机器学习面试题集结号

一般技术面有以下一些环节：自我介绍，项目介绍，算法提问（推公式），数据结构提问（写代码）；自我介绍：一般尽量简短，主要讲清楚自己的研究方向，所取得成就以及优势所在即可；项目介绍：简历上的项目一定要熟悉，介绍时候分三部曲：项目背景，项目方案，项目成果；对项目中涉及到的一些技术点一定要很熟悉；算法提问：一般是问常见机器学习模型原理或者一些机器学习常见问题的解决方案（比如正负样本不平衡之类的），所以常见的机器学习模型一定要很清楚原理，必须会推公式，能知道工程实现的一些trick的话，那你就离sp不远了；
统计学习的核心步骤：模型、策略、算法，你应当对logistic、SVM、决策树、KNN及各种聚类方法有深刻的理解。能够随手写出这些算法的核心递归步的伪代码以及他们优化的函数表达式和对偶问题形式。

代码算法：基本算法（如快排等，需要熟练掌握） + 剑指Offer（面试经常出相似的题） + LeetCode（剑指Offer的补充，增强动手能力）2. 机器学习：李航《统计学习方法》（读3遍都不为过啊！） + Coursera Stanford《Machine Learning》（讲得很基础，但是没有告诉你所以然） + Coursera 台湾大学《机器学习高级技法》（里面详解了SVM，Ensemble等模型的推导，优劣）3. 请详细地回忆自己做过的项目，项目用了什么算法，为什么用它，有什么优缺点等。如果没项目经验可以参加天猫大数据比赛和Kaggle比赛。4. 教你如何迅速秒杀掉：99%的海量数据处理面试题。（基本每次都有一道海量数据处理的面试题）

数据结构算法水题+常用机器学习算法推导+模型调优细节+业务认识

在面试过程中，除了基础的东西要掌握，可以适当地向面试官展示你的一些其他的亮点，比如跟面试官谈论某些最近 paper 的进展以及一些技术方面的想法等，突出自己的与众不同；

掌握 常见的机器学习模型（线性回归，逻辑回归， SVM ， 感知机；决策树，随机森林， GBDT ， XGBoost ； 贝叶斯， KNN ， K-means ， EM 等）；掌握常见的机器学习理论（过拟合问题，交叉验证问题，模型选择问题，模型融合问题等）；掌握常见的深度学习模型（ CNN ， RNN 等）；这里的掌握指的是能够熟悉推导公式并能知道模型的适用场景；推荐书籍：《统计学习方法》《机器学习》《机器学习实战》《 UFLDL 》自然语言处理：掌握常见的方法（ tf-idf ， word2vec ， LDA ）；了解推荐以及计算广告相关知识；推荐书籍：《推荐系统实践》《计算广告》通过参加数据挖掘竞赛熟悉相关业务场景，常见的比赛有 Kaggle ， 阿里天池， datacastle 等；

比如LR,FFM,SVM,RF,KNN，EM，Adaboost,PageRank，GBDT，Xgboost，HMM，DNN，推荐算法，聚类算法，图像，自然语言，等等机器学习领域的算法，这些基本都会被问到
哪些优化方法，随机梯度下降，牛顿拟牛顿原理
常见分类模型（ svm ， 决策树，贝叶斯等）的优缺点，适用场景以及如何选型；

不容错过的机器学习试题与术语

深入理解机器学习，要的是能够透过现象看清算法背后的本质，而且要有着自己对它的思考，不是仅仅停留在表面，哪怕把推导，特性什么的都背一遍也没有用。因为只要换个问题场景，问个开放性问题，如果面试者没有对算法有深入理解，没有很强的learning sense，很容易就被问倒了，这也是高水平的面试官喜欢问的，因为面试者水平会暴露的一览无余。然而深入理解算法，需要的是长期的学习，思考，以及丰富的实践经验，我觉得这也是这个岗位隐藏的高门槛所在，同时也是许多自学的同学和本科生所缺乏的。

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；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；

