; 研究生课程则不妨使用全书.

书中除第 1 章外, 每章都给出了十道习题. 有的习题是帮助读者巩固本章学习, 有的是为了引导读者扩展相关知识. 一学期的一般课程可使用这些习题, 再辅以两到三个针对具体数据集的大作业. 带星号的习题则有相当难度, 有些并无现成答案, 谨供富有进取心的读者启发思考.

本书在内容上尽可能涵盖机器学习基础知识的各方面, 但作为机器学习入门读物且因授课时间的考虑, 很多重要、前沿的材料未能覆盖, 即便覆盖到的部分也仅是管中窥豹, 更多的内容留待读者在进阶课程中学习. 为便于有兴趣的读者进一步钻研探索, 本书每章均介绍了一些阅读材料, 谨供读者参考.

笔者以为, 对学科相关的重要人物和事件有一定了解, 将会增进读者对该学科的认识. 本书在每章最后都写了一个与该章内容相关的小故事, 希望有助于读者增广见闻, 并且在紧张的学习过程中稍微放松调剂一下.

书中不可避免地涉及大量外国人名, 若全部译为中文, 则读者在日后进一步阅读文献时或许会对不少人名产生陌生感, 不利于进一步学习. 因此, 本书仅对一般读者耳熟能详的名字如“图灵”等加以直接使用, 对故事中的一些主要人物给出了译名, 其他则保持外文名.

机器学习发展极迅速, 目前已成为一个广袤的学科, 罕有人士能对其众多分支领域均有精深理解. 笔者自认才疏学浅, 仅略知皮毛, 更兼时间和精力所限, 书中错谬之处在所难免, 若蒙读者诸君不吝告知, 将不胜感激.

周志华

2015 年 6 月

主要符号表

x	标量
$\boldsymbol{x}$	向量
$\mathbf{x}$	变量集
$\mathbf{A}$	矩阵
$\mathbf{I}$	单位阵
$\mathcal{X}$	样本空间或状态空间
$\mathcal{D}$	概率分布
D	数据样本（数据集）
$\mathcal{H}$	假设空间
H	假设集
$\mathfrak{L}$	学习算法
$(\cdot,\ \cdot,\ \cdot)$	行向量
$(\cdot;\ \cdot;\ \cdot)$	列向量
$(\cdot)^{\mathrm{T}}$	向量或矩阵转置
$\{\cdots\}$	集合
$\lvert\{\cdots\}\rvert$	集合$\{\cdots\}$中元素个数
$\lVert\cdot\rVert_p$	L_p 范数, p 缺省时为 L_2 范数
$P(\cdot),\ P(\cdot \mid \cdot)$	概率质量函数, 条件概率质量函数
$p(\cdot),\ p(\cdot \mid \cdot)$	概率密度函数, 条件概率密度函数
$\mathbb{E}_{\cdot \sim \mathcal{D}}[f(\cdot)]$	函数 $f(\cdot)$ 对 $\cdot$ 在分布 $\mathcal{D}$ 下的数学期望; 意义明确时将省略 $\mathcal{D}$ 和(或) $\cdot$
$\sup(\cdot)$	上确界
$\mathbb{I}(\cdot)$	指示函数, 在 $\cdot$ 为真和假时分别取值为 $1, 0$
$\mathrm{sign}(\cdot)$	符号函数, 在 $\cdot < 0,\ = 0,\ > 0$ 时分别取值为 $-1, 0, 1$

目　录

第1章 绪 论

1.1 引言

傍晚小街路面上沁出微雨后的湿润, 和煦的细风吹来, 抬头看看天边的晚霞, 嗯, 明天又是一个好天气. 走到水果摊旁, 挑了个根蒂蜷缩、敲起来声音浊响的青绿西瓜, 一边满心期待着皮薄肉厚瓤甜的爽落感, 一边愉快地想着, 这学期狠下了工夫, 基础概念弄得清清楚楚, 算法作业也是信手拈来, 这门课成绩一定差不了!

希望各位在学期结束时有这样的感觉. 作为开场, 我们先大致了解一下什么是“机器学习”(machine learning).

