大量面经总结(包括牛客网的和我听来的)

## 前言

谢谢牛客网帮助我成功拿到心仪的offer（自然语言算法工程师），也感觉各位大佬分享的面经，所以想回馈一波。在这期间我找到很多面经资料，自己用代码过滤整理了出来。我个人觉得这个资料是十分有用的，我希望也能帮助到各位。祝大家也能够早日找到心仪的工作！

## 目录

1. Python
2. C++
3. 智商题
4. 大数据
5. 计算机基础
6. 概率题
7. HR常问问题
8. 开放题
9. 机器学习
10. 编程题

## Python

* Python的元组和列表的区别。
* a = [1, 2, 3, 4], b = a, b[0] = 100, 请问print(a)结果是什么
* list是怎样实现的。
* list有哪几种添加元素的方法，能否从表头插入元素？
* 如何提高Python的运行效率
* 如何获取list中最后一个元素
* 常用的数据结构及应用场景（list，dict，tuple）

## C++

* Makefile文件，提示未定义的引用，是什么原因（我答的是使用C库忘记加extern，其实应该是没有在makefile指定编译顺序）
* STL中set怎么实现的，假设有“I like love”三个词，如何存。每个节点是直接指向这个单词的指针吗）
* STL中vector是怎样实现的
* [const](http://blog.csdn.net/eric_jo/article/details/4138548)
* [虚函数](http://www.cnblogs.com/longlybits/articles/2386175.html)
* c++如何实现一个接口？（抽象类、纯虚函数）
* c++的数据成员的可见性，继承到子类之后的可见性（这里我是分了不同继承方式讨论的），子类友原函数对父类private能否可见。
* g++中-L,-I,-l的作用，有什么区别。-l指定链接库的时候，如何a库依赖b库，是否a库必须放在b库前面
* 传递一个指针进某函数体内，为什么不能对它重新分配空间，如果想要分配，应该怎么做？（指针的指针）
* 如何想让变量a=100的时候中断，如何写gdb代码
* 如何用gdb调试core文件，
* 对stl的了解程度，map的内部实现原理，为什么选择红黑树，红黑树的由来，与平衡二叉树的区别
* 拷贝构造函数和重载=符分别在什么情况下被调用，实现有什么区别
* 是否有用C++写过实际的工程项目。
* 程序有错误如何调试（回答打log,如何段错误，gdb调试core文件）
* 虚函数的目的，虚函数和模板类的区别，如何找到虚函数
* 说一下TreeMap的实现原理？红黑树的性质？红黑树遍历方式有哪些？如果key冲突如何解决？

## 智商题

* 100张牌，每次只能抽一张，抽过的牌会丢掉，怎么选出最大的牌。
* 36匹马，6条跑道，选出最快3匹，最少赛多少场？
* 5个海盗抢到了100颗宝石，每一颗都一样的大小和价值连城。他们决定：抽签决定自己的号码（1，2，3，4，5）。首先，由1号提出分配方案（你抽到1号），然后大家5人进行表决，当且仅当超过半数的人同意时，按照他的提案进行分配，否则将被扔入大海喂鲨鱼。 如果1号死后，再由2号提出分配方案，依此类推。条件：每颗宝石都是一样的价值。海盗都想保命，尽量多得宝石，尽量多杀人。问题：你会提出怎样的分配方案才能够使自己的收益最大化？
* 一个人要过一座80米的桥，每走一米需要吃一颗豆子，他最多可以装60颗豆子，问最少需要吃多少颗豆子才能走完桥？证明一下为什么你给的答案是最少的？桥长81米呢？当桥长n米，最多装m颗的时候结果用公式怎么表示？
* 一个绳子烧完需要1个小时，假设所有绳子的材质都不一样，也不均匀，怎么取出1小时加 15分钟。
* 把1~9这9个数填入九格宫里,使每一横、竖、斜相等。
* 有100个黑球，100个白球。两个桶，桶的容量无限，每个球都可以任意放在任何一个桶中，没有限制，请设计一种分配方法，使得白黑球分配到两个桶之后, 某个人从某个桶中取出的球是白球的概率最大化。（这个人去第一个桶取球的概率是1/2,第二个桶也是1/2）
* 有1亿个货物，不能单个单个检测，只能通过两两对比来找出其中的次品，请设计一个算法来找出次品。
* 有25匹马 ，5个跑道，一次只能比5匹马，得到跑得最快的前3，至少需要比几次？
* 有3盏灯，房间外有3个开关，你只有1次机会进入房间，怎么判断哪个开关对应哪盏灯？
* 给一堆螺母和螺栓，它们可以一一对应，但是现在顺序乱了，只能用螺母和螺栓比较，将它们一一对应起来。

