

机器学习从入门到小试牛刀

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！只要你能找到对的学习方法。本课程讲师在 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。

- 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

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 ¥

表 2: 课程贡献奖金

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

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。

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 ¥

课程咨询及报名，可以联系以下老师：

- 微信 jhzkf05

QQ group : 619381339



目录

前言	i
0.1 背景	i
0.2 关于本课程	i
0.3 课程详细信息	ii
第一章 课前准备	1
1.1 学习环境准备	1
1.1.1 python	1
1.1.2 numpy	1
1.1.3 pandas	1
1.1.4 scikit-learn	1
1.1.5 TensorFlow	2
1.1.6 cityhash	2
1.2 提前预习资料	2
1.2.1 梯度下降法	2
第二章 机器学习简介	3
2.1 为什么要用机器学习	3
2.2 机器学习的发展历史和应用现状	6
2.3 机器学习的学习路径和就业现状	11
2.4 机器学习任务的类型	15
2.5 本课程学习目标	19
第三章 python 入门	21
3.1 python 简介	21
3.2 安装环境和版本	23
3.3 基本语法及常用模块	26
3.4 高级工具	40
第四章 Logistic Regression	45
4.1 垃圾短信数据集（课程特有数据集）	45
4.2 分类问题	48
4.3 特征的概念	51
4.4 sigmoid 函数	53
4.5 模型优化目标	55
4.6 模型评估指标	58
第五章 梯度下降法	61
5.1 梯度的概念	61
5.2 梯度下降法	63
5.3 随机梯度下降法	71
5.4 mini-batch 梯度下降法	74

第六章 LR 模型的训练算法	77
6.1 常用的目标函数	77
6.2 LR 模型训练算法的推导	81
第七章 实践：LR 模型的 python 实现	85
7.1 梯度下降法的实现	85
7.2 随机梯度下降法的实现	93
7.3 mini-batch 梯度下降法的实现	105
第八章 python 进阶-高级工具的使用	119
8.1 pandas	119
8.2 numpy	131
8.3 优化 LR 实现	153
第九章 python 进阶-机器学习工具的使用	157
9.1 机器学习工具 sklearn	157
第十章 实战：垃圾短信识别小试牛刀-Part 1	179
10.1 机器学习解决问题的流程	179
10.2 训练集、验证集和测试集	188
10.3 特征工程	193
10.4 分词工具 jieba	197
第十一章 实战：垃圾短信识别小试牛刀-Part 2	209
11.1 文本编码处理经验谈	209
11.2 文本的特征表示	212
11.3 使用 LR 解决垃圾短信识别问题	215
11.4 训练误差与测试误差	223
11.5 过拟合与模型选择	229
第十二章 BAT 如何应用 LR 解决实际工程问题	235
12.1 工业界实际项目如何组织开发	235
12.2 千亿特征从何而来	240
12.3 Feature Hash 如何实现	243
12.4 为什么工业界选择使用大规模的离散特征	248
12.5 调参工程师和特征工程师从何而来	256
第十三章 多分类	259
13.1 使用 LR 解决多分类问题	259
13.2 softmax	264
13.3 LR 和 softmax 解决多分类问题的区别	272
第十四章 朴素贝叶斯	275
14.1 概率论基础	275
14.2 朴素贝叶斯法的学习与分类	284
14.3 朴素贝叶斯法的参数估计	289

第十五章 实战：垃圾短信识别小试牛刀-Part 3	297
15.1 问题分析	297
15.2 朴素贝叶斯的实现	299
15.3 效果评估与优化	305
第十六章 机器学习模型的评价指标	309
16.1 准确率	309
16.2 召回率	310
16.3 F 值	311
16.4 混淆矩阵	312
16.5 ROC 曲线和 AUC 值	314
16.6 NDCG	320
第十七章 决策树基础	323
17.1 决策树模型与学习	323
17.2 信息论基础	326
17.3 特征选择	333
第十八章 决策树生成及剪枝	339
18.1 决策树的生成	339
18.2 决策树的剪枝	344
18.3 CART 算法	354
第十九章 实战：垃圾短信识别小试牛刀-Part 4	359
19.1 问题分析	359
19.2 sklearn 中的决策树	360
19.3 模型训练及参数选择	363
19.4 效果评估与优化	369
第二十章 神经网络	375
20.1 神经网络的发展历史及应用现状	375
20.2 感知机模型	384
20.3 感知机学习策略	387
20.4 前馈神经网络	390
第二十一章 反向传播算法	399
21.1 BP 算法的推导	399
21.2 梯度正确性的验证方法	406
21.3 BP 算法的 python 实现	409
第二十二章 TensorFlow 入门	417
22.1 TensorFlow 是什么	417
22.2 计算图	422
22.3 Tensor	425
22.4 TensorFlow 实现神经网络	428

第二十三章 TensorFlow 解决手写数字识别问题	431
23.1 深度学习与深层神经网络	431
23.2 MNIST 数字识别问题	434
23.3 神经网络的优化	439
第二十四章 实战：垃圾短信识别小试牛刀-Part 5	443
24.1 问题分析	443
24.2 搭建网络结构	444
24.3 模型训练及参数选择	446
24.4 效果评估与优化	450
第二十五章 word2vec-Part 1	457
25.1 统计语言模型	457
25.2 词向量简介	460
25.3 word2vec 是什么	462
25.4 word2vec 示例	466
第二十六章 word2vec-Part 2	471
26.1 Bengio 神经网络语言模型	471
26.2 CBOW 模型详解	475
26.3 Hierarchical Softmax	485
第二十七章 word2vec-Part 3	491
27.1 Negative Sampling	491
27.2 Skip-Gram 模型	497
第二十八章 TensorFlow 进阶	509
28.1 TensorFlow 实现 word2vec	509
第二十九章 实战：垃圾短信识别小试牛刀-Part 6	523
29.1 Embedding 是什么	523
29.2 Embedding 的应用	525
29.3 Embedding 向量的使用	532
29.4 模型的训练和评估	537
第三十章 RNN	543
30.1 RNN 的网络结构	543
30.2 BPTT 算法	546
第三十一章 LSTM	557
31.1 LSTM 网络结构深度剖析	557
31.2 TensorFlow 实现 RNN 模型	572
第三十二章 Encoder-Decoder 框架	579
32.1 Encoder-Decoder 框架介绍	579
32.2 NMT 模型	583
32.3 Attention 模型	590
第三十三章 课程总结及展望	599
33.1 深度学习其它 topic	599
33.2 课程总结	609
33.3 行业前沿	614
33.4 进阶学习建议	619

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](#)