DC 共享平台——数据分析岗笔试面试题

本材料是由 DataCastle 从谷歌、微软、facebook、百度等企业的网络公开招聘题中进行精选，并附上详细解析，适合应聘数据分析岗位的求职者，

1.一般，K-NN 最近邻方法在( )的情况下效果较好

A.样本较多但典型性不好

B.样本较少但典型性好

C.样本呈团状分布

D.样本呈链状分布

答案：B

解析：

样本呈团状颇有迷惑性，这里应该指的是整个样本都是呈团状分布，这样 kNN 就发挥不出其求近邻的优势了，整体样本应该具有典型性好，样本较少，比较适宜。

2.一个包里有 5 个黑球，10 个红球和 17 个白球。每次可以从中取两个球出来，放置在外面。那么至少取_____次以后，一定出现过取出一对颜色一样的球。

A.16

B.9

C.4

D.1

答案：A

解析：

考虑最坏的情况，前 10 次取出的都是红球+白球的组合，后 5 次取出的都是黑球+白球的组合，最后只剩下两个白球，则再取 1 次必取出相同颜色的球,因此总计 16 次。

3.用直接插入排序方法对下面 4 个序列进行排序(由小到大),元素比较次数最少的是( )

A.94,32,40,90,80,46,21,69

B.32,40,21,46,69,94,90,80

C.21,32,46,40,80,69,90,94

D.90,69,80,46,21,32,94,40

答案：C

解析：

插入排序的原理是将第 i 个数插入到已经排列好的数据中 ,因此原序列越有序，比较次数越少

4.下面有关分类算法的准确率，召回率，F1 值的描述，错误的是？

A.准确率是检索出相关文档数与检索出的文档总数的比率，衡量的是检索系统的查准率

B.召回率是指检索出的相关文档数和文档库中所有的相关文档数的比率，衡量的是检索系统的查全率

C.正确率、召回率和 F 值取值都在 0 和 1 之间，数值越接近 0，查准率或查全率就越高

D.为了解决准确率和召回率冲突问题，引入了 F1 分数

答案：C

解析：

对于二类分类问题常用的评价指标是精准度（precision）与召回率（recall）。通常以关注的类为正类，其他类为负类，分类器在测试数据集上的预测或正确或不正确，4 种情况出现的总数分别记作：

TP——将正类预测为正类数

FN——将正类预测为负类数

FP——将负类预测为正类数

TN——将负类预测为负类数

由此：

精准率定义为：P = TP / (TP + FP)

召回率定义为：R = TP / (TP + FN)

F1 值定义为： F1 = 2 P R / (P + R)

精准率和召回率和 F1 取值都在 0 和 1 之间，精准率和召回率高，F1 值也会高，不存在数值越接近 0 越高的说法，应该是数值越接近 1 越高。

5.Naive Bayes 是一种特殊的 Bayes 分类器,特征变量是 X,类别标签是 C,它的一个假定是:()

A.各类别的先验概率 P(C)是相等的

B.以 0 为均值，$\sqrt{\frac{2}{2}}$ 为标准差的正态分布

C.特征变量之间是相互独立的

D.P(X|C)是高斯分布

答案：C

解析：

朴素贝叶斯的条件就在于假设每个变量相互独立

6.下列不是 SVM 核函数的是：

A.多项式核函数

B.logistic 核函数

C.径向基核函数

D.Sigmoid 核函数

答案：B

解析：

SVM 核函数包括线性核函数、多项式核函数、径向基核函数、高斯核函数、幂指数核函数、拉普拉斯核函数、ANOVA 核函数、二次有理核函数、多元二次核函数、逆多元二次核函数以及 Sigmoid 核函数

7.(多选)数据清理中，处理缺失值的方法是?