## 大数据

* 100亿数字，怎么统计前100大的？
* 10亿个url，每个url大小小于56B，要求去重，内存4G。
* 1KW句子算相似度（还是那套分块+hash/建索引，但是因为本人不是做这个的，文本处理根本说一片空白，所以就不误导大家了），之后就是一直围绕大数据的题目不断深化。
* Q1：给定一个1T的单词文件，文件中每一行为一个单词，单词无序且有重复，当前有5台计算机。请问如何统计词频？
* Q2：每台计算机需要计算200G左右的文件，内存无法存放200G内容，那么如何统计这些文件的词频？
* Q3：如何将1T的文件均匀地分配给5台机器，且每台机器统计完词频生成的文件只需要拼接起来即可（即每台机器统计的单词不出现在其他机器中）
* 一个大文件A和一个小文件B，里面存的是单词，要求出在文件B中但不在文件A中的单词。然后大文件A是无法直接存到内存中的。
* 一道题目是如果有一个人注册一个qq，如何保证这个qq号码和之前已存在的qq号码不重复呢？
* 扔硬币，连续出现两次正面即结束，问扔的次数期望
* 有100W个集合，每个集合中的word是同义词，同义词具有传递性， 比如集合1中有word a, 集合2中也有word a, 则集合1，2中所有词都是同义词，对这100W个集合进行归并，同义词都在一个集合当中。
* 有几个 G 的文本，每行记录了访问 ip 的 log ，如何快速统计 ip 出现次数最高的 10 个 ip，如果只用 linux 指令又该怎么解决；
* 海量数据的topk问题

## 计算机基础

* Linux下的一些指令，$$（进程id），$?（上一条命令退出时状态），怎么查看进程，按照内存大小，CPU占用排序等等。
* Linux的命令：pwd、ln、which
* Linux线程通信
* hash表是怎么实现的？有冲突的时候怎么处理？
* linux 文件词频统计
* 介绍一下hash，怎么解决冲突。
* 你说一下hashmap的原理
* 内存泄露出现原因。
* 悲观锁乐观锁
* 把两个表按id合并怎么搞？
* 数据库transaction
* 浅拷贝深拷贝
* 第二题是两题 sql ，涉及 join,group by,max,min,sum,count 等操作的结合，以及同个题目多种写法。
* 线程安全是什么意思？新线程什么情况下会影响原有线程？
* 网络基础TCP三次握手
* 计算机网络：描述他发一句hello world到我这边显示，中间经历了哪些过程，我从应用层开始一层层往下分析答的，主要说http和tcp，网络层和链路层有些忘，但主要的几个协议和子网划分什么的也答了，面试官比较满意
* 词向量的推导，混合高斯，linux硬链接，三次握手，linux inode
* 进程线程的区别

## 概率题

* 100人坐飞机，第一个乘客在座位中随便选一个坐下，第100人正确坐到自己坐位的概率是？
* X是一个以p的概率产生1,1-p的概率产生0的随机变量，利用X产生1/2概率是0,1/2概率是1的随机变量。
* X，Y均服存于 [0,1] 的均匀分布，求X+Y。
* 一个国家重男轻女，只要生了女孩就继续生，直到生出男孩为止，问这个国家的男女比例？
* 一个有7个格子的环，三种颜色染色，相邻不能颜色重复，问多少种方案
* 一个袋子里有很多种颜色的球，其中抽红球的概率为1/4，现在有放回地抽10个球，其中7个球为红球的概率是多少？
* 一枚硬币，扔了一亿次都是正面朝上，再扔一次反面朝上的概率是多少？
* 一道概率题，54张牌，平均分成三堆，大小王在同一堆的概率？
* 一道概率题，一个六位的密码，由0~9组成，问你正过来看和倒过来看密码是一样的概率。
* 一道组合数学题。10盏灯，灭三盏，两头的必须亮着，不能灭掉相邻的两盏灯，问组合数？
* 三个硬币，分别是正正，反反，正反。随机抛一个硬币，结果是正面，问选的是那个硬币
* 个人玩游戏，100个球，每次挑5个，如何保证必胜。52张牌，四个人抽，黑桃A和红桃A同时在一个人手里的概率。
* 好像是问有70%的人喜欢玩游戏，30%的人不喜欢玩游戏，现在推送的资源必须是50%游戏，50%非游戏。问怎么分配比较合理。
* 有n个elements和1个Compare(A, B)函数，用Compare函数作为排序算法中的比较算子给elements排序。Compare函数有p的可能比较错。排序完取Top m个元素，本来就在Top m并被正确分在Top m的元素个数是x。问x的数学期望。
* 有两个随机数产生器，R1以0.7的概率产生1，以0.3的概率产生0，而R2以0.3的概率产生1，0.7的概率产生0.问如何组合这两种产生器，使新得到的随机数产生器以0.5的概率产生1，0.5的概率产生0。随机数产生器可复用。
* 有两枚硬币A和B，A正面的概率为0.6，B正面的概率为0.5.现在扔了一枚硬币显示为正面，问：该枚硬币是A的概率是多少？
* 概率题：有种癌症，早期的治愈率为0.8，中期的治愈率为0.5，晚期的治愈率为0.2.若早期没治好就会转为中期，中期没治好就会变成晚期。现在有一个人被诊断为癌症早期，然后被治愈了，问他被误诊为癌症的概率是多少？
* 给一个函数，返回0和1，概率为p和1-p，请你实现一个函数，使得返回01概率一样。
* 给定一个分类器p，它有0.5的概率输出1，0.5的概率输出0。Q1：如何生成一个分类器使该分类器输出1的概率为0.25，输出0的概率为0.75？Q2：如何生成一个分类器使该分类器输出1的概率为0.3，输出0的概率为0.7？
* 问了一个概率题 54张牌，分成6份，每份9张牌，大小王在一起的概率