机器学习&数据挖掘笔记_16（常见面试之机器学习算法思想简单梳理）

机器学习面试准备（持续更新）--- 优秀博文传送门，收集优秀资源

深度学习岗位面试问题整理笔记

阿里妈妈：团队人不多，主要做ctr预估。对于比赛经历貌似不太看重，比较care工程、科研经历。面试官也会查考的比较有深度，从对机器学习、统计学习的理解、普通线性回归到广义线性回归的本质、深度学习的本质、AUC、给出的建议就是面试阿里妈妈的话，提前对计算广告这块做些准备，然后面试的时候，尽量用数学语言、伪代码的方式进行清晰地阐述，当然这也跟面试官有关。

一、机器学习算法

- SVM：
  - 简单介绍SVM（详细原理）：从分类平面，到求两类间的最大间隔，到转化为求间隔分之一，等优化问题，然后就是优化问题的解决办法，首先是用拉格拉日乘子把约束优化转化为无约束优化，对各个变量求导令其为零，得到的式子带入拉格朗日式子从而转化为对偶问题， 最后再利用SMO（序列最小优化）来解决这个对偶问题。
  - SVM的推导，解释原问题和对偶问题，KKT限制条件，软间隔问题，解释支持向量、核函数（哪个地方引入、画图解释高维映射，高斯核可以升到多少维，如何选择核函数），引入拉格朗日的优化方法的原因，最大的特点，损失函数解释，
  - SVM与LR最大区别，LR和SVM对于outlier的敏感程度分析，逻辑回归与SVM的区别
  - 加大训练数据量一定能提高SVM准确率吗?

- 与感知器的联系和优缺点比较
- 如何解决多分类问题、可以做回归吗，怎么做
- 它与其他分类器对比的优缺点，它的速度
- 解释对偶的概念。
- 机器学习有很多关于核函数的说法，核函数的定义和作用是什么？https://www.zhihu.com/question/24627666
- 支持向量机(SVM)是否适合大规模数据？https://www.zhihu.com/question/19591450
- SVM和逻辑斯特回归对同一样本A进行训练，如果某类中增加一些数据点，那么原来的决策边界分别会怎么变化？https://www.zhihu.com/question/30123068
- 各种机器学习的应用场景分别是什么？例如，k近邻,贝叶斯，决策树，svm，逻辑斯蒂回归和最大熵模型。https://www.zhihu.com/question/26726794
- Linear SVM 和 LR 有什么异同？https://www.zhihu.com/question/26768865

- LR

  - LR推导（伯努利过程，极大似然，损失函数，梯度下降）有没有最优解？
  - LR可以用核么？可以怎么用？l1和l2正则项是啥？lr加l1还是l2好？加哪个可以用核（加l2正则项，和svm类似，加l2正则项可以用核方便处理）
  - LR可以用来处理非线性问题么？（还是lr啊 只不过是加了核的lr 这里加核是显式地把特征映射到高维 然后再做lr）怎么做？可以像SVM那样么？为什么？
  - 为什么LR需要归一化或者取对数，为什么LR把特征离散化后效果更好，为什么把特征组合之后还能提升，反正这些基本都是增强了特征的表达能力，或者说更容易线性可分吧
  - 美团技术团队《Logistic Regression 模型简介》https://tech.meituan.com/intro_to_logistic_regression.html
  - SVM和logistic回归分别在什么情况下使用？https://www.zhihu.com/question/21704547
  - 逻辑斯蒂回归能否解决非线性分类问题？https://www.zhihu.com/question/29385169
  - 为什么LR可以用来做CTR预估？https://www.zhihu.com/question/23652394
  - 逻辑回归估计参数时的目标函数 （就是极大似然估计那部分），逻辑回归估计参数时的目标函数 （呵呵，第二次） 逻辑回归估计参数时的目标函数 如果加上一个先验的服从高斯分布的假设，会是什么样（天啦。我不知道，其实就是在后面乘一个东西，取log后就变成加一个东西，实际就变成一个正则项）
  - 逻辑回归估计参数时的目标函数逻辑回归的值表示概率吗？（值越大可能性越高，但不能说是概率）
  - 手推逻辑回归目标函数，正类是1，反类是-1，这里挖了个小坑，一般都是正例是1，反例是0的，他写的时候我就注意到这个坑了，然而写的太快又给忘了，衰，后来他提醒了一下，改了过来，就是极大似然函数的指数不一样，然后说我这里的面试就到这了。
  - 看没看过scikit-learn源码LR的实现？（回头看了一下是调用的liblinear，囧）

