如何自学人工智能

博格坎普

1. 引言

本文的初衷是记录下本人是如何自学数据分析和人工智能的，希望对想转行的兄弟姐妹有所帮助，尽量不要去走我走过的弯路。毕竟转行是一个痛苦的过程，需要勇气与毅力，希望正在转行中的兄弟姐妹要坚持到底，遇到孤独寂寞或失望无助时也不要轻言放弃，因为我们已经没有回头路了!

1.1. **目标.** 首先目标是相当明确的，那就是通过不懈地努力找到一份数据分析相关的工作。

1.2. **心理.** 当遇到挫折的时候，一定要学会坚强。当遇到外部环境干扰的时候，一定要学会静心。当内心产生波动的时候，一定要学会沉着冷静。虽然会有辗转反侧失眠的夜晚，但是当太阳升起的时候，我会面带微笑继续刷题。对那些过去的过去，学会放下，不要再纠结于过往而浪费时间精力了。至少我们还有诗和远方，和往事干杯，爱过，不后悔!

1.3. **方法.** 充分利用各种网络资源，尽量在最短的时间内掌握数据分析相关的知识，瞄准特定公司的需求，量身定制增强自己的技术，最有效地投简历得到面试机会，并最终拿到工作机会。学习过程中应该以做具体项目为中心，加强简历的制作修改，侧重应用而非理论，以商业应用为导向，技术务必结合业务。

2. 编程基础

作为脚本语言，Python 是比较容易入门的，而且有很多成熟的库，所以是新手首选的数据分析编程语言。某些公司也可能要求用 R，也很容易学，可视化方面比较强大。时间允许的话，学习基本机器学习算法，并自己编程实现各种类和库。请充分利用 Github 上的资源。

2.1. **Python.** 对 Python 语言的基础知识必须掌握，对数据分析来说不需要太高级的技巧，但是基本的脚本语言特性还是要知道的，简单的编个程序实现日常的数据应用功能就行。首先 Python 语言基础的书籍，网络资源很多，以下是我读过的几本书。

- 《Python Crash Course》[4] 是本畅销书。个人认为此书是鸡肋，原因是前面的内容比较简单，后面的三个项目选择有严重的问题。如果想学数据分析的话，不需要知道如何上手做游戏或网络开发，但是可视化的项目又太简单。总的来说，我只能给它一颗星（*）。

- 《Dive into Python 3》[7] 两颗星（**）。此书比较中庸，可能是由于出版时间比较久了，内容选取上比较一般。从本书中可以学到一些基础知识，比如正则表达式（Regular Expressions），迭代器（Iterators），单元测试（Unit Testing），各种常见文件格式的操作（I/O）等。

Date: April 22, 2018.

- 《Effective Python》[9] 三颗星（***）。有一定 Python 基础后，此书可以当作高手进阶读物，它秉持了 Effective 系列书籍的特色。从本书中可以学到类和元类（classes and metaclasses），进程线程和同步（concurrancy），内建模块（Built-in modules）等相关知识。

学习的最好方法就是通过做项目，以下是几个简单例子：

(1) 信息系统（类 + 函数 + 数据库）
(2) 棋牌类游戏（Blackjack, 21 点）
(3) 可视化 (经济，商业数据)

2.2. **数据分析.** 有了一定的 Python 基础之后，就可以学习数据分析了，主要是学习常用的 Python 库，比如 Numpy, Pandas, Matplotlib，Seaborn 等。目标是掌握基本的数据采集，清洗与整理，输入与输出，可视化方法等，为进一步的机器学习准备好素材。

Numpy 建立在向量化的思想上，采用 C 语言的底层实现，可以快速计算类似多维数组（ndarray）的数据类型，另一个常用的功能是随机数的生成。Pandas 建立在 DataFrame（常翻译为数据框）数据结构的基础上，具有强大的数据处理能力。Matplotlib 具有类似于 Matlab 的绘图功能，seaborn 是其一个进化版本，提供更多的可视化选项。

- 《Python for Data Analysis》[5] 四颗星（****）。本书对常用库 Numpy, Pandas, Matplotlib/Seaborn 做了比较详细的介绍，是一本不错的参考书。本书也介绍了时间序列（Time Series）相关的内容，最后一章给了一些数据分析的综合应用。