## HR常问问题

* 为什么不读博、对读博报以什么态度。
* 为什么选择百度，谷歌百度都给你offer你选哪个。
* 为什么选择跨专业学计算机？
* 为什么选择阿里
* 以后可能要学习很多新技术，你怎么看。
* 你平时喜欢做什么？看过哪些书？最近在看什么书？
* 你觉得最有挑战的项目是什么。
* 你觉得最难忘的事情是什么？
* 你认为你的优（缺）点是什么。
* 你还有什么想问的？
* 加班怎么看。
* 印象最深刻的事？
* 压力最大的情况是什么时候。
* 在面试过程中觉得自己那些当面有进步
* 场景分析题，有一个任务给你，要求一个月完成，但是以目前的能力一个月完成不了，现在你知道有一个同事擅长这部分工作，但是他有自己的活，帮助你就可能耽误他的进度，问你咋办。
* 大学令你觉得最不爽的事情是什么
* 如何学习的？
* 如何看待加班。
* 实习期间项目，在组内担任的角色，是否熟悉其他组员的工作。
* 家庭教育观念？
* 家里什么情况？独生子女？
* 将来的职业规划？
* 工作地点
* 工作地点的问题
* 平时有什么兴趣爱好。
* 我觉得我会先去专心钻研技术，到达一定的
* 最后问了一下我兴趣爱好
* 有什么问题问我。
* 有没其他offer
* 有没有想过去创业公司
* 现在在哪里实习？实习主要做些什么？
* 简单介绍一下自己
* 聊聊offer情况，有什么考虑之类的。
* 聊聊实验室生活。
* 能不能来北京
* 自己有什么优点缺点？
* 自己本科生和研究生相比有哪些进步
* 要求用两个字评价大学生涯。
* 讲一下你觉得你突出的地方，有亮点的地方。
* 评价一下你自己的优点缺点？
* 详细介绍项目。
* 说下你的优缺点
* 说说你的经历。
* 说说你自己的性格。
* 说说研究生阶段最有成就的事，遇到问题具体怎么解决的。
* 请你说一下你对应聘该岗位的优势。
* 遇到的最大挫折是什么。
* 问你的职业规划，遇到挑战怎么处理，有没有之前和同事发生过较大分歧。