A.估算

B.整例删除

C.变量删除

D.成对删除

答案：A,B,C,D

解析：

数据清理中，处理缺失值的方法有两种：

删除法：

1）删除观察样本

2）删除变量：当某个变量缺失值较多且对研究目标影响不大时，可以将整个变量整体删除

3）使用完整原始数据分析：当数据存在较多缺失而其原始数据完整时，可以使用原始数据替代现有数据进行分析

4）改变权重：当删除缺失数据会改变数据结构时，通过对完整数据按照不同的权重进行加权，可以降低删除缺失数据带来的偏差

查补法：均值插补、回归插补、抽样填补等

成对删除与改变权重为一类

估算与查补法为一类

8.在 Logistic Regression 中,如果同时加入 L1 和 L2 范数,会产生什么效果()

A.可以做特征选择,并在一定程度上防止过拟合

B.能解决维度灾难问题

C.能加快计算速度

D.可以获得更准确的结果

答案：A

解析：

L 1 范数具有系数解的特性，但是要注意的是，L 1 没有选到的特征不代表不重要，原因是两个高相关性的特征可能只保留一个。需要通过交叉验证，确定哪个特征重要。

为什么 L1，L2 范数可以防止过拟合呢？

在代价函数后面加上正则项，L 1 即是 Lasso 回归，L 2 是岭回归

但是它为什么能防止过拟合呢？

奥卡姆剃刀原理：能很好的拟合数据且模型简单

模型参数在更新时，正则项可使参数的绝对值趋于 0 ，使得部分参数为 0 ，降低了模型的复杂度（模型的复杂度由参数决定），从而防止了过拟合。提高模型的泛化能力。

9.有两个样本点，第一个点为正样本,它的特征向量是(0,-1);第二个点为负样本,它的特征向量是(2,3),从这两个样本点组成的训练集构建一个线性 SVM 分类器的分类面方程是( )

A.2x+y=4

B.x+2y=5

C.x+2y=3

D.2x-y=0

答案：B

解析：

SVM 要找到间隔最大的分类平面，这里即求两点(0,-1),(2,3)的垂直平分线。

斜率为：-1/((3+1)/(2-0))=-1/2

中点为：(1,2)

所以，分类超平面为：x+2y=5

10.执行完下列语句段后,i 值为()

```
int f(int x){
    return ((x>0)?x*f(x-1):2)
}
int i;
i=f(f(2));
```

A.4

B.48

C.8

D.无限递归

答案：B

解析：

f(x) 当 x 大于 0 时，返回 x*f(x-1)，否则返回 2

f(0) = 2

f(1) = 1 * f(0) = 2

f(2) = 2 * f(1) = 4

f(3) = 3 * f(2) = 12

f(4) = 4 * f(3) = 48

f(4) = f(f(2)) = 48

11.连续存储设计时,存储单元的地址()

A.一定连续

B.一定不连续

C.不一定连续

D.部分连续,部分不连续

答案：C

解析：

1.存储单元的地址，考察的是存储结构：

2.存储结构的含义是：数据元素在计算中的存储形式。

3.线性表的存储结构分为顺序存储和链式存储：（1）顺序存储为逻辑上相邻且物理地址也连续，以数组形式出现，可以取任意下标访问，是一种随机存取的存储结构；（2）链式存储是逻辑上相邻但是物理地址不一定连续，以链表的形式出现，必须从头开始访问，是一种顺序存取的存储结构。

所以：答案选 C,存储单元的地址（物理地址）不一定连续

12.麦秋时节，庄园主雇了个力大无穷的农民来帮他收割田里的麦子。收获的劳动量很大，农民必须在七天之内收割完。庄园主答应每天给他一块金块作工钱。但是这七块相等的金子是连在一起的，然而工钱是必须每天结清的。农民不愿意庄园主欠帐，而庄园主也不肯预付一天工钱。请问最少掰金子几次可以完成上述任务？

A.2

B.3

C.4

D.7

答案：A

解析：

1、第一天，庄园主掰 1 块金给农民，付第一天的。农民：1；庄园主：6

2、第二天，庄园主从剩下的 6 块里面掰下 2 块给农民，并收回第一天的一块。农民：2；庄园主：1+4

3、第三天，庄园主将手里的一块散金给农民。农民：2+1；庄园主：4

4、第四天，庄园主收回农民手里的三块金，并把手里的金给农民。农民：4；庄园主：2+1

5、第五天，庄园主把手里的 1 块散金给农民。农民：4+1；庄园主：2

6、第六天，庄园主收回农民手里的 1 块散金，将 2 块金给农民。农民：4+2，庄园主：1

7、第七天，庄园主将手里的金块给农民。农民：7，庄园主：0

综上，最少掰了两次可以搞定

13.(多选)算法一般都可以用哪几种控制结构组合而成？

A.顺序

B.选择

C.递归

D.循环

答案：A,B,D

解析：

算法一般不用递归，因为太消耗时间。

14.用下面的 T-SQL 语句建立一个基本表：

CREATE TABLE Student (Sno CHAR (4) PRIMARY KEY,

Sname CHAR (8) NOT NULL,

Sex CHAR (2) ,

Age INT)

