机器学习

* 有n（n未知且很大）个query，从中随机选择k个，要求每个query被选择的概率相等，均为k/n?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | … | K |

假设第i个（i>k）个query到来，随机从[1, i]中选择一个随机数r, 如果r大于k则直接丢弃，否则较少第r个数据和第i个数据

证明：

第i个数据被换入数组的概率是k/i, 进一步的，数组中第j个数字被选中替换的概率是1/k,从而第j个数字被替换掉的概率是k/i\*1/k=1/i

归纳：假设经过i-1轮后，任意元素x在数组中的概率是k/(i-1), 则在第i轮中，该元素被替换的概率为1/i,其不被替换的概率为(i-1)/i, 综上，该元素在第i轮后仍然在数组中的概率为k/(i-1)\*(i-1)/I = k/i, 结论得证。

* 外部排序（归并）

腾讯面试：

1亿2000万行，如何分布式用SVD分解

LR、决策树、GDBT、SVM，SM

个人经历感觉，面试过程必备下面四个方面：

1.研究生一般会先问项目，所以对项目的描述一定要重点突出，细节部分要了如指掌，要将项目的背景提升到一个大的高度，最重要的是体现自己的工作量（包括遇到的困难，如何解决）。

2.基础知识要熟悉，这里说的并不是所有的基础知识，常问也就那么几个，为了面试针对性的复习就好了（后面有干货）。

3.手写代码技能，个人感觉要在代码中体现两点，边界条件考虑；能体现思路清晰，不一定每个函数都得实现，但算法思路一定要体现。另外还有必要的注释会帮助面试官理解你的思路。

4.最后来一碗鸡汤：对自己想要的东西，要不顾一切的去追求，不要轻易放弃。

谢谢牛友听我啰嗦完，下面是本次腾讯面试的题目汇总，有些是身边同学提供的，没有给出答案，相信牛友们会自己找的。

面试题汇总（干货）

计算机网络相关

1 TCP三次握手、四次挥手

2 TCP滑动窗口机制

3 TCP拥塞控制机制

4 socket模型

C/C++相关

1 虚析构、模板和宏

2 虚函数实现机制

3 vector与list的区别，map是如何实现的，查找效率是多少

4 extern 关键字有什么用

5 malloc和new的区别，能否malloc(1.2G)

linux以及操作系统相关

1 内存池实现

2 进程间通信机制

3 Linux ps命令，以及看内存当前使用状态的命令

4 进程与线程的区别，共享的数据

5 进程的内存空间

算法与数据结构 （手写代码实现）

1 大整数加、减、乘、除、求模运算实现

2 很多整数，找其中出现次数最多的那个数

3 单链表翻转（两个指针如何实现）、查找、删除、插入以及双向链表、有序链表合并

4 判断一个整数是否是2的整数次幂.(n&(n-1))

5 二分查找（注意边界条件）

6 常见排序算法的实现以及稳定性（快排跟归并考的很多）

7 字符串翻转（O(n)）、匹配（KMP算法）

8 最**长递增子序列**（nlogn的算法）

9 链表判断是否有环，环的入口，两个链表是否相交（快慢指针）。

10 指定一个数组，求2个数的和等于指定的和（某一个数），如果是3,4,5，n个等于个的和（某一个数）呢？（可以看作背包问题）

11 跳台阶问题

其他

1 红黑树的性质以及插入和删除

2 解析XML文件

3 千万级的用户，提供一个服务，该服务有很多模块，现在有一个底层模块需要优化，问怎么实现，在不影响其他服务模块以及用户体验的情况下。（面IEG）

4 卡特兰数以及公式推导（应多很多）

5 未知大小的文件，翻转整个文件

6 如果内存中有个cache存储qq号和最近登录时间问怎么样做hit和淘汰

7 检测短信敏感词

8 大数据问题

9 C++、java和PHP有什么本质区别