## 开放题

* 2016年每个项目有个上线和下线时间段，统计每天在线的项目数量
* 一堆问题和答案的pair，算它们的相关性
* 一面现场面，自我介绍加挑一个项目细讲，还有场景题，第一题是QQ添加好友按名称搜索时，怎么区别广告号，诈骗号；
* 为什么之前没有深度网络出现（数据量不够+机器性能）
* 为今日头条设计一个热门评论系统，支持实时更新。
* 从项目中在哪一方面体会最深。
* 假设一个文档，连续的K个词，认为是一个时间窗口，一个时间窗口的词有关系，如何得到所有的时间窗口。
* 假设你拥有一切搜索数据，问怎么在不同场景下进行推荐，具体场景忘了（核心点：共线性、语义相似度、主题聚类等等）
* 假设有100W个单词，如何存储（我答的是trie树，面试官问每个节点会有很多子节点，每个子节点是一个指针，占用8个字节，如何节省空间，我说不知道，面试官提示双数组trie树）
* 假设要对一场nba球赛进行自动解说，会遇到哪些困难，又该怎么解决呢？
* 做过哪些项目？项目中遇到哪些难点，你是怎样解决的？
* 关于集群调度的一些经验 trick 掌握多少；
* 分词时，为了提高效率，怎么存储词典？（键树）如何压缩存储？
* 在微信的场景下，如何判断用户的职业？开放问题
* 场景题如何鉴别淘宝上卖假货的商家，价格维度可以用什么策略等
* 如何做一个新闻推荐
* 如何在语料中寻找频繁出现的字串，分析复杂度。
* 如何用尽可能少的样本训练模型同时又保证模型的性能；
* 如何预测双十一支付宝的负载峰值。
* 对推荐算法的未来看法。
* 平面上有n个点，让你设计一个数据结构，能够返回这个这n个点中距离某特定点最近的一个点。一开始讲了下kd树，然而太复杂面试官不满意，就讲了一个类似GeoHash的方案。
* 建立一个数据结构，基于此写一段程序用于存储sparse vector，同时编写一个函数实现两个sparse vector的相加运算
* 很多单词，如何计算单词之间的相似度（或者对单词进行分类）
* 怎么预测降雨量。
* 我只有一大批实体词， 如何对他们进行聚类（无监督聚类）， 如何找出这些词中， 哪些词之间有关系， 是强关系还是弱关系， 具体是什么关系，（如刘德华和朱丽倩 属于娱乐分类， 是强关系， 关系为夫妻）
* 拼车软件是如何定价的以及如何优化。
* 推荐算法（基于用户的协同过滤，基于内容的协同过滤）
* 推荐系统的冷启动问题如何解决
* 文本挖掘中，分词算法？如何选取特征？如何进行相似度计算，文本聚类结果如何评估？
* 无给定条件，预测蔬菜价格。
* 有100W个集合，每个集合中有一些词，对于每个集合，找出他是哪些集合的真子集。
* 有一堆已经分好的词，如何去发现新的词？
* 比赛相关问题提特征特征选择等
* 海量的 item 算文本相似度的优化方法；
* 特征工程经验。
* 用两分钟介绍自己的项目，创新点在哪里。
* 用户给三个item（query），如何给出查询网页。
* 第三题是如何鉴别实施诈骗的QQ用户；
* 第二题是微信朋友圈内容的安全鉴别；
* 第四题是如何做反作弊，例如公众号的刷阅读量。
* 系统设计题，给一个query，如何快速从10亿个query中找出和它最相似的 （面试官说可以对每个query找1000个最相似的，存起来，每天离线更新）
* 线性代数：特征线性依赖，出现冗余，会导致什么问题？
* 给一堆数据找找到最佳拟合的直线，数据有较多噪声
* 给你一个系统（面试官好像是无人车部门的），后台的逻辑已经实现了，但是前端加载很慢，怎么检测。
* 给你两个文件a和b，大小大概100M，两个文件每行一个整数，要求找到两个文件中相同的整数，存到文件c里，问我怎样尽快的完成这项工作？
* 给出一个算法实现如何确定快递邮件上的地址，要求从国家到省市到县到乡镇的一个识别，要求效率高（有陷阱，比如有的人把县写到市的前面，有人喜欢写地域名称的省略词比如安徽省写成安徽或者皖）。
* 给定淘宝上同类目同价格范围的两个商品A和B，如何利用淘宝已有的用户、商品数据、搜索数据、评论数据、用户行为数据等所有能拿到的数据进行建模，判断A和B统计平均性价比高低。统计平均性价比的衡量标准是大量曝光，购买者多则高。
* 给很多单词，统计某个子串出现次数，我给的方法还是用Trie，只不过一个单词要分成多个插入到Trie数中就行了。
* 给很多单词，要求统计出现某个前缀出现次数。
* 统计全球会弹钢琴的人数，我用机器学习的思路答的，面试官还比较满意
* 自己项目中有哪些可以迁移到其他领域的东西。
* 讲了讲自己在深度学习的认识，问的问题是几个具体场景的设计，包括怎么从海量数据中提取热点问题。
* 设计 LRU 系统
* 设计一个合理的电梯调度策略，调度两个电梯 ，考虑满足基本的接送需求，满足能耗最小，满足用户等待时间最短
* 设计一个系统可以实时统计任意ip在过去一个小时的访问量；
* 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）
* 设计实现一个git diff
* 说一下最能代表你技术水平的项目吧？
* 项目：具体问了特征怎么做的。
* （难到我了，我想的方法不好，面试告诉我了他的想法，类似于一个进程调度问题，每一时刻只可能有一个用户按按钮，把这条指令接收，判断当前电梯能否满足，能满足就执行，不能满足则放入一个队列里，实际情况还要细化）

