一面：

1、xgboost和gbdt的区别？这方面问的很细，比如说xgboost可以并行加速是怎么进行的，每次分裂叶子节点是怎么决定特征和分裂点的。

2、LR手推，包括极大似然的概念，为什么极大似然的时候可以相乘，满足独立同分布？然后梯度下降更新权值，这里我自己自作主张的说了下mini-batch 梯度下降，SGD等，说手写过这里的代码，想展示下自己。

3、结合项目，讲下怎么进行特征选择，特征工程方面，比如说非数值型怎么处理，one-hot后维度高怎么。问你用的xgboost也许要对特征进行标准化吗？这里也考得xgboost吧，我感觉xgboost其实对特征的的预处理要求不是那么高，讲了下原理，就没再问了。

4、结合我的项目，用gan网络去噪，解释了下conditional gan的原理。损失函数的创新点。

5、画了下googlenet的结构。讲了下googlenet跟之前的网络的不同。

6、googlenet中为什么采用小的卷积核？答，两个3x3相当于一个5x5这样可以降低参数量。

7、代码题，反转链表，树的层次遍历（说思路，用了quene),全排列(123,321,213)。代码部分写的很慢，但基本也都由思路，写出来了。

8、LR和SVM的区别。

9、SVM为什么可以处理非线性问题。

一面过了，主要是面试官人很好。说我代码要加强，有思路但是写得很慢。

二面：1、问了爬虫的项目，和特征选择方面的问题，主要是高维的特征怎么进行降维。反爬虫的机制。

2、依然是xgboost和rt和gbdt三联问。这部分卤煮觉得自己答得挺全面，所以两次的面试官都挺满意的。

3、回归决策树和分类决策树分裂节点的时候怎么处理。

4、各种排序的时间复杂度（最好最坏和平均），空间复杂度，写了归并排序（这个写的很快）。面试小哥哥说你代码写的还挺好，楼主其实是背的，只能尴尬的说这个排序都是很基础的，小哥哥说你觉得太简单了？楼主说你不能因此考我难得，面试官居然就没怎么问代码了。

5、开放题，楼主不是做nlp的但是问了中国到中华人民共和国这种模糊搜索怎么办，小哥哥一直在引导，人真的很好。

6、手推LR（楼主真的很幸运，两次题目居然一样，开心的答了）。  
7、协同推荐方面，usercf和itemcf。

二面比较轻松了，和面试官聊的很愉快，人真的很好，没想到也过了。

三面：三面感觉是个技术大牛在问，很深入

1、亿级文件，每一行是一个字符串，单个文件中，字符串没有重复，两个文件中取交集。露珠真不会，引导下说了hash的原理等。

2、亿级用户推荐视频。

3、高维稀疏的特征怎么处理。答降维。问会不会embedding，说听过word2wec里面解决词向量，但是具体不会。

4、xgboost什么的深入问了下。

5、代码题，反转数组求最小。

<https://www.nowcoder.com/discuss/14953>

<https://www.nowcoder.com/discuss/94247>

word2vec原理，cbow+skip-gram，采用softmax计算输出概率，为了减少计算量，采用层级sofemax即哈夫曼树以及负抽样方法改进，其中哈夫曼树的流程，叶子结点，跟节点，节点都代表了什么，为什么利用哈夫曼树，近的更便于找到，贪心思想，并且复杂度降低，哈夫曼树二分类，左右树怎么分的，为什么左为负1，右为正0，能不能换，可以换，只是一种定义，层级softmax不用可不可以，正常的softmax也可以，计算量亿万级而已

F1计算公式，BOW，tfidf的意义

pagerank与倒排索引

手写k-means的伪代码和代码

svm原理与推导

L1和l2正则，为什么l1产生稀疏矩阵，l2可以防止过拟合

SVM和gbdt区别，介绍了svm和gbdt原理，从集成学习角度讲的，gbdt为什么效果比svm效果好，能够构建高度非线性的分类函数，而svm采用rgb核函数人工固定核函数，后调节参数，gbdt不也要调参数吗，都有哪些参数需要调节，

3.与Gbdt相似的方法有哪些，rf，adaboost,xgboost，分别讲了区别与联系，其中xgboost用了列抽样，为什么列抽样就能防止过拟合，会不会lightgbm

