机器学习从入门到小试牛刀

Ryan
1651372471@qq.com

2017 年 9 月 26 日[1]

[1]文档由解惑者学院创建于 2017 年 9 月 13 日，机器学习交流 QQ 群（总群）：194047142

前言

0.1 背景

人工智能的浪潮正席卷而来，而机器学习作为最热门的研究领域，为人工智能的各类奇怪的应用提供了强有力的支撑。从早期的规则系统，到统计学习方法逐渐显现其威力；从感知机、SVM到提升方法，再到近些年流行的深度学习、强化学习，机器学习的理论越来越丰富和高级。人脸识别、语音识别、无人驾驶逐渐为大众所熟知，AlphaGo 的横空出世震惊了整个世界。与此同时，不仅互联网行业产生了大量的机器学习相关岗位的需求，其它行业也产生了大量的岗位需求。金融领域的智能投顾、量化交易以及反金融欺诈等等应用大量引入机器学习的技术，通信行业开始引入机器学习技术去改进原有的算法，医疗、法律、生物等等领域都逐渐引入机器学习的技术去改进原有的产品和技术。人工智能已经成为国家的战略，未来机器学习技术将会像互联网一样，慢慢渗透到各行各业，对各类人才的需求将会越来越大。

行业发展产生了大量的岗位需求，与此形成鲜明对比的是，整个行业真正有能力和技术的人很少很少，造成只要懂一些机器学习的技术，找工作随便挑的现象。应届毕业生拿到 30W 的年薪已经是白菜价，四五十万也不少见。各行各业的人都想从事人工智能领域的相关工作，但各种奇怪的名词和媒体夸张的报道让不熟悉这个领域的互联网从业者有些不知所措，究竟入门有没有那么难?本课程将为大家揭开机器学习神秘的面纱。

0.2 关于本课程

其实学习机器学习很简单！只要你能够找到对的学习方法。本课程讲师 Ryan 在 BAT 从事多年的机器学习相关的工作，从在校园里研究信息检索、数据挖掘和自然语言处理相关领域的知识，到加入一线互联网名企参与、主导大规模的机器学习项目，具有丰富的行业技术背景。与此同时，在企业内部给同事分享机器学习相关的技术，给想进入这个行业的同学做指导，深知当年自己走过的弯路对于新入行的人来说，不需要再走一遍，**有捷径！**

本课程由讲师日常工作中做的分享资料整理而来，经由反复给初学者讲解，再加上空余时间不断的丰富和改善，历时一年多。**课程面向的对象是还未入门、刚入门的在校学生以及已经工作但是机器学习基础薄弱的同学**。如果你想转行，或者想引入机器学习到现有的工作中，亦或是对机器学习的很多知识掌握的不牢，那么本课程非常适合你!

本课程具有以下特点:

- 精选了对于入门者来说最重要的那部分知识，并且具有很强的逻辑性。掌握了这部分知识后，再学习机器学习的其它知识，都是水到渠成的事情，**so easy!**
- 在授课的过程中，对于找工作面试过程中可能会考察的知识点会着重讲解，并会介绍面试官想要考察的能力以及希望得到的解答。
- 提供独有的垃圾短信数据集，供课程中对学到的知识点进行实战和验证，课程也会提供示例

代码供学员学习。此外，**课程中会组织比赛，对于比赛获得较好名次的同学，我们会颁发奖金！**

- 上课直播过程中，授课老师会对学员提出的问题做实时的解答。课后，学员提出的问题，也会有专门的老师给予解答。
- 对于优秀的学员，可以提供大公司实习或者工作的内推机会。

0.3 课程详细信息

课程一共 32 个课时，每个课时包括两个部分：（1）课程知识讲解；（2）学员答疑。每周两次课，每次两个课时，分别是**周三和周六晚上 19：30**，课程时长为 2 个月。如果个别情况下，讲师时间需要调整，会提前一周通知。对于错过讲课时间的同学，我们**会提供录制好的课程视频**，供学员反复观看。**课程收费为 1288￥，现在开始接受报名。**

课程正式授课时间为：**2017 年 10 月 14 日周六 19：30**。

课程会提供垃圾短信数据集，供学员实战。另外，我们会保留一小部分数据集（此部分不对外公开）用于评估学员的模型结果好坏，最终学员按照评分排序，名次前列的同学将会得到课业奖金，具体情况如下表1。**报名成功后，我们就会提供数据集供该学员提前实验，增加其获得奖金的可能性。**

表 1: 课业奖金明细

名次	奖金
第 1 名	1288￥
第 2 名	644￥
第 3 名	200￥

课程最后一次课结束后一周，会对学员提交的结果进行评估，最终选取**前 5 名**进入答辩环节。答辩环节形式为：评估结果公示后一周内，奖金候选人需要提交自己的项目材料，材料包括代码、文档说明、答辩 PPT 等。课程所有学员学习和评价候选人的方案，投票决定最终名次。为了防止意外情况出现，**Ryan 老师具有一个投票权，该投票权相当于是 5 个学员的投票数量，Ryan 老师也可以放弃使用该投票权。**