回头看第一段话, 我们会发现这里涉及很多基于经验做出的预判. 例如, 为什么看到微湿路面、感到和风、看到晚霞, 就认为明天是好天呢? 这是因为在我们的生活经验中已经遇见过很多类似情况, 头一天观察到上述特征后, 第二天天气通常会很好. 为什么色泽青绿、根蒂蜷缩、敲声浊响, 就能判断出是正熟的好瓜? 因为我们吃过、看过很多西瓜, 所以基于色泽、根蒂、敲声这几个特征我们就可以做出相当好的判断. 类似的, 我们从以往的学习经验知道, 下足了工夫、弄清了概念、做好了作业, 自然会取得好成绩. 可以看出, 我们能做出有效的预判, 是因为我们已经积累了许多经验, 而通过对经验的利用, 就能对新情况做出有效的决策.

上面对经验的利用是靠我们人类自身完成的. 计算机能帮忙吗?

机器学习正是这样一门学科, 它致力于研究如何通过计算的手段, 利用经验来改善系统自身的性能. 在计算机系统中, “经验”通常以“数据”形式存在, 因此, 机器学习所研究的主要内容, 是关于在计算机上从数据中产生“模型”(model) 的算法, 即“学习算法”(learning algorithm). 有了学习算法, 我们把经验数据提供给它, 它就能基于这些数据产生模型; 在面对新的情况时(例如看到一个没剖开的西瓜), 模型会给我们提供相应的判断(例如好瓜). 如果说计算机科学是研究关于“算法”的学问, 那么类似的, 可以说机器学习是研究关于“学习算法”的学问.

[Mitchell, 1997] 给出了一个更形式化的定义: 假设用 P 来评估计算机程序在某任务类 T 上的性能, 若一个程序通过利用经验 E 在 T 中任务上获得了性能改善, 则我们就说关于 T 和 P, 该程序对 E 进行了学习.

本书用“模型”泛指从数据中学得的结果. 有文献用“模型”指全局性结果(例如一棵决策树), 而用“模式”指局部性结果(例如一条规则).

例如[Hand et al., 2001].

1.2 基本术语

要进行机器学习, 先要有数据. 假定我们收集了一批关于西瓜的数据, 例如(色泽=青绿; 根蒂=蜷缩; 敲声=浊响), (色泽=乌黑; 根蒂= 稍蜷; 敲声=沉闷), (色泽=浅白; 根蒂=硬挺; 敲声=清脆), ……, 每对括号内是一条记录, "="意思是"取值为".

这组记录的集合称为一个"数据集" (data set), 其中每条记录是关于一个事件或对象(这里是一个西瓜)的描述, 称为一个"示例" (instance) 或"样本" (sample). 反映事件或对象在某方面的表现或性质的事项, 例如"色泽""根蒂""敲声", 称为"属性" (attribute) 或"特征" (feature); 属性上的取值, 例如"青绿""乌黑", 称为"属性值" (attribute value). 属性张成的空间称为"属性空间" (attribute space)、"样本空间" (sample space)或"输入空间". 例如我们把"色泽""根蒂""敲声"作为三个坐标轴, 则它们张成一个用于描述西瓜的三维空间, 每个西瓜都可在这个空间中找到自己的坐标位置. 由于空间中的每个点对应一个坐标向量, 因此我们也把一个示例称为一个"特征向量" (feature vector).

有时整个数据集亦称一个"样本", 因为它可看作对样本空间的一个采样; 通过上下文可判断出"样本"是指单个示例还是数据集.

一般地, 令 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$ 表示包含 m 个示例的数据集, 每个示例由 d 个属性描述(例如上面的西瓜数据使用了 3 个属性), 则每个示例 $\boldsymbol{x}_i = (x_{i1}; x_{i2}; \ldots; x_{id})$ 是 d 维样本空间 $\mathcal{X}$ 中的一个向量, $\boldsymbol{x}_i \in \mathcal{X}$, 其中 x_{ij} 是 $\boldsymbol{x}_i$ 在第 j 个属性上的取值(例如上述第 3 个西瓜在第 2 个属性上的值是"硬挺"), d 称为样本 $\boldsymbol{x}_i$ 的"维数" (dimensionality).

从数据中学得模型的过程称为"学习" (learning)或"训练" (training), 这个过程通过执行某个学习算法来完成. 训练过程中使用的数据称为"训练数据" (training data), 其中每个样本称为一个"训练样本" (training sample), 训练样本组成的集合称为"训练集" (training set). 学得模型对应了关于数据的某种潜在的规律, 因此亦称"假设" (hypothesis); 这种潜在规律自身, 则称为"真相"或"真实" (ground-truth), 学习过程就是为了找出或逼近真相. 本书有时将模型称为"学习器" (learner), 可看作学习算法在给定数据和参数空间上的实例化.

训练样本亦称"训练示例" (training instance) 或"训练例".