- 为什么LR需要归一化或者取对数，为什么LR把特征离散化后效果更好，为什么把特征组合之后还能提升，反正这些基本都是增强了特征的表达能力，或者说更容易线性可分吧
- naive bayes和logistic regression的区别http://m.blog.csdn.net/blog/muye5/19409615

- L1和L2
  - L2正则化，为什么L2正则化可以防止过拟合？L1正则化是啥?
  - 深度学习里面怎么防止过拟合？（data aug；dropout；multi-task learning）如何防止过拟合，我跟他列举了4中主要防止过拟合方法：Early Stopping、数据集扩充、正则化法以及dropout，还详细跟他说了每种方法原理及使用的场景，并解释我在哪些项目里具体用到了这些方法，
  - 机器学习中使用「正则化来防止过拟合」到底是一个什么原理？为什么正则化项就可以防止过拟合？https://www.zhihu.com/question/20700829
  - 机器学习中常常提到的正则化到底是什么意思？https://www.zhihu.com/question/20924039
  - 什么是正则项，L1范式，L2范式区别是什么，各自用在什么地方？L1 与 L2 的区别以及如何解决 L1 求导困难；
  - L1正则为什么能让系数变为0？L1正则怎么处理0点不可导的情形？（这个谁会？近端梯度下降）
  - L0，L1，L2正则化(如果能推导绝对是加分项，一般人最多能画个等高线，L0是NP问题)其实上面的这些问题基本都能在《李航：统计学习方法》《周志华：机器学习》里面找到，能翻个4，5遍基本就无压力了
  - 避免过拟合策略、如何提高模型泛化能力、L1与L2正则区别，优缺点、生成式，判别式模型、深度学习这块了解多少、
  - 如何克服过拟合，欠拟合
  - L1 与 L2 的区别以及如何解决 L1 求导困难；

- 树模型
  - rf ， gbdt 的区别； gbdt ， xgboost 的区别（烂大街的问题最好从底层原理去分析回答）
  - 介绍决策树，谈了3种决策树及其区别和适应场景
  - 决策树处理连续值的方法；简单介绍决策树几种算法，有什么区别?
  - 决策树基本模型介绍？决策树算法中缺失值怎么处理？决策树算法在应用中有什么值得注意的地方。SVM、LR、决策树的对比？GBDT 和 决策森林 的区别？决策树的特性？（3 ）决策树处理连续值的方法；
  - 解释下随机森林和gbdt的区别。gbdt的boosting体现在哪里。解释下随机森林节点的分裂策略，以及它和gbdt做分类有什么区别？哪个效果更好些？为什么？哪个更容易过拟合？为什么？ 问了随机森林的损失函数，和lr的优缺点对比， adaboost和随机森林的

