Central University of Finance and Economics

中央财经大学

数据挖掘课程

个人思考

吴宇翀

2017310836

WUYUCHONG.COM

指导老师：马景义

2020 年 5 月 23 日

目录

1 方向展望

在知乎对机器学习[1] 的讨论中，共同点是大家都认为，大致有三种不同的观点：

restart[2] 认为**机器学习专家不能停留在应用层面，而应该具备很强的工程能力，才能将算法落地**。

微调[3] 则认为**对于绝大部分人而言，努力成为交叉领域的熟手，远比成为计算机科学家要现实且有意义**。她认为工业界需要能将模型应用于专业领域，对大部分人来说，需要有自己精通的某个业务领域，而对于机器学习，只需要做到会应用即可。

紫衫[4] 认为，专精一个领域非常重要，能将**学术模型转化为工业模型更加重要**。在很多领域特别是高危领域，模型的优化十分不易，要求很强的学术能力，同时工程能力也不可或缺。同时，他特别提到，在工业界的经验非常重要，因为相比学术界，工业界的数据吞吐量、业务规模更大，许多问题更加复杂，能够获得的经验更多。

2 个人思考

我觉得这些观点都非常有道理，不过的确对大多数人来说，可能最终都不会从事十分顶尖的算法研究，应用层面可能还是大多数人所看中的。对我来说，也是这样，更加倾向于将机器学习用于其它领域，如商业和金融方面，而非专业地去研究算法，如机器学习在**金融风控和衍生品领域**的应用对我非常有吸引力。

[1] 未来 3~5 年内，哪个方向的机器学习人才最紧缺？https://www.zhihu.com/question/63883507

[2] https://www.zhihu.com/people/bo-ge-si-bao

[3] https://www.zhihu.com/people/breaknever

[4] https://www.zhihu.com/people/zi-shan-43

3 能力欠缺方面

相比于学术界，工业界可能不会太过于追求算法模型的效果，而是更多地在意模型的稳定性，更加强调将模型落地。所以更加具有工程思维，全面地去考虑实际环境中可能遇到的各种兼容性等问题，同时自己能够写一些测试用例，这些**工程能力都是我目前所欠缺的**。许多的 Machine Learning Software Engineer 都是计算机出身，而作为一个非计算机出身的统计学学生, 这工程能力方面的确是我不易跨过的一道槛，也是我期末作业选择进行程序编写的原因之一。

4 未来计划

大三暑期开始，我将到滴滴出行数据科学部实习一年，希望能在较成熟的互联网公司开阔眼界，对数据挖掘、机器学习上有进一步的理解，同时锻炼自己的工程能力。同时，我希望能够通过此次实习，去确定数据科学、机器学习方向是否是我未来希望从事的，如果这真的是我所热爱的领域，那么未来很可能会申请到美国留学修两年数据科学硕士。

5 合作邀请

之前您说 scikit-learn 上有很多不足的地方，我们可以做些改进，我觉得这是一个为开放社区做贡献的好机会。我想，如果可以的话，我们可以去完善您觉得有问题的算法，然后我们可以在 github 上提 issue，如果可以顺利通过测试，那么能够成为 scikit-learn 的 contributors 中的一员，我想相比于其它选题，这是更加有趣且有贡献的选择，虽然许多的 issue 最终没有得到采纳或无法通过严苛的测试，但这至少是一件非常有挑战性的事情。

同时，如果您有一些新颖的算法，或是一些算法改进，我们可以一起将它实现出来，不一定要在 scikit-learn 上发布，我们也可以写一个网站作好 SEO 直接放在搜索引擎上进行推广。鉴于我的知识水平有限、眼界不够开阔，还请您在方向和技术细节上多多指导。希望能和老师一起合作，共同向社区做一点微小的贡献！

6 致谢

最后，非常感谢马景义老师提供这次思考总结的机会。同时也感谢老师这一学期**数据挖掘**课程的精心授课和答疑，希望有机会能继续向老师学习，跟老师一起做项目。