## 机器学习

* Boost算法
* 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与随机森林比较。
* 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）
* KKT条件用哪些，完整描述
* KNN（分类与回归）
* L1 与 L2 的区别以及如何解决 L1 求导困难。
* L1和L2函数。
* L1和L2正则相关问题。
* L1和L2正则项，它们间的比较
* 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
* 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？
* LR和SVM有什么区别，libsvm和liblinear有什么区别。
* Logistics与随机森林比较
* Logistics（推导）
* Logistic回归的推导，怎么得到objective function。
* SVM与随机森林比较
* SVM为什么要引入拉格朗日的优化方法。
* SVM原问题和对偶问题关系？
* SVM在哪个地方引入的核函数, 如果用高斯核可以升到多少维。
* SVM怎么防止过拟合
* SVM的目标函数。常用的核函数。
* SVM的过程，讲了推导过程，可能表达不清晰，都是泪
* bagging、adaboost、boosting
* em 与 kmeans 的关系；
* k-means的k怎么取等等
* k-means算法初始点怎么选择？你的项目里面推荐算法是怎么实现的？
* kmeans的原理，优缺点以及改进。
* k折交叉验证中k取值多少有什么关系
* l2惩罚项是怎么减小Overfitting的？l1，l2等范数的通式是什么？他们之间的区别是什么？在什么场景下用什么范数？l1在0处不可导，怎么处理？
* randomforest,GBDT
* rf, gbdt, xgboost的区别。
* softmax公式
* 为什么要做数据归一化？
* 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）
* 什么情况下一定会发生过拟合？
* 什么是贝叶斯估计
* 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
* 介绍SVD、SVD++
* 会哪些机器学习算法
* 信息熵公式
* 假设面试官什么都不懂，详细解释 CNN 的原理；
* 决策树原理
* 决策树处理连续值的方法。
* 决策树如何防止过拟合
* 决策树过拟合哪些方法，前后剪枝
* 分类模型可以做回归分析吗？反过来可以吗？
* 分类模型和回归模型的区别
* 判别模型，生成模型
* 各个模型的Loss function，牛顿学习法、SGD如何训练。
* 因为面我的总监是做nlp的,所以讲了很多rnn、lstm、还有HMM的东西。不算很熟，但是接触过，以前稍微看过一些相关论文，所以还是勉强能聊的。
* 在平面内有坐标已知的若干个点P0...Pn，再给出一个点P，找到离P点最近的点。
* 在模型的训练迭代中，怎么评估效果。
* 如何减少参数（权值共享、VGG的感受野、GoogLeNet的inception）
* 如何防止过拟合（增加数据，减少模型复杂度->正则化）
* 对于同分布的弱分类器，求分类器均值化之后的分布的均值跟方差。
* 对于机器学习你都学了哪些？讲一个印象深的。
* 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
* 归一化方式
* 手写k-means的伪代码。
* 手写k-means的伪代码和代码。（Code）
* 手撕svm硬软间隔对偶的推导
* 手撕逻辑回归（损失函数及更新方式推导）
* 接着写一下信息增益的公式。
* 推一下bp算法等等
* 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。
* 数据挖掘各种算法，以及各种场景下的解决方案
* 是否了解mutual infomation、chi-square、LR前后向、树模型等特征选择方式。
* 是否了解线性加权、bagging、boosting、cascade等模型融合方式
* 有哪些常见的分类器，简单介绍下原理
* 机器学习与深度学习的区别
* 机器学习基础（线性回归与逻辑回归区别等）
* 机器学习：几种树模型的原理和对比，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； k-means 聚类的原理以及缺点及对应的改进；
* 梯度下降牛顿拟牛顿原理
* 梯度下降的优缺点。
* 深度学习和普通机器学习有什么不同？
* 深度学习有很大部分是CNN，给他用通俗的语言解释下卷积的概念，解释下CNN中的优势及原因
* 激活函数的选择（sigmoid->ReLu->LReLU->PReLU）
* 然后20分钟内手写k-means
* 牛顿法、随机梯度下降算法和直接梯度下降算法的区别？
* 牛顿法推导
* 特征选择的方法
* 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）
* 聊聊SVM，这段说了好久，从基本的线性可分到不可分，相关升维，各种核函数，每个是如何实现升。以及出现了XX问题，分析是样本的原因还是其他原因。针对不同情况，采取什么解决方案较好。
* 自己实现过什么机器学习算法
* 解决过拟合的方法有哪些？
* 解释 word2vec 的原理以及哈夫曼树的改进。
* 解释一下过拟合和欠拟合，有哪些方法防止过拟合。
* 让我一步一步地构造决策树，怎么计算信息熵、信息增益、然后C4.5 ID3 CART的区别，还说了一下优缺点
* 详细讨论了样本采样和bagging的问题
* 说一下Adaboost，权值更新公式。当弱分类器是LR时，每个样本的的权重是w1，w2...,写出最终的决策公式。
* 说了一下bagging跟boosting。
* 说明L1L2正则的效果与为什么形成这种情况（L1正则稀疏，L2正则平滑，之后说明就是画图说明正则化）
* 过拟合的解决方法；
* 选个你熟悉的机器学习方法 ，着重介绍一下产生原因，推导公式，背后统计意义什么等等
* 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。
* 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？
* 问了会不会RNN,LSTM。
* 问了很多数据挖掘的基础知识，包括SVM,逻辑回归、EM、K-means等，然后给我很多场景问我遇到这些情况我要怎么来处理数据，怎么进行建模等等，问得很细
* 随机梯度下降，标准梯度
* 随机森林和GBDT的区别？LR的参数怎么求解？有没有最优解？
* 随机森林（Bagging+CART）