比较，为了防止随机森林过拟合可以怎么做，是否用过随机森林，怎么用的。
    - 随机森林和GBDT的区别？CART（回归树用平方误差最小化准则，分类树用基尼指数最小化准则）
    - GBDT（利用损失函数的负梯度在当前模型的值作为回归问题提升树算法中的残差的近似值，拟合一个回归树）
    - 随机森林（Bagging+CART）
    - SVM与随机森林比较
    - 改变随机森林的训练样本数据量，是否会影响到随机森林学习到的模型的复杂度
    - Logistics与随机森林比较
    - GBDT与随机森林比较随机森林的学习过程；随机森林中的每一棵树是如何学习的；随机森林学习算法中CART树的基尼指数是什么？
    - RF 与 GBDT 区别，原理优缺点适用场景分析，哪个具备交叉验证功能等
    - 接着写一下信息增益的公式。之后就是问机器学习相关算法，说了一下bagging跟boosting，之后问了GBDT（没做过，只能说说大体思路）。（2） rf， gbdt 的区别；gbdt， xgboost 的区别；
    - 说说xgboost、gbdt区别、Tree-based Model如何处理连续型特征。
    - 让我把一个完整的数据挖掘流程讲一下，从预处理，特征工程，到模型融合。介绍常用的算法，gbdt和xgboost区别，具体怎么做预处理，特征工程，模型融合常用方式，融合一定会提升吗？
    - 介绍LR、RF、GBDT，分析它们的优缺点，是否写过它们的分布式代码
    - XGB和GBDT区别与联系也会经常问到：https://www.zhihu.com/question/41354392/answer/128008021?group_id=773629156532445184
    - CART（回归树用平方误差最小化准则，分类树用基尼指数最小化准则）、Logistics（推导）、GBDT（利用损失函数的负梯度在当前模型的值作为回归问题提升树算法中的残差的近似值，拟合一个回归树）
    - 在面试过程中主动引导面试官提问，比如面试官让你讲解 gbdt 原理时，这会你可以跟他说，一般说起 gbdt，我们都会跟 rf 以及 xgboost 一块讲，然后你就可以主动地向面试官输出你的知识；面试并不是死板地你问我答，而是一种沟通交流，所以尽可能地把面试转化成聊天式的对话，多输出自己一些有价值的观点而不是仅仅为了回答面试官的问题；
    - 几种树模型的原理和对比，
    - 特征选取怎么选？ 为什么信息增益可以用来选特征？
    - 信息熵和基尼指数的关系(信息熵在x=1处一阶泰勒展开就是基尼指数)

- K-means
    - k-means 聚类的原理以及缺点及对应的改进；kmeans 算法的优缺点。。。。
    - kmeans 的原理，优缺点以及改进；
    - em 与 kmeans 的关系；

- kmeans 代码；
- 说说 Kmeans 算法， Kmeans 算法 K 怎么设置、适用什么样数据集、怎么评价 Kmeans 聚类结果、 Kmeans 有什么优缺点？你的项目中使用 Kmeans 遇到哪些问题，怎么解决的？
- 用 EM 算法推导解释 Kmeans。
- KMeans的算法伪代码
- 如何判断自己实现的 LR、Kmeans 算法是否正确?
- 如何优化kmeans算法
- 如何用hadoop实现k-means
- 手写k-means的伪代码（就6行）

- 集成学习

  - bagging和boosting是怎么做的和他们的比较
  - 详细讨论了样本采样和bagging的问题
  - 聊的比较多的是如何知道一个特征的重要性，如何做ensemble哪些方法比较好。聊了聊计算广告方面FM，embedding。
  - 常见融合框架原理，优缺点，bagging，stacking，boosting，为什么融合能提升效果
  - 是否了解线性加权、bagging、boosting、cascade等模型融合方式
  - K-means起始点http://www.cnki.com.cn/Article/CJFDTotal-DNZS200832067.htm

- 贝叶斯

  - 朴素贝叶斯分类器原理以及公式，出现估计概率值为 0 怎么处理（拉普拉斯平滑），缺点；
  - 解释贝叶斯公式和朴素贝叶斯分类。
  - 贝叶斯分类，这是一类分类方法，主要代表是朴素贝叶斯，朴素贝叶斯的原理，重点在假设各个属性类条件独立。然后能根据贝叶斯公式具体推导。考察给你一个问题，如何利用朴素贝叶斯分类去分类，比如：给你一个人的特征，判断是男是女，比如身高，体重，头发长度等特征的的数据，那么你要能推到这个过程。给出最后的分类器公式。
  - 那你说说贝叶斯怎么分类啊？比如说看看今天天气怎么样？我：blabla，，，利用天气的历史数据，可以知道天气类型的先验分布，以及每种类型下特征数据（比如天气数据的特征：温度啊，湿度啊）的条件分布，这样我们根据贝叶斯公式就能求得天气类型的后验分布了。。。。面试官：en（估计也比较满意吧）那你了解关于求解模型的优化方法吗？一般用什么优化方法来解？
  - 贝叶斯分类器的优化和特殊情况的处理

- 深度学习

  - 解释一下CNN、介绍CNN、卷积公式，以及特点，假设面试官什么都不懂，详细解释CNN 的原理；问CNN的细节特点，哪些特点使得CNN这么好用，哪些场景用CNN可