可以插入到表中的元组是（ ）

'A.5021'，'刘祥'，男，21

B.NULL，'刘祥'，NULL，21

'C.5021 '，NULL，男，21

'D.5021'，'刘祥'，NULL，NULL

答案：D

解析：

“男”为字符串类型，要添加引号，所以 A,C 排除，主键不能为 NULL，排除 B

15.一个查询语句执行后显示的结果为：

1 班 80

2 班 75

3 班 NULL

，则最有可能的查询语句是（ ）

A.SELECT AVG(成绩) FROM 成绩表 WHERE class<=3

B.SELECT AVG(成绩) FROM 成绩表 WHERE class<=3 GROUP BY class

C.SELECT AVG(成绩) FROM 成绩表 WHERE class<=3 order by class

D.SELECT AVG(成绩) FROM 成绩表 HAVING class <=3 GROUP BY class

答案：B

解析：

select 之后如果是聚合函数则 group by 分组会显示 null 的结果,而 order by 不会显示 null 的结果，having 用在分组之后

16.在一个单链表中，q 的前一个节点为 p，删除 q 所指向节点，则执行

A.delete q

B.q->next=p->nerx;delete p;

C.p-next=q->next;delete p;

D.p->next=q->next;delete q;

E.delete p;

F.q->next=p->next;delete q

答案：D

解析：

让 p 指向 q 的下一个节点再删除 q

17.（多选）有一个单向链表，头指针和尾指针分别为 p，q，以下哪项操作的复杂度不受队列长度的影响？

A.删除头部元素

B.删除尾部元素

C.头部元素之前插入一个元素

D.尾部元素之后插入一个元素

答案：A,C,D

解析：

单链表删除元素需要找到尾部元素的前一个元素，与队列长度有关，因此删除尾部元素时，虽然给出了尾指针，但是单链表删除还要知道前一节点,所以还是要遍历一遍才能知道尾指针前一节点 既与队列长度有关

18.设有表示学生选课的三张表，学生 S（学号，姓名，性别，年龄，身份证号），课程 C（课号，课名），选课 SC（学号，课号，成绩），则表 SC 的关键字（键或码）为（　　）。

A.课号，成绩

B.学号，成绩

C.学号，课号

D.学号，姓名，成绩

答案：C

解析：

学号是学生表 S 的主键，课号是课程表 C 的主键，所以选课表 SC 的关键字就应该是与前两个表能够直接联系且能唯一定义的学号和课号，所以选择 C。

19.S 市 A，B 共有两个区，人口比例为 3：5，据历史统计 A 的犯罪率为 0.01%，B 区为 0.015% ,现有一起新案件发生在 S 市 ,那么案件发生在 A 区的可能性有多大？

A.37.5%

B.32.5%

C.28.6%

D.26.1%

答案：C

解析：

在 A 区犯案概率：P(C|A)=0.01%

在 B 区犯案概率：P(C|B)=0.015%

在 A 区概率：P(A)=3/8

在 B 区概率：P(B)=5/8

犯案概率：P(C)=（3/8*0.01%+5/8*0.015%)

则犯案且在 A 区的概率：P(A|C)=P(C|A)*P(A)/P(C)=0.01%*(3/8)/（3/8*0.01%+5/8*0.015%)≈28.6%

20.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?

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

21.若元素 a,b,c,d,e,f 依次进栈，允许进栈、退栈操作交替进行。但不允许连续三次进行退栈操作，则不可能得到的出栈序列是（ ）

A.d，c，e，b，f，a

B.c，b，d，a，e，f

C.b，c，a，e，f，d

D.a，f，e，d，c，b

答案：D

解析：

最后入栈的元素越早出栈，就越可能违背题目中不得连续出栈的要求，因为后入栈的元素之前的元素大部分都只能跟在该元素之后依次弹出。

如 D 中最后入栈的 f 第二个出栈，那么其前的 bcde 就只能连续出栈了。

22.单链表中,增加一个头结点的目的是为了()