## 编程题

* 1~n这n个数现在去掉两个，如何找到去掉的两个数。 假设去掉的两个数是a和b，那么通过求和，平方和可以知道a+b和a^2+b^2，然后解方程就行了。
* char a[4] = {1, 2, 3, 4}; char \*b = a; b[0] = 100; 请问输出a的结果是什么？
* 一个 N\*M 的矩阵，从左上走到右下最小需要（N+M）步走完，问一共有多少种走法。
* 一个严格递增的数组，将前缀取一部分放在后面，在修改后的数组上找到最小的数。(剑指Offer原题)
* 一个大写字符串如ABABB（len<1000），代表游客进游乐场的顺序及从哪个入口进入，要求每个入口(不多于26个入口)从第一个游客直到该入口的最后一个游客，检票员都不能离开，问最少检票人数K。
* 一个字符数组中，每个字符都出现了3次，只有一个出现了2次，如果快速找出这个出现2次的？
* 一个字符矩阵，只可能是R,G,B三种字符。判断是否满足某个条件。这个条件是每种符号连成一个长方体，三个长方体长宽一致,且横着平行
* 一个广告，它有一个id，一个上线时间，一个下线时间，现在我有很多这样的广告，如果现在给你一个时间，告诉我有多少个广告在这个时间在线的
* 一个数据流中，如何采样得到100个数，保证采样得到的100个数是随机的？
* 一个数组中某个数出现次数大于一半，最快找出该数。
* 一个数组只有一个数字是单独出现，其他出现了三次。
* 一个数组存着1-1000连续的整数，假如我取出其中一个数，怎么能快速找到（用类二分查找）
* 一个数组存着负数与正数，将正数放在前面，负数放在后面
* 一个运算序列只有+、\*、数字，计算运算序列的结果。（Code）
* 一堆ip地址区间，不会重叠，来一个新的ip地址，看它在不在，在哪个区间。
* 一维数组，swap 其中的几对数字（每个数字只属于一次 swap 操作），实现查找（与二分有关）；
* 一维有序数组，经过循环位移后，最小的数出现在数列中间，如果原数组严格递增或递减，如何找这个最小数；
* 一维有序数组，经过循环位移后，最小的数出现在数列中间，如果原数组严格递增，如何找这个最小数。
* 一维有序数组，经过循环位移后，最小的数出现在数列中间，如果原数组非严格递增或递减，如何找这个最小数；
* 一维有序数组，经过循环位移后，最小的数出现在数列中间，数组可能是递增、递减、递减后递增、递增后递减四种情况，递增递减都是非严格的，如果有转折点，返回转折点的值，否则返回-1；
* 一道题：给定一个整数数组，里面有两个数相同，其他数都是不同的，如何尽快找到这两个数（答，用hash表，O（N），有更好的方法么？）
* 一题是多位数用链表存储（ e.g. 123 用 1->2->3 存储），实现相加功能函数
* 不创建临时产量换两个数
* 两个同样大小有序数组求中位数，写代码
* 两个大整数相乘。（Code）
* 两棵树相加——对应位置两棵树都有值则相加，对应位置只有一棵树有值则取该值；
* 中序遍历二叉树，利用O(1)空间统计遍历的每个节点的层次。（Bug Free Code）
* 中缀表达式转逆波兰表达式，逆波兰表达式求值；
* 为分析用户行为，系统常需存储用户的一些 query ，但因 query 非常多，故系统不能全存，设系统每天只存 m 个 query ，现设计一个算法，对用户请求的 query 进行随机选择 m 个，请给一个方案，使得每个 query 被抽中的概率相等，并分析之，注意：不到最后一刻，并不知用户的总请求量。
* 二分查找
* 二分查找，查找target，在区间[start，end]之间，如果有重复元素，返回最后一个下标，其他情况返回-1
* 二叉树前序递归遍历算法（手写代码）
* 二叉树的前中后遍历
* 二叉树的文件存储，也就是序列化。
* 二叉树遍历，描述下层序遍历。
* 二维数组，每行递增，每列递增，任意交换其中的两数，发现并恢复。
* 二维数组，每行递增，每列递增，实现查找。
* 二维数组，每行递增，每列递增，求第k大的数。
* 什么样的数据结构可以满足多次插入删除，取最小数，给出时间复杂度。
* 介绍二叉树前序遍历非递归遍历算法（手写代码）
* 介绍大顶堆和小顶堆
* 从一组数中找出和为sum的三个数（leetcode原题，先sort再找，并且剪枝），写代码，四个数呢？说思路。
* 假设有个M\*N的方格，从最左下方开始往最右上方走，每次只能往右或者往上，问有多少种走法，假设中间有若干个格子不能走，又有多少种走法。
* 允许两个元素交换一次的最大连续子序列和。
* 全排列
* 全排列。
* 冒泡排序(手写代码)