以，抽象一下这些场景的特征，可以降采样但仍能保持主要信息；局部连接可以保证获取局部信息；权值共享保证高效，DNN和CNN相比有哪些区别，用过RNN么？画一下RNN的图，你在深度学习过程中遇到过哪些问题？如果出现过拟合你怎么办？dropout是什么？它有什么用？你会怎么用它？当全连接跟dropout连着用需要注意什么？你之前过拟合怎么解决的？如果本身training loss就很大你怎么办？如果数据不变，怎么调整网络结构解决这个问题？（batch normalization）梯度消失知道么？为什么会出现梯度消失？dnn和rnn中的梯度消失原理一样么？dnn中是哪个部分导致梯度消失？（激活层如sigmoid）rnn中怎么解决梯度消失问题？（lstm的结构相对普通RNN多了加和，为避免梯度消散提供了可能。线性自连接的memory是关键。）讲一下CNN吧，有哪些重要的特点？CNN可以处理哪些场景？为什么CNN要用权值共享？（每个卷积核相当于一个特征提取器，它的任务是匹配局部图像中的特征，权值共享后，匹配的特征方式都是一样的，提取若干特征后就知道学习的是啥了）CNN里面哪些层？讲一下卷积。卷积的形式是啥样？给定一个输入，算输出的feature map大小。卷积有啥用？池化有啥用？有哪些池化方式？池化除了降采样还有啥用？（就不知道了）还有哪些层你用过？讲讲dropout。dropout内部是怎么实现只让部分信号通过并不更新其余部分对于输入的权值的？讲讲BN（BatchNormalization）为什么好？全连接有什么用处？知道RNN么？讲讲RNN大致的实现思路。知道梯度消失么？为什么会出现梯度消失？RNN里的梯度消失一般怎么处理？细讲下lstm的结构，这样设计为什么好？（门关闭，当前信息不需要，只有历史依赖；门打开，历史和当前加权平均）你觉得梯度消失靠引入一些新的激活层可以完全解决么？为什么？

- 问了做的比赛里面使用tensorflow的细节，LSTM里调参的细节
- 用过哪些库或者工具，mkl，cuda这些会用吗？
- 有一个弱分类器和大量未被标记过的图像数据，如何人工标记图像来对分类器进行提升
- 介绍下RNN和它的优缺点
- 让我推导BP反向传播、随机梯度下降法权重更新公式
- 卷积神经网络结构特点、各参数对模型结果影响、项目进展遇到的难题、推导BP神经网络参数更新方式、随机梯度下降法（SGD）优化函数存在的缺点以及拟牛顿法在优化函数使用上更有优势、修改Caffe开源框架、开源社区代码贡献量就跟我聊了很多行业发展趋势及问题，知道目前深度学习的一个趋势，也了解到最新行业发展动态，改进相机智能化程度，也聊到了美颜相机美颜效果以及小米相机人脸分类、年龄检测等等不足之处，了解到新兴行业大佬商汤科技和旷视科技（face++脸草）在研究的热门方向
- 看到有deep learning相关的项目，就问了deep learning 相关问题：如何减少参数（权值共享、VGG的感受野、GoogLeNet的inception ），激活函数的选择（sigmoid->ReLu->LReLU->PReLU ），为什么之前没有深度网络出现（数据量不够+机器性能），由数据引申到数据不平衡怎么处理（10W正例，1W负例，牛客上有原题），
- 后面问了下DNN原理，应用，瞎扯一通……
- 你了解神经网络吗？我：了解一些，讲感知机，然后是BP网络。简单讲了一下原理。
- 图像处理题：如何找相似图片。我说用感知哈希算法，计算汉明距离，他说这种方法精度不行；我说那就用SIFT算法吧，他说SIFT效果还可以，但计算有点繁重，有没有轻量

级的方法？我想起来去年在美图秀秀实习时，曾经做过一种图像滤波算法，有一步是把像素点用K-means聚类。我就说先把图片灰度化，然后用K-means聚类，把聚类后的各个中心点作为一张图片的特征向量如果两张图片的特征向量相近则说明这两张图片相似。貌似我这个答案有点出乎他的意料，他意味深长地说了个“行吧~~~~”（个人觉得颜色直方图匹配是个他期待的常规回答）
    - 介绍卷积神经网络，和 DBN 有什么区别?
    - Deep CNN, Deep RNN, RBM的典型应用与局限，看Hinton讲义和Paper去吧
    - 神经网络,plsi的推导
    - 验证码图片的去噪和提取字符
    - 有限状态自动机,然后要我画状态转移图.

