面试机器学习、大数据岗位时遇到的各种问题

2016-08-10 要学习更多点这→ 数据挖掘DW

作者：@太极儒

自己的专业方向是机器学习、数据挖掘，就业意向是互联网行业与本专业相关的工作岗位。各个企业对这类岗位的命名可能有所不同，比如数据挖掘/自然语言处理/机器学习算法工程师，或简称算法工程师，还有的称为搜索/推荐算法工程师，甚至有的并入后台工程师的范畴，视岗位具体要求而定。

机器学习、大数据相关岗位的职责

自己参与面试的提供算法岗位的公司有 BAT、小米、360、飞维美地、宜信、猿题库 等，根据业务的不同，岗位职责大概分为：

- **平台搭建类**

> 数据计算平台搭建，基础算法实现，当然，要求支持**大样本量、高维度数据**，所以可能还需要底层开发、并行计算、分布式计算等方面的知识；

- **算法研究类**

> **文本挖掘**，如领域知识图谱构建、垃圾短信过滤等；
>
> **推荐**，广告推荐、APP 推荐、题目推荐、新闻推荐等；
>
> **排序**，搜索结果排序、广告排序等；
>
> 广告投放效果分析；
>
> 互联网信用评价；
>
> 图像识别、理解。

- **数据挖掘类**

> **商业智能**，如统计报表；

> **用户体验分析**，预测流失用户。

以上是根据本人求职季有限的接触所做的总结。有的应用方向比较成熟，业界有足够的技术积累，比如搜索、推荐，也有的方向还有很多开放性问题等待探索，比如互联网金融、互联网教育。在面试的过程中，一方面要尽力向企业展现自己的能力，另一方面也是在增进对行业发展现状与未来趋势的理解，特别是可以**从一些刚起步的企业和团队那里，了解到一些有价值的一手问题。**

以下首先介绍面试中遇到的一些真实问题，然后谈一谈答题和面试准备上的建议。

面试问题

> 1. 你在研究/项目/实习经历中主要用过哪些机器学习/数据挖掘的算法？
> 2. 你熟悉的机器学习/数据挖掘算法主要有哪些？
> 3. 你用过哪些机器学习/数据挖掘工具或框架？

- **基础知识**

> 无监督和有监督算法的区别？
>
> SVM 的推导，特性？多分类怎么处理？
>
> LR 的推导，特性？
>
> 决策树的特性？
>
> SVM、LR、决策树的对比？
>
> GBDT 和 决策森林 的区别？
>
> 如何判断函数凸或非凸？
>
> 解释对偶的概念。
>
> 如何进行特征选择？
>
> 为什么会产生过拟合，有哪些方法可以预防或克服过拟合？
>
> 介绍卷积神经网络，和 DBN 有什么区别？
>
> 采用 EM 算法求解的模型有哪些，为什么不用牛顿法或梯度下降法？
>
> 用 EM 算法推导解释 Kmeans。
>
> 用过哪些聚类算法，解释密度聚类算法。
>
> 聚类算法中的距离度量有哪些？
>
> 如何进行实体识别？

> 解释贝叶斯公式和朴素贝叶斯分类。
>
> 写一个 Hadoop 版本的 wordcount。
>
> ……

- **开放问题**

> 给你公司内部群组的聊天记录，怎样区分出主管和员工？
>
> 如何评估网站内容的真实性（针对代刷、作弊类）？
>
> 深度学习在推荐系统上可能有怎样的发挥？
>
> 路段平均车速反映了路况，在道路上布控采集车辆速度，如何对路况做出合理估计？采集数据中的异常值如何处理？
>
> 如何根据语料计算两个词词义的相似度？
>
> 在百度贴吧里发布 APP 广告，问推荐策略？
>
> 如何判断自己实现的 LR、Kmeans 算法是否正确？
>
> 100亿数字，怎么统计前100大的？
>
> - ……

答题思路

- **用过什么算法？**

> 最好是在**项目/实习的大数据场景**里用过，比如推荐里用过 CF、LR，分类里用过 SVM、GBDT；
>
> 一般用法是什么，是不是自己实现的，有什么比较知名的实现，使用过程中**踩过哪些坑**；
>
> 优缺点分析。