- 《Python Data Science Handbook》[10] 四颗星 (****)。本书在 Pandas 和 Matplotlib 方面和 [5] 有所重叠，但是也对其做了部分的补充。本书的特色是最后一章用比较简洁的语言介绍了经典机器学习 (Machine Learning) 的基础知识及应用。

3. 人工智能

具备了 Numpy, Scipy, Matplotlib 等的准备知识之后，就可以开始学习使用 scikit-learn 进行机器学习（Machine Learning）了。Scipy 是另一个强大的算法库和数学工具包，其中有线性代数，最优化等科学计算的算法实现。Scikit-learn (sklearn) 是建立在 Numpy 和 Scipy 基础上的一个开源机器学习库，它整合了很多经典机器学习算法的实现，提供了非常方便实用的 API （Application Programming Interface）。更高级的算法，比如神经网络，不包括在经典机器学习算法的范畴之内。

- Machine Learning course CS229，三颗星（***）。这是 Stanford 大学开设的机器学习课程，最早由 Andrew Ng 讲授，其视频和课堂笔记可以下载。本课程注重算法的理论基础，工程应用稍有提及，可作为机器学习的理论准备知识来学习。此课程其实是门经典机器学习算法课，其中的数学主要涉及线性代数和概率论。

- 《Introduction to Machine Learning with Python》[6] 四颗星（****）。本书对各种常见机器学习算法的应用，scikit-learn 的实现作了详细的介绍，是一本不错的参考书。最后一章还简单介绍了文本及自然语言处理（NLP）的基础知识。

掌握了经典机器学习算法和实现之后，就可以尝试各种人工智能的简单应用了。举例，

(1) Scikit-learn 自带的数据集上的分类，回归，聚类等练习。比如，Boston 房价预测，糖尿病可能性的分类问题等。
(2) Kaggle compititions, Kaggle 上有很多数据集，给新手提供了各种实战的练习机会。比如，经典的泰坦尼克号幸存者问题，自行车共享预测问题，工资预测等等。

4. 深度学习

神经网络 (Neural network) 和深度学习（Deep learning）是最近几年比较热门的人工智能方向，主要在视频或图像处理，机器视觉（computer vision），自然语言处理（NLP）等领域有诸多应用。常用的 Python 库包括 TensorFlow 和 Keras 等，它们利用显卡上的图像处理器（GPU）比多核中央处理器（CPU) 有更高效的运算速度与结果。相对于经典机器学习算法，深度学习对特定的复杂问题，比如图像处理，有非常大的优势。但深度学习的理论还在不断的更新发展中，很多神经网络算法在标准库里面是没有实现的。

- 《Deep Learning》[3] 三颗星（***）。本书对深度学习的基础知识做了非常详细的介绍和讨论，是很好的一本理论参考书。

- 《Python Machine Learning》[8] 四颗星（****）。本书内容非常丰富，不但包括了经典机器学习算法及其实现，而且还包括了常用神经网络算法及其实现。每个重要算法，还有相应的理论基础作为补充。最后部分介绍了卷积神经网络（CNN）在图像处理上的简单应用，以及循环神经网络（RNN）在自然语言处理上的简单应用，包括了 Tensorflow 和 Keras 的实现。

- 《Hands-On Machine Learning with Scikit-Learn and TensorFlow》[2] 四颗星（****）。本书第二部分对 Tensorflow 的实现做了具体的介绍。

References

[1] David Beazley and Brian K. Jones. *Python Cookbook, 3rd Ed.* O'Reilly Media, 2013.
[2] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2017.
[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, 2016.
[4] Eric Matthes. *Python Crash Course: A Hands-On, Project-Based Introduction to Programming*. No Starch Press, 2015.
[5] Wes McKinney. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, 2nd Ed*. O'Reilly Media, 2017.
[6] Andreas Müller and Sarah Guido. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, 2016.
[7] Mark Pilgrim. *Dive into Python 3*. Apress, 2009.
[8] Sebastian Raschka and Vahid Mirjalili. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, 2nd Ed*. Packt Publishing, 2017.
[9] Brett Slatkin. *Effective Python: 59 Specific Ways to Write Better Python*. Addison-Wesley Professional, 2015.
[10] Jake VanderPlas. *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media, 2016.

E-mail address: danli091981@gmail.com