- 聚类
    - 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?

- 优化
    - 梯度下降的优缺点；主要问最优化方面的知识，梯度下降法的原理以及各个变种（批量梯度下降，随机梯度下降法， mini 梯度下降法），以及这几个方法会不会有局部最优问题，牛顿法原理和适用场景，有什么缺点，如何改进（拟牛顿法）
    - 常用优化算法：1.梯度下降法：又有随机梯度下降和负梯度下降，2.牛顿法 主要是问了各自的优缺点，速度，能不能得到全局最优解，牛顿法的二次收敛等
    - 问你如果有若干个极小值点，如何避免陷入局部最优解。
    - 它们间的牛顿学习法、SGD如何训练，
    - 如何判断函数凸或非凸?
    - 线性回归的梯度下降和牛顿法求解公式的推导
    - 最速下降法和共轭梯度法 wolfe条件 最速下降法和共轭梯度法的收敛速度如何判断
    - 深刻理解常用的优化方法：梯度下降、牛顿法、各种随机搜索算法（基因、蚁群等等），深刻理解的意思是你要知道梯度下降是用平面来逼近局部，牛顿法是用曲面逼近局部等等。

- 推荐系统
    - 介绍SVD、SVD++
    - 推荐系统的冷启动问题如何解决
    - 深度学习在推荐系统上可能有怎样的发挥?
    - 推荐系统的算法中最近邻和矩阵分解各自适用场景
    - 白板写SVD/SVD++公式，SGD迭代更新p，q矩阵公式，SVD/SVD++优化方法
    - 对推荐算法的未来看法；
    - 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，

- 我面的推荐，问了各类协同过滤的好与坏。
- 问了一个很有意思的问题，现实应用中的Top-N推荐问题和学术研究中的评分预测问题之间有什么不同。问我ItemCF的工程实现，面对大数据如何实现，又追问了有没有什么工程优化算法。这个问题我没答好，一开始我说了一个MapReduce模型，他问能不能更快一点，我就卡那了。。。最后面试官告诉我，不能只从算法角度分析，要从系统设计分析，利用内存来减小MapReduce的吞吐量。（当然也许从MapReduce那一刻开始我就输了也不一定）
- 推荐系统的算法中最近邻和矩阵分解各自适用场景http://www.doc88.com/p-3961053026557.html

- PCA
  - 那你对pca了解吗？我：了解啊，面试官：那讲一下pca是用来干嘛的？我：pca啊，可以用来分析主方向啊，降维啊，特征筛选啊，具体方法是用svd分解得到特征值矩阵和特征向量矩阵，然后根据不同的任务对选择特征值或向量进行计算。

- EM
  - 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?

- NLP
  - 用过哪些 NLP 算法项目中用过哪些机器学习算法。
  - 海量的 item 算文本相似度的优化方法；
  - 解释 word2vec 的原理以及哈夫曼树的改进
  - 二面面试官主要跟我聊简历上的几个项目，他好像不能理解词向量的形式，反复解释了很多遍，问的问题都比较简单，有TF-IDF,余弦相似度，分词工具等等。
  - 然后我说我做过LDA，问我，Dirichlet Distribution的定义和性质，并问我，为什么它和multinomial distribution是共轭的，顺便问了我啥叫共轭分布。

- 关联分析：
  - 项目中涉及到频繁模式挖掘，于是问了一下如何实现的？ 用的是 Apriori算法，描述他的原理过程，关键字眼：支持度，支持度计数，k项候选频繁项集，怎么从k项到k+1项等，连接剪枝过程。