4.深度学习了解哪些算法（dnn,rnn,lstm），对embedding了解么，嵌入式处理，word2vec？  
  
word2vec原理与思想，层级softmax与负采样的思想，负采样是随机采样的吗，会不会抽到同一个？ LSTM的思想，门控消除梯度弥散，遗忘门输入门输出门等，传统DNN怎么改进梯度弥散的现象，更换激活函数ReLU，初始化参数，还有其他的方法么？加入BN层  
  
word2vec维度怎么选的（窗口大小为5，设置为随机效果更好），有没有自己实现word2vec，在使用lstm训练时，用的框架还是自己实现的，网络设置了多少层，为什么？每一层都是什么含义，都调节了哪些参数，最后效果怎样。

word2vec怎么训练的，有没有没有得到的词向量，比例多少，word2vec原理，LSTM+CNN怎么做的，谁前谁后，为什么，效果提升多少，用的什么工具（TensorFlow），分词用的什么工具（jieba），怎么做的新词发现（左右熵互信息），写互信息的公式，为什么互信息可以进行新词发现，原理

4.写SVM常用核函数、LR的公式，损失函数类型，为什么用sigmoid函数进行非线性映射（从二项分布的伯努利方程角度）

Word2vec原理，CBOW与Skip\_gram，二者不同，传统使用softmax全连接的计算工作量大，改进方法层级softmax，负抽样，各自的原理，如何进行梯度上升迭代（过程类似LR），哈夫曼树的叶子结点概率怎么算，节点如何进行分裂的，最终所有叶子结点的概率和是否等于1，在使用过程中，负抽样使用多还是层级softmax多，为什么（负抽样，计算简单），哈夫曼树是如何构建出来的

Word2vec原理，损失函数，怎么优化，怎么训练模型的，用的什么语料库，如何扩充的  
  
LSTM原理与流程，从dnn到rnn到lstm的转变，lstm的三个门控是如何设计的，具体的计算，上一时刻保留的加本时刻更新的传递给输出门，lstm如何解决梯度弥散的问题，为什么相乘变成相加就能解决  
  
改进现有的分类算法、聚类算法等，提出一种新算法，从哪个角度切入（讲了感知机到SVM的改进，GBDT到xgb的改进，kmeans到结合语义的改进）以及从梯度优化的角度，提高迭代速度（梯度下降到牛顿法、拟牛顿法、SGD、MBGD，动量，NAG，自适应的Adagrad、RMSprop、Adam），问了Adam的特点，xgb的改进点

为什么把CNN结构放到LSTM之前，效果为什么比单独使用LSTM差，为什么不考虑CNN+LSTM+CNN，论文里提到CNN对单字特征提取效果较好（没考虑到，再调整）

4.分词时用的什么工具，引入互信息与左右熵的方法，考虑内部聚合度和外部聚合度，从而决定是否进行切分，从而实现新词发现，提高分词精度。

5.对于不平衡数据如何处理：上采样，下采样，分配权重，SMOTE采样的过程。

6.从头至尾介绍垃圾邮件识别的过程，从数据获取至生成报告，重点考察思考问题的广度，数据不平衡时怎么办（如数据量很大时怎么处理，数据量小时怎么处理，如何将文本数据转化为特征向量，用什么方法，对于邮件的标题及正文是否进行相同的处理，权重是否相同，如何找到关键特征等），数据特征维度远高于数据量时，从样本采样、向量维度降维、分类算法的重新选择三方面改进，尽量多考虑系统性设计的思想

4. RF和GBDT的区别；RF怎么解决的过拟合问题；

5. 决策树的分裂策略：ID3，C4.5，Gini指数，选一个讲一下；

6. 熵了解吗？

7. 讲一下XGBoost；

8. 正则化项了解吗？

9. 优化算法： 牛顿法和拟牛顿法，过程？（这个不会，李航老师的统计机器学习附录有详细讲解）；

10. 熟悉深度学习吗?

