1. **什么是 ab test [业务]**

Ab test，即有两个即将面对大众的设计版本(A 和 B)。通过小范围发布，得到并比较这两个版本之间你所关心的数据(转化率，业绩，跳出率等)，最后选择效果最好的版本。

具体来说，在对产品进行 ab test 时，我们可以为同一个优化目标（如优化购买转化率）制定两个方案，选出一部分用户，让 50%的用户使用 A 方案，另外 50%的用户使用 B 方案，统计对比不同方案的转化率、点击量、留存率等指标，以判断不用方案的优劣并进行决策，从而提升转化率。

2. **估算北京市有多少家餐馆 [估算]**

解析：此题思路与上两题一致

角度一（需求层面）

北京市餐馆数量=每顿饭在餐厅吃饭的人数÷每家餐厅的客容量

已知，北京有常驻人口约 2000 万人。每人每两天约有 1 顿饭在餐厅吃，每顿饭在餐厅吃饭的人数为 2000÷6=333 万人。

根据常识，一家小型餐厅可以容纳约 30 人，一家中大型餐厅可以容纳约 100 人。小型餐厅与中大型餐厅的比例为 5：1，那么，小型餐厅约有 6.66 万家，大型餐厅约有 1.33 万家

由此，北京市共有餐馆 6.66+1.33≈8 万家

角度二（供给层面）

北京市餐馆数量=北京市餐饮业从业人员÷餐馆平均员工数

北京市餐营业从业人员共有约 26 万人。假设一家小型餐馆有员工 4 人，一家中大型餐馆有员工 20 人，小型餐厅与中大型餐厅的比例为 5：1。那么，经过计算，小型餐厅有 3.3 万家，中大型餐厅共有 0.65 万家。

由此，北京市共有餐馆 3.3+0.65≈4 万家

结果分析：根据以上两个角度的估算，北京市共有餐馆约 4~8 万家。仍有一些因素可能导致误差，如外卖行业的发展使得许多无店面餐馆迅速发展，从而影响餐厅平均员工数、餐厅容纳人数等数据，可将外卖单独列出进行讨论

3. **统计教授多门课老师数量并输出每位老师教授课程数统计表。[SQL]**

（设 class 表中字段为 id,teacher,class）

解析：

```
Select count(*)

From class

Group by teacher
```

```
Having count(*) >1;

Select teacher , count(*) as NumClass
From class
Group by teacher;
```

4~5 题 [SQL]
用户登录日志表为 user_id, log_id, session_id, plat, visit_date
4. 用 sql 查询近 30 天每天平均登录用户数量
5. 用 sql 查询出近 30 天连续访问 7 天以上的用户数量

解析：

1）近三十天每天平均登录用户数量

```
select visit_date, count(distince user_id)
group by visit_date
```

2）近 30 天连续访问 7 天以上的用户数量

```
select t1.date
from table t1, table t2, ..., table t7
on t1.visit_date = (t2.visit_date+1) and t2.visit_date = (t3.visit_date+1)
and ... and t6.visit_date = (t7.visit_date+1)
```

6. 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。[算法]

解析：聚类分析是一种无监督的学习方法，根据一定条件将相对同质的样本归到一个类总。
聚类方法主要有：
a. 层次聚类
b. 划分聚类：kmeans
c. 密度聚类
d. 网格聚类
e. 模型聚类：高斯混合模型
k-means 比较好介绍，选 k 个点开始作为聚类中心，然后剩下的点根据距离划分到类中；找到新的类中心；重新分配点；迭代直到达到收敛条件或者迭代次数。 优点是快；缺点是要先指定 k，同时对异常值很敏感

7. 当不知道数据所带标签时，可以使用哪种技术促使带同类标签的数据与带其他标签的数据相分离？（ ）[算法]
A. 分类
B. 聚类
C. 关联分析
D. 隐马尔可夫链
答案：B

解析：聚类为非监督学习。**是否有监督（supervised），就看输入数据是否有标签（label）。输入数据有标签，则为有监督学习，没标签则为无监督学习。**

8. 某超市研究销售纪录数据后发现，买啤酒的人很大概率也会购买尿布，这种属于数据挖掘的哪类问题？（ ）[算法]

