**2011Alibaba数据分析师（实习）试题解析**

投稿人/作者: wrchow 发布时间：2012-04-30 11:48:03   [投稿到ChinaKDD](http://www.chinakdd.com/post.html)

**一、异常值是指什么？请列举1种识别连续型变量异常值的方法？**

异常值（Outlier） 是指样本中的个别值，其数值明显偏离所属样本的其余观测值。在数理统计里一般是指一组观测值中与平均值的偏差超过两倍标准差的测定值。  
Grubbs’ test（是以Frank E. Grubbs命名的），又叫maximum normed residual test，是一种用于单变量数据集异常值识别的统计检测，它假定数据集来自正态分布的总体。  
未知总体标准差σ，在五种检验法中，优劣次序为：t检验法、格拉布斯检验法、峰度检验法、狄克逊检验法、偏度检验法。

**二、什么是聚类分析？聚类算法有哪几种？请选择一种详细描述其计算原理和步骤。**

聚类分析(cluster analysis)是一组将研究对象分为相对同质的群组(clusters)的统计分析技术。 聚类分析也叫分类分析(classification analysis)或数值分类(numerical taxonomy)。聚类与分类的不同在于，聚类所要求划分的类是未知的。  
聚类分析计算方法主要有： 层次的方法（hierarchical method）、划分方法（partitioning method）、基于密度的方法（density-based method）、基于网格的方法（grid-based method）、基于模型的方法（model-based method）等。其中，前两种算法是利用统计学定义的距离进行度量。  
k-means 算法的工作过程说明如下：首先从n个数据对象任意选择 k 个对象作为初始聚类中心；而对于所剩下其它对象，则根据它们与这些聚类中心的相似度（距离），分别将它们分配给与其最相似的（聚类中心所代表的）聚类；然 后再计算每个所获新聚类的聚类中心（该聚类中所有对象的均值）；不断重复这一过程直到标准测度函数开始收敛为止。一般都采用均方差作为标准测度函数. k个聚类具有以下特点：各聚类本身尽可能的紧凑，而各聚类之间尽可能的分开。  
其流程如下：  
（1）从 n个数据对象任意选择 k 个对象作为初始聚类中心； 　　　　  
（2）根据每个聚类对象的均值（中心对象），计算每个对象与这些中心对象的距离；并根据最小距离重新对相应对象进行划分； 　　  
（3）重新计算每个（有变化）聚类的均值（中心对象）；  
（4）循环（2）、（3）直到每个聚类不再发生变化为止（标准测量函数收敛）。  
优 点：本算法确定的K 个划分到达平方误差最小。当聚类是密集的，且类与类之间区别明显时，效果较好。对于处理大数据集，这个算法是相对可伸缩和高效的，计算的复杂度为 O(NKt)，其中N是数据对象的数目，t是迭代的次数。一般来说，K<<N，t<<N 。  
缺点：1. K 是事先给定的，但非常难以选定；2. 初始聚类中心的选择对聚类结果有较大的影响。

**三、根据要求写出SQL  
表A结构如下：**  
**Member\_ID** （用户的ID，字符型）  
**Log\_time** （用户访问页面时间，日期型（只有一天的数据））  
**URL** （访问的页面地址，字符型）  
**要求：提取出每个用户访问的第一个URL（按时间最早），形成一个新表（新表名为B，表结构和表A一致）**

**create** table B as **select**Member\_ID, min(Log\_time), URL from A **group by**Member\_ID ;

**四、销售数据分析  
以下是一家B2C电子商务网站的一周销售数据，该网站主要用户群是办公室女性，销售额主要集中在5种产品上，如果你是这家公司的分析师，  
a) 从数据中，你看到了什么问题？你觉得背后的原因是什么？  
b) 如果你的老板要求你提出一个运营改进计划，你会怎么做？  
表如下：一组每天某网站的销售数据**  
[http://saslist.com/wrchow/files/2011/03/ali-300x41.jpg](http://saslist.com/wrchow/files/2011/03/ali.jpg)

a) 从这一周的数据可以看出，周末的销售额明显偏低。这其中的原因，可以从两个角度来看：站在消费者的角度，周末可能不用上班，因而也没有购买该产品的欲望；站在产品的角度来看，该产品不能在周末的时候引起消费者足够的注意力。  
b) 针对该问题背后的两方面原因，我的运营改进计划也分两方面：一是，针对消费者周末没有购买欲望的心理，进行引导提醒消费者周末就应该准备好该产品；二是，通过该产品的一些类似于打折促销等活动来提升该产品在周末的人气和购买力。

**五、用户调研  
某公司针对A、B、C三类客户，提出了一种统一的改进计划，用于提升客户的周消费次数，需要你来制定一个事前试验方案，来支持决策，请你思考下列问题：  
a) 试验需要为决策提供什么样的信息？  
c) 按照上述目的，请写出你的数据抽样方法、需要采集的数据指标项，以及你选择的统计方法。**

a) 试验要能证明该改进计划能显著提升A、B、C三类客户的周消费次数。

b) 根据三类客户的数量，采用分层比例抽样；  
需要采集的数据指标项有：客户类别，改进计划前周消费次数，改进计划后周消费次数；  
选用统计方法为：分别针对A、B、C三类客户，进行改进前和后的周消费次数的，两独立样本T-检验（two-sample t-test）。