学习算法通常有参数需设置, 使用不同的参数值和(或)训练数据, 将产生不同的结果.

如果希望学得一个能帮助我们判断没剖开的是不是"好瓜"的模型, 仅有前面的示例数据显然是不够的. 要建立这样的关于"预测" (prediction) 的模型, 我们需获得训练样本的"结果"信息, 例如"((色泽=青绿; 根蒂=蜷缩; 敲声=浊响), 好瓜)". 这里关于示例结果的信息, 例如"好瓜", 称为"标记" (label); 拥有了标记信息的示例, 则称为"样例" (example). 一般地, 用

将"label"译为"标记"而非"标签", 是考虑到英文中"label"既可用作名词、也可用作动词.

若将标记看作对象本身的一部分，则“样例”有时也称为“样本”.

$(\boldsymbol{x}_i, y_i)$ 表示第 i 个样例，其中 $y_i \in \mathcal{Y}$ 是示例 $\boldsymbol{x}_i$ 的标记，$\mathcal{Y}$ 是所有标记的集合，亦称“标记空间”(label space)或“输出空间”.

若我们欲预测的是离散值，例如“好瓜”“坏瓜”，此类学习任务称为“分类”(classification)；若欲预测的是连续值，例如西瓜成熟度 0.95、0.37，此类学习任务称为“回归”(regression). 对只涉及两个类别的“二分类”(binary classification)任务，通常称其中一个类为“正类”(positive class)，另一个类为“反类”(negative class)；涉及多个类别时，则称为“多分类”(multi-class classification)任务. 一般地，预测任务是希望通过对训练集 $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$ 进行学习，建立一个从输入空间 $\mathcal{X}$ 到输出空间 $\mathcal{Y}$ 的映射 $f : \mathcal{X} \mapsto \mathcal{Y}$. 对二分类任务，通常令 $\mathcal{Y} = \{-1, +1\}$ 或 $\{0, 1\}$；对多分类任务，$|\mathcal{Y}| > 2$；对回归任务，$\mathcal{Y} = \mathbb{R}$，$\mathbb{R}$ 为实数集.

亦称“负类”.

学得模型后，使用其进行预测的过程称为“测试”(testing)，被预测的样本称为“测试样本”(testing sample). 例如在学得 f 后，对测试例 $\boldsymbol{x}$，可得到其预测标记 $y = f(\boldsymbol{x})$.

亦称“测试示例”(testing instance)或“测试例”.

我们还可以对西瓜做“聚类”(clustering)，即将训练集中的西瓜分成若干组，每组称为一个“簇”(cluster)；这些自动形成的簇可能对应一些潜在的概念划分，例如“浅色瓜”“深色瓜”，甚至“本地瓜”“外地瓜”. 这样的学习过程有助于我们了解数据内在的规律，能为更深入地分析数据建立基础. 需说明的是，在聚类学习中，“浅色瓜”“本地瓜”这样的概念我们事先是不知道的，而且学习过程中使用的训练样本通常不拥有标记信息.

否则标记信息直接形成了簇划分；但也有例外情况，参见 13.6 节.

根据训练数据是否拥有标记信息，学习任务可大致划分为两大类：“监督学习”(supervised learning)和“无监督学习”(unsupervised learning)，分类和回归是前者的代表，而聚类则是后者的代表.

亦称“有导师学习”和“无导师学习”.

需注意的是，机器学习的目标是使学得的模型能很好地适用于“新样本”，而不是仅仅在训练样本上工作得很好；即便对聚类这样的无监督学习任务，我们也希望学得的簇划分能适用于没在训练集中出现的样本. 学得模型适用于新样本的能力，称为“泛化”(generalization)能力. 具有强泛化能力的模型能很好地适用于整个样本空间. 于是，尽管训练集通常只是样本空间的一个很小的采样，我们仍希望它能很好地反映出样本空间的特性，否则就很难期望在训练集上学得的模型能在整个样本空间上都工作得很好. 通常假设样本空间中全体样本服从一个未知“分布”(distribution) $\mathcal{D}$，我们获得的每个样本都是独立地从这个分布上采样获得的，即“独立同分布”(independent and identically distributed，简称 *i.i.d.*). 一般而言，训练样本越多，我们得到的关于 $\mathcal{D}$ 的信息

更确切地说，是“未见示例”(unseen instance).

现实任务中样本空间的规模通常很大(例如 20 个属性，每个属性有 10 个可能取值，则样本空间的规模已达 10^{20}).