A. 关联规则发现
B. 聚类
C. 分类
D. 自然语言处理

答案：A

解析：**关联规则分析主要用于分析 Market-Based Problems，比如你在商城里买了一本书，然后商城会推荐你买另一本书（这本书是通过推算买了上本书的人又会买的一本）。类似的例子有很多。简单来说说求 A 发生同时，B 发生的概率。**

9. 决策树中不包含一下哪种结点（ ）[算法]

A,根结点（root node）
B,内部结点（internal node）
C,外部结点（external node）
D,叶结点（leaf node）

答案：C

解析：

在决策树构造过程中，选择什么属性作为根节点、子节点，什么时候停止到达叶节点是三个根本问题，也是组成决策树的三个主要元素；而“外部结点”是“二叉树”模型中的概念。关于决策树的具体内容欢迎大家回顾群内 Day3 的机器学习算法模型讲解哦~

10~11 题 以下代码是否报错 [Python]

10.

```
list= [ ‘a’ ,’ e’ ,’ i’ ,’ o’ ,’ u’ ]
    print list [8:]
```

解析：

输出为[]。 访问一个列表的以超出列表成员数作为开始索引的切片将不会导致 IndexError，并且将仅仅返回一个空列表。

11.

```
def foo (i= []):
    i.append (1)
    return i
foo ()
foo ()
```

解析：

输出为 [1],[1,1]

新的默认列表仅仅只在函数被定义时创建一次。当 foo 没有被指定的列表参数调用的时候，其使用的是同一个列表。

12. 简述 hadoop 和 mapreduce 原理，以及常见的应用场景 [Python]

Hadoop 原理：采用 HDFS 分布式存储文件，MapReduce 分解计算，其它先略

应用场景：

底层：存储层，文件系统 HDFS，NoSQL Hbase

中间层：资源及数据管理层，YARN 以及 Sentry 等

上层：MapReduce、Impala、Spark 等计算引擎

顶层：基于 MapReduce、Spark 等计算引擎的高级封装及工具，如 Hive、Pig、Mahout。

MapReduce 原理：

a. map 阶段：读取 HDFS 中的文件，解析成<k,v>的形式，并对<k,v>进行分区（默认一个区），将相同 k 的 value 放在一个集合中

b. reduce 阶段：将 map 的输出 copy 到不同的 reduce 节点上，节点对 map 的输出进行合并、排序

应用场景：

MR 是应大数据的背景产生，其解决的问题的共性为：大问题可以被分解为许多子问题，且这些子问题相对独立，将这些子问题并行处理完后，大问题也就被解决。是用来分治、分解的思想。

URL 访问率统计，分布式 grep，分布式排序，倒序索引构建，Web 连接图反转等

13. 请创建一个函数检查一个词是否具有回文结构。[Python]

解析：

使用 Python 进行编写

```
def huiwen(str):
if len(str)==1:
return True
```

```
else:
return str[0]==str[-1] and huiwen(str[1:-1])
```

14. **已知中国人的血型分布约为 A 型：30%，B 型：20%，O 型：40%，AB 型：10%，则任选一批中国人作为用户调研对象，希望他们中至少有一个是 B 型血的可能性不低于 90%，那么最少需要选多少人？[概统]**

A. 7
B. 9
C. 11
D. 13

答案：C
解析：
一个人不是 B 型的概率是 1- 0.2 = 0.8
n 个人全不是 b 型的概率是 0.8^n，所以 n 个人至少有一个是 b 型的概率是 1-0.8^n
要这个概率不低于 0.9，就需要 0.8^n < 0.1
n 的最小值是 11

15．简要介绍最大似然估计［概统］

解析：最大似然估计是利用已知的样本结果，反推最有可能（最大概率）导致这样结果的参数值。

其原理可以简要概括为：极大似然估计提供了一种给定观察数据来评估模型参数的方法，即：“模型已定，参数未知”。通过若干次试验，观察其结果，利用试验结果得到某个参数值能够使样本出现的概率为最大，则称为极大似然估计。

如果想要了解更多可以戳下面这个链接：

https://blog.csdn.net/zengxiantao1994/article/details/72787849