11. 讲一下word2vec；

12 算法题： 找出数组中第K大的数。有几种方法？时间复杂度分别是多少？

高频问题简单总结一下，希望能够帮助大家。  
1.gbdt,xgboost,lgbm的区别(阿里，头条)  
2.梯度下降法，牛顿法，拟牛顿法区别(阿里)  
3.SGD,ADAM区别(百度)  
4.什么是梯度消失，饱和，如何改善(阿里)  
5.lr的推导(腾讯)  
6.SVM目标函数，为什么转为对偶(腾讯，百度)  
7.定义class mlp(头条)  
8.kd tree(腾讯)  
9.FFM的优化(百度)  
10.解释RESNET(百度，阿里)  
11.mapreduce思想(腾讯)  
12.解释BN(头条，百度)  
13.非结构化文本处理方法(阿里)  
14.bagging.boosting.stacking区别(阿里)  
15.CNN与RNN的区别(阿里)  
16.如何防止过拟合(头条)

怎么获得一棵完全二叉树的最后一个结点

激活函数的优缺点

介绍下xgboost

rnn结构图

xgboost和随机森林各自的特点，差别

词嵌入，N-Gram

命名实体识别，CRF

CNN和RNN各自在文本方面的特点，什么时候用。

BERT用没用过

LR和SVM的区别（写出各自目标函数）

boosting方法（先讲了总的boosting方法,然后讲了gbdt,手推其原理）

讲讲word2vec

说一下attention实现

介绍下结构，CRF层怎么用的

gbdt和rf区别，如何理解gbdt中的gb，从偏差方差角度讲一下

深度学习过程中学习曲线产生震荡，分析原因

2、LR为什么用sigmoid函数。这个函数有什么优点和缺点？为什么不用其他函数？

3、SVM原问题和对偶问题关系？

4、KKT条件有哪些，完整描述

6、有一堆已经分好的词，如何去发现新的词？

面试官给的提示：用这个词和左右词的关系。**互信息**新词的左右比较丰富，有的老词的左右也比较丰富。还要区分出新词和老词。

7、L1正则为什么可以把系数压缩成0，坐标下降法的具体实现细节

8、spark原理

9、spark Executor memory 给16G  executor core 给2个。问每个core分配多少内存

2、对于机器学习你都学了哪些？讲一个印象深的

说了SVM原理，拉格朗日法，对偶问题，以及好处。

3、SVM怎么防止过拟合

说了SVM里面的松弛变量。不知道对不对

4、我主动出击，有另一大类算法决策树，说不管是LR还是SVM都不能直观的感受到决策依据。而决策树易于理解，能够直观的感受到决策依据。

说了划分依据：信息增益（说了信息熵的来源，等概率时熵最大）、信息增益率、基尼系数。

说了划分方法（基于信息增益的）

说了C4.5比较ID3的优点。

5、决策树如何防止过拟合

剪枝，前剪枝和后剪枝。说了REP剪枝。C4.5是悲观剪枝

6、项目没问，说从上位面试官了解了。

7、撸代码

求连续子数组最大乘积，还让考虑边界问题（最后问了：连乘有可能导致溢出，存不下了）

4、K-means聚类个数选择，做什么样的试验来确定K

5、两个4G的文件（每个文件可能有重复），里面全都是数字。现有内存1G，求这两个文件的交集。

2个4G的文件，分别hash成10个子文件，一个400M。

把一个子文件存储到hash表中，作为key。遍历另一个文件，看这个数字是否存在于刚才的hash表中。存在即可输出。

作者：crazyhoney  
链接：[https://www.nowcoder.com/discuss/32008?type=0&order=0&pos=8&page=2](https://www.nowcoder.com/discuss/32008?type=0&order=0&pos=8&page=2" \t "_blank)  
来源：牛客网

3、说了SVM

4、为什么要把原问题转换为对偶问题？

因为原问题是凸二次规划问题，转换为对偶问题更加高效。

5、为什么求解对偶问题更加高效？

我答了，因为只用求解alpha系数，而alpha系数只有支持向量才非0，其他全部为0.