此外，为鼓励学员主动积极的学习，并为课程的其它学员做出积极的贡献，我们决定对这部分积极的学员进行奖励。奖金明细见表格2。名次决定方式如下：首先由学员主动报名，产生候选人（如果报名人数过多，则由 Ryran 老师进行删选），最终名次由全体学员投票产生。

表 2: 课程贡献奖金

名次	奖金
第 1 名	300￥
第 2 名	200￥
第 3 名	100￥

其它优惠及奖励活动：

- 已经报名的学员，推荐其他未报名的同学报名，每成功一名，奖励 50￥（请注意告知报名的学员报名时，注明是你推荐的，并填写你的 QQ 号，否则奖金可能无法兑现。报名结束后，会统一统计，并做奖金的发放）。

- 未报名的同学，如果由已经报名的学员推荐，学费优惠 50￥（报名时需要指定推荐你报名的学员的 QQ 号）。

课程咨询及报名，可以联系以下老师：

- Ryan 老师：QQ 1651372471，微信 Cheung9507

图 1: Ryan 的微信号

- 唐老师：QQ 1935699063，微信 jiehuozhe-tang

图 2: 唐老师的微信号

目录

第一章　课前准备

1.1　学习环境准备

1.1.1　python

目前，Python 有两个版本，一个是 2.x 版，一个是 3.x 版，这两个版本是不兼容的，因为现在 Python 正在朝着 3.x 版本进化，在进化过程中，大量的针对 2.x 版本的代码要修改后才能运行，所以，目前有许多第三方库还暂时无法在 3.x 上使用。

我的建议是，版本的选择主要取决于做哪方面的工作，如果做 web 开发相关，建议使用 Python3；但是如果是做科学计算、机器学习方面的工作，建议选择 Python2（Python2.6 以上）。Python3 相比 Python2 进行了许多改进，性能更好，值得推荐，但是鉴于目前有少部分库仅支持 Python2.x，另外，Python 也提供了 2to3 的转换工具，可以比较简单的将 Python2 代码迁移到 Python3。在绝大部分 Linux 系统上默认自带的 Python 版本是 2.x 版本，Centos6 以下默认自带 Python2.6 版本，并且在 Mac 上自带的也是 Python2.x 版本。

为了保证你的程序能用到大量的第三方库，本课程仍以 2.x 版本为基础，确切地说，是 2.7 版本。请确保你的电脑上安装的 Python 版本是 2.7.x（我们用的版本是 2.7.12），这样，你才能无痛学习整个课程。

对于刚入门 python 的同学来说，建议学习一下廖雪峰写的教程：Python 2.7 教程。

1.1.2　numpy

numpy 是一个用 python 实现的科学计算包，是 python 的一种开源的数值计算扩展，这种工具可用来存储和处理大型矩阵。建议在上课前安装好此 python 工具包。

官方链接：`http://www.numpy.org`

1.1.3　pandas

pandas 是基于 numpy 的一种工具，该工具是为了解决数据分析任务而创建的。pandas 纳入了大量库和一些标准的数据模型，提供了高效地操作大型数据集所需的工具。pandas 提供了大量能使我们快速便捷地处理数据的函数和方法。建议阅读：

英文版：10 Minutes to pandas

翻译版：十分钟搞定 pandas

1.1.4　scikit-learn

scikit-learn 是一个常用的 python 实现的机器学习工具包。建议在上课前安装好此 python 工具包。

官方链接：`http://scikit-learn.org/stable/`

1.1.5 TensorFlow

Google 出品的深度学习工具，也是最常用的深度学习工具之一。建议在上课前安装好此工具。

官方链接：https://www.tensorflow.org

1.1.6 cityhash

大规模 LR 模型中，很关键的一个功能就是需要对原始特征或者组合类特征进行 hash，cityhash 是一个非常优秀的开源工具，有兴趣的同学可以提前用用这个工具。

代码地址：https://github.com/google/cityhash

1.2 提前预习资料

1.2.1 梯度下降法

梯度下降法是机器学习领域最常见和常用的优化算法，在上课前，希望大家能够提前学习这个算法。下面是一些学习材料（这些材料几乎包含了所有的梯度下降法的变种，其难度和覆盖面远远超过了本课程，有余力的同学可以学完所有的资料，学不完也没关系，课程中能用到的只是一小部分）：

- An overview of gradient descent optimization algorithms
- An overview of gradient descent optimization algorithms
- Introduction to Gradient Descent Algorithm (along with variants) in Machine Learning