- hadoop
  - 简单介绍 MapReduce 原理，有没有看过源码，说说 Map 阶段怎么实现的，
  - MapReduce 实现统计出现次数最多的前 100 个访问 IP.
  - MapReduce 实现统计不重复用户 ID,MapReduce 实现两个数据集求交集。
  - HBase 行健怎么设计,spark 性能一般优化方法,spark streaming 和 storm 区别.给了一

张笔试题， 10 道选择，一道大题。选择题是 java 基础知识，大题一个有三问：根据场景写出 Hive 建表语句； Hsql 从表中查询；
    - 用MapReduce写好友推荐，在一堆单词里面找出现次数最多的k个
    - 用分布式的方法做采样怎么保证采样结果完全符合预期?
    - 后面又问了Hadoop,Spark,storm下面的产品，原理，适用场景，
    - 写一个 Hadoop 版本的 wordcount。

- HMM
    - 实现 hmm 的状态转移代码；

- 杂项
    - 讲机器学习中常用的损失函数有哪些？交叉熵有什么好处？（凸优化问题）
    - 判别模型与生成模型的本质区别是什么
    - 分类模型和回归模型的区别，分类模型可以做回归分析吗？反过来可以吗？（我回答是分类不可以做回归，回归倒是可以做分类，不知道对不对）
    - k折交叉验证 中k取值多少有什么关系 （我不知道，随便答，然后面试官后面问我知道bias和variance吗？估计是和这两个东西有关， 知乎上有个问题讨论了k值大小与bias和variance的关系）
    - 解释局部相关性
    - 特征选择的方法；
    - 在模型的训练迭代中，怎么评估效果；
    - 特征选择方法有哪些(能说出来10种以上加分)，之后和面试官仔细聊了一下特征选择的问题，我介绍了了解的几种基本的特征选择思路（错误率选择、基于熵的选择、类内类间距离的选择）；
    - 有没有接触过机器学习的前沿，深度学习看过paper没有？（并没有）
    - 如何用尽可能少的样本训练模型同时又保证模型的性能；
    - 你读哪些期刊会议的论文？你遇到的比较有意思的算法?
    - 生成模型，判别模型
    - 线性分类和非线性分类各有哪些模型
    - 比较各个模型的Loss function，
    - 设计一个结构存取稀疏矩阵 （面试官最后告诉我了一个极度压缩的存法，相同行或列存偏差，我当时没听懂，还不懂装懂，最后还是没记住）
    - PageRank原理，怎么用模型来查找异常用户，我讲了一大堆我的理解，然后面试官一句你怎么不用规则把我噎到了……
    - 无监督和有监督算法的区别?
    - 经典算法推导(加分项)，原理，各个损失函数之间区别，使用场景，如何并行化，有哪些关键参数
    - 什么叫判别模型什么叫生成模型。

- 先针对项目十分细致地询问了各种细节，然后就问我如何处理数据中的噪声点、数据清洗算法（正好自己做了一个算法）、如何选择特征等。
- 校招TST内推，面过了2面，还是跟之前那个有点类似的游戏开发的安全部门，因为我也玩LOL，又问到怎么来判断玩家有没有作弊之类的问题，这次我小心翼翼的说用模型怎么做，用规则怎么做，感觉这次聊的都挺开心的。
- 是否了解A/B Test以及A/B Test结果的置信度特征工程经验是否了解mutual infomation、chi-square、LR前后向、树模型等特征选择方式
- 深刻理解各种算法对应采用的数据结构和对应的搜索方法。比如KNN对应的KD树、如何给图结构设计数据结构？如何将算法map-red化
- 矩阵的各种变换，尤其是特征值相关的知识。分布式的矩阵向量乘的算法
- 线性分类器与非线性分类器的区别及优劣；特征比数据量还大时，选择什么样的分类器？对于维度很高的特征，你是选择线性还是非线性分类器？对于维度极低的特征，你是选择线性还是非线性分类器？如何解决过拟合问题？L1和L2正则的区别，如何选择L1和L2正则?
- 项目中的数据是否会归一化处理，哪个机器学习算法不需要归一化处理
- 并行计算、压缩算法LDA http://www.doc88.com/p-1621945750499.html