6、alpha系数有多少个？

我答了：样本点的个数

7、避免过拟合的方法

答了：决策树剪枝、L2正则和L1正则

8、为什么L1正则可以实现参数稀疏，而L2正则不可以？

答了：L1正则因为是绝对值形式，很多系数被压缩为0,。而L2正则是很多系数被压迫到接近于0，而不是0

9、为什么L1很多系数可以被压缩为0，L2是被压缩至接近于0？

答了：图像上，L1正则是正方形，L2正则是圆形。

L1正则的往往取到正方形顶点，即有很多参数为0

L2正则往往去不到圆形和参数线的交点，即很多分量被压缩到接近于0

SQL中inner join 和outer join的区别？

14、试图给他说说SPARK，结果被严词拒绝（开玩笑的）。。。说时间紧迫，还是他来问吧。。。

15、Kmeans中，现在给你n个样本点不在欧式空间中，无法度量距离。现在给了一个函数F，可以衡量任意两个样本点的相似度。请问Kmeans如何操作？

答：想了一会，比如K=4的聚类。

1、首先，随机去4个点，作为初始类簇中心。

2、计算所有样本点与这4个点的F相似度。根据相似程度，把所有样本点分到4个类中。

3、在这4个类中，计算每一个样本点 i 到该类其他样本点的相似度和Si。取Si最大的那个点作为这个类的中心。

4、重复2、3步骤，直到类中心不再变化或者循环次数达到目标。

我给您说说SVM吧，自学的时候留下很深的印象（试图抓住主动权~）

SVM是基于。。。说着手动起来写SVM的损失函数

A （打断）为什么样本点到决策面是 1/||w||

B 手推向量点到决策面的表达式（麻蛋，竟然一时紧张忘了。。。没推出来）

A 点到直线距离公式记得吧？

B 嗯嗯，又没写出来。只能说之前推过，现在一紧张忘了。。。

A 这个也无关紧要，继续

B 继续说SVM

A （打断）知道LR吧，知道LR和SVM有什么不同吗？

B 知道，首先这两个算法的分类思想不同，LR是基于概率推导的，SVM是基于最大化几何间隔的

A （打断）写一下，LR的损失函数

B 手写出来。其实这个sigmoid函数由那个什么族分布（真的忘了名字，其实是：指数族分布），加上二项分布导出来的。损失函数是由最大似然估计求出的。

A 怎么由最大似然估计导出的？推导一下

B 最大似然估计就是求让已知事件发生的概率最大的参数。

假设有5个样本，每一个的类别是yi，由LR计算出的概率是h(x)。那么每一个样本预测正确的概率为：

(H(x)^yi)\*((1-h(x))^(1-yi)) ----

（刚开始一紧张，把h(x)和yi写反了）面试官说是这样吗？你这样全为0，我感觉你在背公式。。。你再看看

我一看，卧槽这竟然写错了。赶紧改过来，然后表明是自己紧张了。

概率连乘后，然后取对数就是LR的损失函数了。

A 为什么损失函数有个负号？

B 这是因为要应用梯度下降法，引入的。不加负号也可以，梯度上升法。这都是一样的。

A OK，继续，LR和SVM有什么区别？

B SVM决策面只由少量的支持向量决定，而LR的话是所有样本都会参与决策面的更新。

A 对，所以说SVM怎么样？

B SVM对于异常点不敏感，而LR敏感。SVM更加健壮，决策面不受非支持向量影响。

A OK

A 知道过拟合吧？

B 知道，在训练集表现好，在测试集表现一塌糊涂。举个例子就是：学生平时考试成绩非常棒，但一到实际应用就很烂。

A 说说常见的过拟合的解决办法

B 数据，样本不够，如果现在的训练集只是所有样本空间的一个小小的部分，那么这个模型的泛化能力就非常差（边画图，边说）

A 嗯嗯，还有呢

B 可以加正则项，L1，L2正则。L1还可以用来选择特征

A 为什么L1可以用来选择特征

B 因为L1的话会把某些不重要的特征压缩为0

A 为什么L1可以把某些特征压缩为0

B 因为（画图）L1约束是正方形的，经验损失最有可能和L1的正方形的顶点相交，L1比较有棱角。所以可以把某些特征压缩为0

A 还有什么过拟合的解决方法

B 神经网络中，dropout方法。就是每层网络的训练，随机的让一半神经元不工作。达到防止过拟合的目的

A 还有吗？

B 决策树中可以用剪枝操作。

B 决策树过拟合，可以用随机森林。。。

A 什么？？？现在一个决策树已经过拟合了，还要再以它为基准训练随机森林？

B 。。。对，你说的对。我想错了。。。

B 我就知道这些方法了。。。

A OK，挑一个项目给我说说吧

B 说项目（不记得中间有没有再提问了。。。）

B 要不我给您说说spark框架吧，之前还用的挺多。

A 嗯（看简历和笔试题中。。。）