A.使单链表至少有一个结点

B.标识表结点中首结点的位置

C.方便运算的实现

D.说明单链表是线性表的链式存储

答案：C

解析：

头结点作用: (1) 对带头结点的链表，在表的任何结点之前插入结点或删除表中任何结点 ,所要做的都是修改前一结点的指针域 ,因为任何元素结点都有前驱结点。若链表没有头结点，则首元素结点没有前驱结点，在其前插入结点或删除该结点时操作会复杂些。 (2) 对带头结点的链表，表头指针是指向头结点的非空指针，因此空表与非空表的处理是一样的。

23.已知二叉树后序遍历序列是 bfegcda，中序遍历序列是 badefcg，它的前序遍历序列是：

A.abcdefg

B.abdcefg

C.adbcfeg

D.abecdfg

答案：B

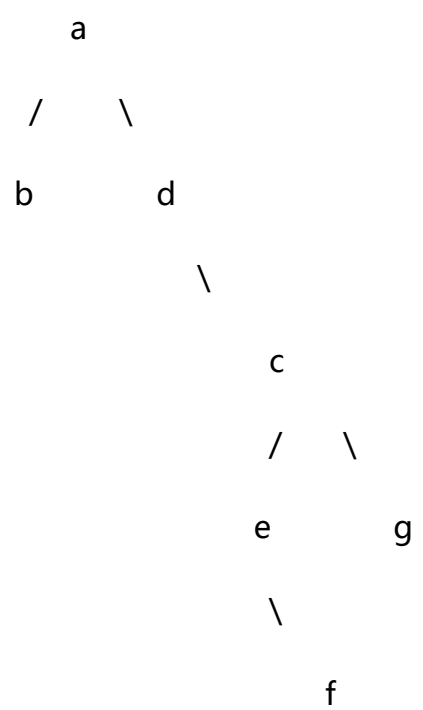
解析：

1.前根序遍历：先遍历根结点，然后遍历左子树，最后遍历右子树。

2.中根序遍历：先遍历左子树，然后遍历根结点，最后遍历右子树。

3.后根序遍历：先遍历左子树，然后遍历右子树，最后遍历根节点。

最后树的形状如下：



24.请创建一个函数检查一个词是否具有回文结构。

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])
```

25.在应用机器学习算法之前纠正和清理数据的步骤是什么？

解析：

- 将数据导入
- 看数据：重点看元数据，即对字段解释、数据来源等信息；导入数据后，提取部分数据进行查看
- 缺失值清洗
  - 根据需要对缺失值进行处理，可以删除数据或填充数据
  - 重新取数：如果某些非常重要的字段缺失，需要和负责采集数据的人沟通，是否可以再获得
- 数据格式清洗：统一数据的时间、日期、全半角等显示格式

- 逻辑错误的数据
  - 重复的数据
  - 不合理的值

- 不一致错误的处理：指对矛盾内容的修正，最常见的如身份证号和出生年月日不对应

不同业务中数据清洗的任务略有不同，比如数据有不同来源的话，数据格式清洗和不一致错误的处理就尤为突出。数据预处理是数据类岗位工作内容中重要的部分。

26.在 K-Means 中如何拾取 k？

解析：

K-Means 算法的最大缺点是不能自动选择分类数 k，常见的确定 k 的方法有：

- 根据先验知识来确定
- $k=\sqrt{\frac{N}{2}}$，N 为样本数
- 拐点法：把聚类结果的 F-test 值对聚类个数的曲线画出来，选择图中的拐点
- 基于信息准则判断，如果模型有似然函数，则可以用 BIC、DIC 来进行决策

具体的 k 的选择往往和业务联系紧密，如希望能将用户进行分类，就有先验的分类要求

27.请定义一下方差。

解析：数据与平均数之差的平方和的平均数，用于衡量随机变量或一组数据离散程度的度量

28.如何对数值预测模型进行评估？

解析：

不同模型作用于同一个数据集的结果是不同的，往往需要对模型性能做出评估，来进行选择，数值预测模型的评估，有以下的常用评估指标：

- 均方误：是最常见的指标，但是容易受到奇异值的影响
- 平均绝对误差
- 方均根差
- 借助图形分析工具

29.什么是中心极限定理（Central Limit Theorem），它的应用方向是什么？

解析：

中心极限定理：讨论随机变量序列部分和分布渐近于正态分布的一类定理，这组定理指出了大量随机变量累积分布函数逐点收敛到正态分布的积累分布函数的条

件。

应用方向：中心极限定理规定，在一些互相独立的随机因素的影响都很微小的情况下，总影响可以看做是服从正态分布的，它是数理统计和误差分析的基础，在自然界与生产中，有着广泛的应用。

30.请解释过拟合，以及如何防止过度拟合。

解析：

过拟合：是指为了得到一致假设而使假设变得过度严格

