

如何使用本书

——写在第十次印刷之际

本书2016年1月底出版,首印5000册一周内竟告售罄;此后8个月重印9次,累积72000册;先后登上亚马逊、京东、当当网等的计算机类畅销书榜首.出乎预料的销量和受欢迎程度,意味着本书读者已大大超出了预设的目标人群,这使作者隐隐产生了些许不安,感觉有必要说一说本书的立场,以及使用本书需注意的一些事项.因此,在第10次印刷之际草就本文.

首先,读者诸君务须注意,本书是一本教科书.

如本书“后记”所述,写作本书的主因是作者要开设“机器学习”课.根据作者的从教经验,若每堂课涉及页码过多,则不少同学由于选修多门功课,在课后或许难有兴趣和精力认真钻研阅读,教师也会因“包袱太重”而失去个人发挥的空间.因此,作为一学期课程的教材,本书篇幅进行了仔细考量:16章正文,每章6~7节,一般不超过25页.研究生课程若每学期18周,则除去习题和答疑时间,基本上每周讲授一章;本科生课程则可进度稍缓,一学期讲授9~10章.囿于此限,作者需对内容材料、以及材料讲述的程度进行取舍;否则若不分巨细,其篇幅可能令读者望而生畏.因此,读者不要指望本书是无所不包、“从入门到精通”的书籍.事实上,对机器学习这个发展极迅速、已变得非常广袤的学科领域,那样的书尚不存在;即便出现,也非数千页不止,不适于用作教科书.

第二,这是一本入门级教科书.

作者以为,入门阶段最需要的是理清基本概念、了解领域概貌.这好比人们到了一个陌生的地方,首先要去找张地图,大致弄清哪里是山、哪里有水、自己身在何处,然后才好到具体区域去探索.读者当然都希望所学“既广且深”,但在有限时间内必先有个折中.在入门阶段,“顾及细微”应该让位于“观其大略”,否则难免只见树木、不见森林.因此,作者试图通过化繁为简的讲述,使读者能在有限的篇幅中感受更多的、应该接触到的内容.一定程度上说,本书的主要目的就是为读者提供一张“初级地形图”,给初学者“指路”,而本书提供的这张“地形图”,其覆盖面与同类英文书籍相较不遑多让.

机器学习中存在多种学派可从其角度阐释其他学派的内容.作者以为,理解学派间的包容等价,在进阶之后对融汇贯通大有裨益,但在入门阶段,先看到各自的本原面貌更为重要.因为没有任何一个学派能完全“碾压”其他,而过早先入为主地强化某学派观念,对理解欣赏其他学派的妙处会埋下隐碍.因此,本书尽可能从材料的“原生态”出发讲述,仅在少数地方简略点出联系.需说明的是,作者试图以相近深度讲述主要内容.读者若感到在某些地方“意犹未尽”,或因作者以为,入门阶段到

此程度已可, 对其他内容的初窥优先于此处的进一步深究. 另外, 机器学习飞速发展, 很多新进展在学界尚无公论之前, 作者以为不适于写入入门级教科书中; 但为了不致于与学科前沿脱节, 本书也简略谈及一些本领域专家有初步共识的相对较新的内容.

第三, 这是一本面向理工科高年级本科生和研究生的教科书.

对前沿学科领域的学习, 必然需有基础知识作为先导. 为便于尽可能多的读者通过本书对机器学习有所了解, 作者已试图尽可能少地使用数学知识, 很多材料尽可能选择易于理解的方式讲述. 若读者感觉书中涉及的数学较深, 且自己仅需对机器学习做一般了解, 则不妨略过细节仅做概观, 否则建议对相关基础知识稍作复习以收全功. 囿于篇幅, 作者对许多材料尽可能述其精要、去其细冗, 所涉数学推导在紧要处给出阐释, 对理工科高年级同学稍下工夫就易自行弄清的繁冗则惜墨不赘.

读者不要指望通过读这本入门级教科书就能成为机器学习专家, 但书中各章分别给出了一些文献指引, 有兴趣的读者不妨据此进一步深造. 另外, 互联网时代之信息获取已相当便利, 读者可以容易地在网上找到机器学习中关于单个“知识点”的内容, 而信息搜索是理工科学生必备的本领, 只需知道自己在“找”什么, 就应该一定能找到材料. 根据本书提供的“地形图”, 读者若渴望对某个知识点进一步探究, “按图索骥”应无太大困难.

第四, 这本书不妨多读几遍.

初学机器学习易陷入一个误区: 以为机器学习是若干种算法(方法)的堆积, 熟练了“十大算法”或“二十大算法”一切即可迎刃而解, 于是将目光仅聚焦在具体算法推导和编程实现上; 待到实践发现效果不如人意, 则又转对机器学习发生怀疑. 须知, 书本上仅能展示有限的典型“套路”, 而现实世界任务千变万化, 以有限之套路应对无限之变化, 焉有不败! 现实中更多时候, 需依据任务特点对现有套路进行改造融通. 算法是“死”的, 思想才是“活”的. 欲行此道, 则务须把握算法背后的思想脉络, 无论创新科研还是应用实践, 皆以此为登堂入室之始. 本书在有限篇幅中侧重于斯, 冀望辅助读者奠造进一步精进的视野心法. 读者由本书初入门径后, 不妨搁书熟习“套路”, 数月后再阅, 于原不经意处或能有新得. 此外, 作者在一些角落融入了自己多年研究实践的些微心得, 虽仅只言片语, 但可能不易得之, 进阶读者阅之或可莞尔.

读者若仅对某几种具体机器学习技术的算法推导或工程实现感兴趣, 则本书可能不太适合; 若仅需机器学习算法“速查手册”, 则直接查看维基百科可能更便利一些.

作者自认才疏学浅, 对机器学习仅略知皮毛, 更兼时间和精力所限, 书中错谬之处甚多, 虽每次印刷均对错处或易误解处做勘误修订, 但仍在所难免, 若蒙读者诸君不吝告知, 将不胜感激.

周志华

2016年9月