* 写 find 函数，在目标串中匹配模式串（要考虑中文字符的情况）
* 写一个二叉树的非递归的后续遍历
* 写一个简单的正则匹配表达式(将文本中的123.4匹配出来)
* 写个动态规划，最长公共子序列
* 判断一个字符串是否为另外一个字符串旋转之后的字符串
* 前k大的数
* 单链表的翻转
* 去掉连续的重复数字，输出新数组，例如：1，2，2，2，1，3，5——> 3，5。
* 去除字符串S1中的字符使得最终的字符串S2不包含’ab’和’c’。（Code）
* 合法括号匹配
* 在一个字符串中，找出最长的无重复字符的字串
* 在二叉树结点结构中加一个指针域，使其指向层次遍历的下一个结点，特别地，每一层的最后一个结点为空。（Code）
* 堆排序(手写代码)
* 堆是怎么调整的。
* 复杂链表的复制。
* 如果给出一个二叉搜索树的后续能不能建立（可以，因为只要将遍历结果排序就可以得到中序结果）。
* 字符串反转（手写代码）
* 字符串移位，给出字符串abc##dfg##gh，实现将所有#移至字符串串头。输出####abcdfggh。
* 字符串转整数
* 字符串，给一个url，求中间的site
* 字符串，给一个url，求中间的site。
* 定义满足$n=x^a+y^b$（$x，y，a，b$是非负整数）的n是神奇数。如$4 = 2^0 + 3^1,17 = 2^3 + 3^2$。输入l和r，请求出闭区间$[l,r]$里，最长的一段不含有神奇数的连续区间长度。$x,y,l,r<=10^{18},x>=2,y>=2$，如$3\ 5\ 10\ 22$，在$[10,22]$区间内，$x=3,y=5$的条件下，区间内[14]是神奇数，所以最长的区间是$[15,22]$长度为$8$，如$2，3，1，10$，在$[1,10]$区间内，$x=2，y=3$的条件下，$2，3，4，5，7，9$都是神奇数，所以最长的区间只有长度$1$。
* 实现栈，使得 添加、删除、max 操作的复杂度为 O(1)。
* 对于一个字符串，请设计一个算法，只在字符串的单词间做逆序调整，也就是说，字符串由一些由空格分隔的部分组成，你需要将这些部分逆序。给定一个原字符串A和它的长度，请返回逆序后的字符串。
* 对于一个字符串，请设计一个算法，将字符串的长度为len的前缀平移到字符串的最后。
* 寻找字符串中第一个只出现一次的字符；
* 将字符串连续重复出现的字符删到只剩一个，这个可以用双指针，时间复杂度n，空间复杂度1。
* 常用排序算法的时间和空间复杂度
* 平衡二叉树是什么
* 归并排序(手写代码)
* 快速排序(手写代码)
* 快速排序+二分查找
* 手写快排(easy)
* 打印数组的组合数。
* 打印螺旋数组；
* 把一个bst转化成一个双向链表。
* 把一个字符串的大写字母放到字符串的后面，各个字符的相对位置不变，不能申请额外的空间。例如AbcDeFGhi ->bceiADFG
* 排序二叉树转双向链表。（Code）
* 描述Dijkstra最短路径算法
* 插入排序(手写代码)
* 数列中找第 k 大的数字（与快排或堆排序有关）
* 数据解压缩，3(a4(ab)) -> aababababaababababaabababab；
* 数组有只有一个数出现一次，其他数都出现三次，找出那个数。
* 旋转数组
* 最少时间复杂度求数组中第k大的数。（Code）
* 最短路径代码。
* 最长公共子串（动态规划有关）；
* 最长公共子序列
* 有一堆无向好友列表 1-2， 3-4, 2-3 之类的，问能不能把这些用户划分两组，组内都不互为好友。
* 有序数组寻找和为某数的一对数字；
* 正数数组，找三个数使积最小，问有多少种选择。
* 母鸡、公鸡和小鸡问题：公鸡五块一只，母鸡三块一只，小鸡一块三只，用100元买100只鸡的所有方法。
* 求double类型的二进制1的个数。
* 求二叉树最近公共祖先(leetcode原题)
* 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）
* 用一个队列，将每个二叉树的root先放入队列。
* 用数组实现队列，各操作的复杂度分析。
* 用速度不同的指针可以判断链表中是否有环，问两速度满足怎样的关系可以保证发现环。
* 直接插入排序写代码
* 看段代码，问输出是啥。(就是段求二进制中1的个数)
* 矩阵求最长连续递增的路径长度
* 矩阵求最长连续递增的路径长度。
* 第一题是链表倒数第 k 节点；第二题是二叉树打印路径，第三题是矩阵中将 0 元素所在行列全置 0 的最优空间解法
