内部资料整理by寒小阳(hanxiaoyang.ml@gmail.com)，请勿外传

1、面试知识板块：

1. 代码算法

   a. 基本算法（如快排等，需要熟练掌握）

   b. 剑指Offer（面试经常出相似的题）

   c. LeetCode（剑指Offer的补充，增强动手能力）

2. 机器学习

   a. 李航《统计学习方法》（精简浓缩）

   b. Coursera Stanford《Machine Learning》（入门，对于NN和模型状态的部分可以参考）

   c. Coursera 台湾大学《机器学习高级技法》（SVM与Ensemble）

3. 项目

   a. 项目用了什么算法，选择它的原因，优缺点等。

   b. 如果没项目经验，相应的天猫大数据比赛和Kaggle比赛。

   c. 项目细节：数据从哪里来，做了什么预处理，特征怎么抽的，怎么想到的，评判标准是什么，模型优化过后的提升。

4. 海量数据处理（Map-Reduce）的面试题

   a. 相关知识参见所给资料

2、面试岗位说明：

每个企业对数据类岗位的命名可能有所不同，有些叫做数据挖掘/自然语言处理/机器学习算法工程师，有些简称算法工程师，有专门方向的比如搜索/推荐算法工程师，中小型的公司有些会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大概分为：

- **平台/数据处理类**
  数据计算平台搭建使用和基本数据清洗，处理和各种统计，要求支持大样本量、高维度数据，可能需要底层开发、并行计算、分布式计算等方面的知识(map-reduce等)；
- **算法研究与应用类**
  文本挖掘，如领域知识图谱构建、社交网络热门主题，核心主体挖掘等；
  推荐，广告推荐、商品推荐、题目推荐、新闻推荐、音乐推荐等；
  排序，搜索结果排序、广告排序，商品排序等；

> 用户信用评价，反作弊与风控；
> 图像识别、理解、图像内容跟踪。
> 商业智能，如统计报表，数据走势预测；
> 传统行业的应用，预测流失用户。

其中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，比如互联网医疗、互联网金融、互联网在线教育。**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。**

3、面试考察能力说明（for攻城师）：

1、数据结构算法水题

2、常用机器学习算法推导

3、模型调优细节

4、业务认识

a. 算法和理论基础

b. 工程实现能力与编码水平

c. 业务理解和思考深度

d. 沟通和表达能力

4、Machine Learning相关问题

- 最小二乘与梯度下降的区别
  - https://www.zhihu.com/question/20822481
- 最优化问题中，牛顿法为什么比梯度下降法求解需要的迭代次数更少?
  - https://www.zhihu.com/question/19723347
- 最小二乘、极大似然、梯度下降有何区别
  - https://www.zhihu.com/question/24900876
- 梯度下降or拟牛顿法
  - https://www.zhihu.com/question/46441403

- 如何判断函数凸或非凸?
  - https://www.zhihu.com/question/49902644
- 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?
  - 建议从简单的算法开始讲
- 无监督和有监督算法的区别，什么是半监督?
  - 训练样本数据和待分类的类别已知，然而训练样本既有标签数据，也有非标签数据；

    无监督学习：k-聚类、主成分分析等；

    有监督学习：支持向量机、线性判别；

    半监督学习：S3VM、S4VM、CS4VM、TSVM；
- 判别模型与生成模型?
  - https://www.zhihu.com/question/20446337
- LR 的损失函数和含义，梯度下降简单推导，特性?
  - 参考课件
- 什么是准确率，召回率，F值，ROC曲线，AUC?
  - https://www.zhihu.com/question/30643044
- 决策树的原理?
  - 参考课件
- SVM 用的什么损失函数，特性?
  - 参考课件
- SVM的kernel，什么时候用什么kernel
  - 参考课件
- SVM、LR、决策树的对比?
  - https://www.zhihu.com/question/21704547
  - https://www.zhihu.com/question/34735588
  - https://www.zhihu.com/question/26726794
- GBDT 和 随机森林 的区别?
  - 见19课课件
- 随机森林有什么优点?
  - a. 对于很多数据集表现良好，精确度比较高；
  - b. 不容易过拟合；

    - c. 可以得到变量的重要性排序；
    - d. 既能处理离散型数据，也能处理连续型数据，且不需要进行归一化处理；
    - e. 能够很好的处理缺失数据；
    - f. 容易并行化
- 多分类怎么处理?
    - 1 vs 1
    - 1 vs rest
    - softmax等
- 样本处理?
    - https://www.zhihu.com/question/30492527
- 正则化?
    - https://www.zhihu.com/question/20700829
- Kmeans优缺点，K的取值，改进
    - https://www.zhihu.com/question/31296149
- Hadoop怎么实现K-means
    - http://www.open-open.com/doc/view/d4657e719c6f45e98e5ffd79aed3a613
- 聚类算法中的距离度量有哪些，一般在什么场景下用?
    - 参考课件
- 解释贝叶斯公式和朴素贝叶斯分类。
    - http://blog.csdn.net/han_xiaoyang/article/details/50616559
    - http://blog.csdn.net/han_xiaoyang/article/details/50629587
    - http://blog.csdn.net/han_xiaoyang/article/details/50629608
- 如何进行特征选择?
    - 参见课件
- 为什么会产生过拟合，有哪些方法可以缓解过拟合?
    - 谈一下模型状态，数据或者正则化角度谈改善方式
- 你用过哪些机器学习/数据挖掘工具或框架?
    - numpy，scipy，pandas，sklearn，xgboost，caffe/Tensorflow/Keras
- 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?

    - 拿GMM为例解释下
- 主体模型里LDA的原理和推导（企鹅家）
    - 参见课程PPT
- 做广告点击率预测，用哪些数据什么算法（BAT）
    - LR GBDT FM FFM NN
- 推荐系统的算法中最近邻和矩阵分解各自适用场景（AT）
    - 参见课件
- 用户流失率预测怎么做（游戏公司/外卖公司等...）
    - https://zhuanlan.zhihu.com/p/22214370
    - https://www.zhihu.com/question/20308082
- 线性分类器与非线性分类器的区别及优劣；
    - https://www.zhihu.com/question/30633734
- 特征比数据量还大时，选择什么样的分类器?
    - 特征稀疏的情况下，其实LR这种分类器也是OK的
    - Random Forest能缓解过拟合
- 对于维度很高的特征，你是选择线性还是非线性分类器?
    - 一般是线性
- 对于维度极低的特征，你是选择线性还是非线性分类器?
    - 一般非线性，对特征做特征映射
- L1和L2正则的区别，如何选择L1和L2正则?
    - https://www.zhihu.com/question/37096933
- 随机森林的学习过程；
    - 参考课件
- 随机森林中的每一棵树是如何学习的；
    - 参考课件
- 随机森林学习算法中CART树的基尼指数是什么?
    - 参考课件