- **熟悉的算法有哪些？**

> 基础算法要多说，其它算法要挑熟悉程度高的说，不光列举算法，也适当说说应用场合；
>
> 面试官和你的研究方向可能不匹配，不过在基础算法上你们还是有很多共同语言的，你说得太高大上可能效果并不好，一方面面试官还是要问基础的，另一方面一旦面试官突发奇想让你给他讲解高大上的内容，而你只是泛泛的了解，那就傻叉了。

- **用过哪些框架/算法包？**

主流的分布式框架如 Hadoop，Spark，Graphlab，Parameter Server 等择一或多使用了解；

通用算法包，如 mahout，scikit，weka 等；

专用算法包，如 opencv，theano，torch7，ICTCLAS 等。

- **基础知识**

对知识进行结构化整理，比如撰写自己的 cheet sheet，我觉得**面试是在有限时间内向面试官输出自己知识的过程**，如果仅仅是在面试现场才开始调动知识、组织表达，总还是不如系统的梳理准备；

从面试官的角度多问自己一些问题，通过查找资料总结出全面的解答，比如如何预防或克服过拟合。

产生背景，适用场合（数据规模，特征维度，是否有 Online 算法，离散/连续特征处理等角度）；

原理推导（最大间隔，软间隔，对偶）；

求解方法（随机梯度下降、拟牛顿法等优化算法）；

优缺点，相关改进；

和其他基本方法的对比；

个人感觉高频话题是 SVM、LR、决策树（决策森林）和聚类算法，要重点准备；

算法要从以下几个方面来掌握:

- 产生背景，适用场合（数据规模，特征维度，是否有 Online 算法，离散/连续特征处理等角度）；
- 原理推导（最大间隔，软间隔，对偶）；
- 求解方法（随机梯度下降、拟牛顿法等优化算法）；
- 优缺点，相关改进；
- 和其他基本方法的对比；

不能停留在能看懂的程度，还要:

- 对知识进行结构化整理，比如撰写自己的 cheet sheet，我觉得**面试是在有限时间内向面试官输出自己知识的过程**，如果仅仅是在面试现场才开始调动知识、组织表达，总还是不如系统的梳理准备；
- 从面试官的角度多问自己一些问题，通过查找资料总结出全面的解答，比如如何预防或克服过拟合。

- **开放问题**

> 由于问题具有综合性和开放性，所以不仅仅考察对算法的了解，还需要足够的实战经验作基础；
>
> **先不要考虑完善性或可实现性**，调动你的一切知识储备和经验储备去设计，有多少说多少，想到什么说什么，方案都是在你和面试官讨论的过程里逐步完善的，不过面试官有两种风格：引导你思考考虑不周之处 or 指责你没有考虑到某些情况，遇到后者的话还请注意**灵活调整答题策略**;
>
> 和同学朋友开展讨论，可以从上一节列出的问题开始。

准备建议

1. **基础算法复习两条线**
    - 材料阅读 包括经典教材（比如 PRML，模式分类）、网上系列博客，系统梳理基础算法知识；
    - 面试反馈 **面试过程中会让你发现自己的薄弱环节和知识盲区，把这些问题记录下来，在下一次面试前搞懂搞透。**
2. 除算法知识，还应适当掌握一些系统架构方面的知识，可以从网上分享的阿里、京东、新浪微博等的架构介绍 PPT 入手，也可以从 Hadoop、Spark 等的设计实现切入。
3. 如果真的是以就业为导向就要在平时注意实战经验的积累，在科研项目、实习、比赛（Kaggle，阿里大数据竞赛等）中摸清算法特性、熟悉相关工具与模块的使用。

总结

如今，好多机器学习、数据挖掘的知识都逐渐成为常识，要想在竞争中脱颖而出，就必须做到

> - 保持学习热情，关心热点；
> - 深入学习，会用，也要理解；
> - 在实战中历练总结；
> - 积极参加学术界、业界的讲座分享，向牛人学习，与他人讨论。


50000+
数据分析爱好者 选择来这里。
他们每天如何利用这里
如何学习？代码如何实现？
在学习什么？python
有什么操作案例？
R语言 hadoop
如何入门？
如何规划？
这里能有什么，不会有人跟你直说。
数据分析师 微信学习公众号：datadw
长按此图片 然后点击 识别二维码 关注