* 第二轮是写出一个算法输出二叉树的 s 序列，何为 s 序列，比如现在有个二叉树 1-2,3-4,5 6,7 这是一颗完全二叉树， S 序列输出就是按照 1237654 这个顺序输出，用两个栈就能实现比较简单。
* 算法题，也只记得一个了：存在一个数组，大小98，里面的元素均为在[1,100]，且无重复， 不申请额外空间的情况下，在时间复杂度为O(N)情况下，找出缺失的两个元素值。
* 给一个n\*n的矩阵，矩阵中满足每行每列都是递增的，要查找矩阵是否存在某个数.(leetcode原题)
* 给一个数组，只有一个元素出现了一次，其他都出现了两次，找出出现一次的数。
* 给一个数组，数组种存在一种数，它的左边都比它小，右边都比它大，找出所有这些数的位置。
* 给一个股票，n天的价格，只能两次买入卖出，而且只能只能先卖再买，问最多赚多少钱？
* 给一个股票，n天的价格，只能进行一次买入和卖出，问最多赚多少钱？
* 给一个股票，n天的价格，可以买入卖出k次，而且只能只能先卖再买，问最多赚多少钱？
* 给一个股票，n天的价格，可以无限次买入卖出，问最多赚多少钱？
* 给了一个链表，第1个结点标号为1，把链表中标号在M到N区间的部分反转。
* 给你一个无重复的数组输出全排列。
* 给你一颗二叉树按层输出每一层输出后都换行
* 给出一个二维矩阵，从（0,0）出发走到右下角，只能向右或向下走，找到一条路径，是这条路径上的总和最大。
* 给出一段代码问代码作用（二进制数据1的个数）
* 给出一颗二叉树，两个叶节点，找到这两个叶节点互连通的一条最短路径。
* 给定一个数组，只有一个元素出现了一次，其他都出现了3次，找出出现一次的数。
* 给定一个数组，有两个元素出现了一次，其他都出现了两次，找出两个出现一次的数。
* 给定一个正整数向量，判断这个向量是否存在一个片段，使得反转这个片段后能够使该向量升序排列。如：[1, 2, 4, 3]，就可以通过反转[4, 3]使得向量变为[1, 2, 3, 4]。
* 给定二叉树的先序跟后序遍历，能不能将二叉树重建（不能，因为先序：父节点-左节点-右节点，后序：左节点-右节点-父节点，两者的拓扑序列是一样的，所以无法建立）
* 给定循环递增数组 $a=[7,8,9,1,2,3]$和一个值$k=2$,返回该值得再数组中的下标。
* 给定数组A[]={1,4,7,...}和一个数T。求和为T的A中的数最少要几个。A中的数可复用。 我写了个递归，面试官不建议使用，因为效率不高。但没有反对。
* 给定数组，寻找 next big（堆排序有关）；
* 给我一个数组［1，2，5，10，20，50，100］，可以从里面取若干个数，要求得出和为100的不同取法有多少？（说出思路）
* 统计数列中的逆序对（归并排序有关）；
* 编程题：实现求正整数平方根整数部分的函数（使用梯度下降）
* 翻转二叉树（Code）
* 若干个二叉树，如何按照层序遍历
* 设 rand （ s ， t ）返回 [s,t] 之间的随机小数，利用该函数在一个半径为 R 的圆内找随机 n 个点，并给出时间复杂度分析。
* 输入一个大长方形，长宽ab，和一堆小长方形。选择两个小长方形，它能放进大长方形，而这个小长方形面积和最大
* 输入一个宿舍楼亮灯描述图，计算把所有灯关掉的最短时间，管理员起点在左下角，只能在最左或最右的楼梯往上一层，不可往下一层。每次往上一层花费1分钟，每次往左或往右一间宿舍花费1分钟，关灯不花时间。输入的高<=15，宽<=100。  
  如图:
* 0010
* 0100

从左下角开始，最短花费时间是先往右（关灯），再往左，再上一层，再往右两次（关灯）完成：5

再如：

001000

000010

000010

最短时间是先往右四次（关灯），往右一次，上一层，往左一次（关灯），往右一次，上一层，往左三次（关灯），完成，12

* 输入两个正数数组，在两个数组分别选一个数，要求第一个数组选的数的下标小于第二个数组选的数的下标。使得两个数的乘积最大。
* 输出字符串中的所有重复子串，例如：abcab，输出: a, b, ab
* 连续子数组最大和
* 迷宫的深度搜索、广度搜索；
* 选取任意数据结构实现添加、删除、随机返回三个功能，分析复杂度。
* 选择排序(手写代码)
* 链表上的快速排序。
* 长度为N的序列Sequence=abc......Z，问有多少不同的二叉树形态中序遍历是这个。（Code）
* 问了一两个算法题，记不清了，只记得其中一个是：找数组中2个出现两次的数字，还有3个两次的数字
* 问了一个1的平方加到100的平方结果
* 非常经典的0-1背包问题