B 开始说。。。说到三分之一

A 好了！ 你不必说了。（大手一挥~）我看你5道笔试题都没写思路，现在把第二题代码写出来

注： 第二题就是检测括号是否匹配

B 我写了啊。。。（给他翻到其中一个的背面）

A 哦，（迅速扫过代码，），为什要把字符压栈呢？不压栈也可以的。

B 是吗？{abc()}这样的也是合法的吗？

A 当然啊（看了一眼题。）

B 好吧，我本来也准备看到字符就丢到，不入栈。但担心这种情况不合法，就给入栈了。

A 嗯，第三题呢？

B 没思路，没写

A 给我说说第四题

第四题：10分钟内，恶意IP访问检测（10分钟内访问次数超过1024即为恶意访问）

B 这是10分钟动态检测的，需要时间刻度精确到秒吗？

A 不需要

B 把10分钟内的<ip,次数>存入hashmap, 再把key,value互换存入treemap。因为treemap是基于key有序的，升序。然后直接拿出来最后一个和1024比较。

A 怎么实现动态的检测，当前检测0-10分钟，那么第11分钟怎么办？

B 把0-10分钟的摘出来，从10分钟内的hashmap中减去，再把10-11分钟内的加上。

我知道这样实现起来，效率应该不高，但这一会我只想到了这个。。。

A 嗯，其实可以这样，把每分钟的分开存储，动态的向后移动，取这10个的总的数据就行。

甚至可以每分钟只存储TOP200的，然后10个分钟的汇总，取TOP1

B 嗯，明白了。

字存在于这个hashet中。就认为有循环依赖。

3、项目中用到了聚类？手写一下Kmeans

4、一般工业界不这样用，用kd-tree加速

5、给你出道题写一下，一个文件每一行有3列（\t分隔），每个字符串是abcd，这种形式，中间有大写有小写。

现认为：abcDe 等于BcaDe （即：不区分大小写，无关顺序）

要求输出： 字符 空格 出现次数 空格 每一种字符（以|分隔）

实例输出： abcde 2 abcDe|BcaDe

3、说SVM

4、好像还说了spark原理

**5、电话中断，面试官线上有BUG，去改BUG了。。。**

6、10分钟后电话来了

7、我主动说：我给您说一下决策树方面的吧

8、面试官说：不用了，来道题。。。

9、一个矩阵都是0,1 且每一行，0都在1前面。求1个数最多的那一行的序号

3、说一下项目中用的Kmeans算法

4、知道哪几种聚类算法，说下原理

5、Kmeans有什么优缺点

6、项目用了RNN，说一下RNN原理

说了RNN原理，顺便说了LSTM/GRU的出现

7、为什么会出现长时依赖的问题

8、LSTM/GRU如何解决长时依赖的问题

9、写代码：

一个有序数组中查找某个数

一开始写了个遍历查找，面试官说，还能再快吗？

然后写了个二分查找

打断，问个扩展题：问答系统，有200W个FAQ，如何用分类模型做分类

用倒排索引，把FAQ的问题分词，每个词对应多个FAQ。新来的query分词，每个词对应的FAQ拉出来。再在这个里面做分类。

说shuffle，说map、reduce分别分配资源，可以细粒度控制资源占用情况，有利于超大任务平稳正常运行。

6、面试官说，其实是HDFS，正是由于有了分布式文件系统，才可以分布式计算

对，分布式文件系统。数据在哪里计算就在哪里，移动数据变成了移动计算。更高效

7、做题

给定二叉树前序、中序遍历结果。求后序遍历结果

8、一维空间中，2个线段，a1-b1 和a2-b2。判断是否两个线段有交集

他想要的答案是：一个线段里面的大坐标，小于等于另一个线段里面的小坐标。

3、用RNN了，说一下原理

说RNN，顺便说了长时依赖问题，介绍了LSTM，GRU

4、说情感分析的项目

5、每个句子都被打上标签正向或者负向情感，如果我想得出句子中的每个词的情感倾向，怎么做？

我不清楚该怎么做，就如下扯乎：

认为每个句子的情感倾向由每个词的情感倾向打分相加而得。

有的词正向：+1，+2，+3...

有的词负向：-1，-2，-3...

经过RNN，每一时刻的输出。。。扯完我现在想都想不通了。。。

3、用RNN了，说一下原理

4、问RNN怎么训练的？

5、RNN的输入是什么呢？

有word2vec训练的词向量库，一个句子分词后，把词都换成对应的向量输入

6、继续说项目

7、项目用到聚类了？介绍一下

8、说文本情感分类项目，文本向量用tf-idf这种有什么问题没有？

有，不能捕获到上下文之间的联系。以后尝试用doc2vec这种。

9、了解bagging和boosting吗？