判断过拟合的方法：一个假设（模型）在训练数据上能够获得比其他假设（模型）更好的拟合， 但是在【训练数据外】的数据集上却不能很好地拟合数据，这就意味着出现了过拟合现象。

解决办法通常有：

- 增大数据量
- 适当放宽阈值
- 交叉验证
- 减少特征
- 正则化特征
- 减少权值

31.请尝试向非技术人员阐释交叉验证（Cross-validation）。

解析：

将数据样本切割成较小的子集，一部分用于训练模型，一部分用于验证模型（训练集的规模比验证集的规模大得多），利用验证集来测试训练得到的模型,主要用于评估模型的性能。通过模型在训练集上的表现和在验证集上的表现差异，来评估模型的泛化能力，和最终确定模型。

常见的有：k-folds 交叉验证，leave-one-out 法。

k-folds: 将初始采样分割成 K 个子样本，一个单独的子样本被保留作为验证模型的数据，其他 K-1 个样本用来训练。交叉验证重复 K 次，每个子样本验证一次，平均 K 次的结果或者使用其它结合方式，最终得到一个单一估测。

leave-one-out 法：只使用原本样本中的一项来当做验证资料， 而剩余的则留下来当做训练资料。 这个步骤一直持续到每个样本都被当做一次验证数据。

交叉验证例子：如要预测学生的成绩，可以将学生数据集分成十份，轮流将其中 9 份做训练 1 份做验证，10 次的结果的均值作为对算法精度的估计，一般还需要进行多次 10 折交叉验证求均值。

32.请解释异方差（heteroskedasticity）是什么，以及如何解决它。

解析：

异方差的概念要从经典线性回归模型的假设开始说起，根据同方差假设，每一个干扰项的方差相同，而异方差则为相反的，常见解决异方差的方法有：

- 对模型变换：适用于异方差的具体形式可以确定的情况，将模型作适当变换有可能消除或减轻异方差的影响

- 加权最小二乘法：隶属于对模型变换方法中的一种，将 lm 模型换成 gls 模型，可以消除异方差

- 对数变换模型：运用对数变换可以使得测定变量值的尺度缩小，可以一定程度上解决异方差问题

请阅读异方差案例，了解是怎样通过模型的变化来修正异方差的。

33.请解释偏差和方差权衡

解析：

首先需要明确的是偏差和方差的概念，对于同一总体，有若干个数据子集，通过训练这些子集可以得到若干的模型，将模型的期望（或平均）预测和我们真实值之间的差定义为偏差。将模型之间的多个拟合预测之间的偏离程度定义为方差。偏差和方差的权衡在于如果一味的追求模型的精确匹配，可以使得偏差降低，但可能会导致不同子集数据训练出的不同模型之间的差异非常大，方差过大，模型的泛化能力较差，容易出现过拟合现象。所以一般来讲偏差和方差是不能兼顾的。

34.65,8,50,15,37,24,( )。括号中的数字是：

A.25

B.26

C.22

D.27

答案：B

解析：

奇数项是按照 15,13,11 的递减差递减数列；偶数项是按照 7,9,11,的递增差增序列。那么答案就是第七项，奇数列就按照减 11 吧，37-11=26

35.在以下不同的场景中,使用的分析方法不正确的有：

A.根据商家最近一年的经营及服务数据,用聚类算法判断出天猫商家在各自主营类目下所属的商家层级

B.根据商家近几年的成交数据,用聚类算法拟合出用户未来一个月可能的消费金额公式

C.用关联规则算法分析出购买了汽车坐垫的买家,是否适合推荐汽车脚垫

D.根据用户最近购买的商品信息,用决策树算法识别出淘宝买家可能是男还是女

答案：B

解析：

预测消费需要用回归模型来做。而不是聚类算法。

36.下面关于 ID3 算法中说法错误的是（ ）

A.ID3 算法要求特征必须离散化

B.信息增益可以用熵，而不是 GINI 系数来计算

C.选取信息增益最大的特征，作为树的根节点

D.ID3 算法是一个二叉树模型

答案：D

解析：

ID3 算法（Iterative Dichotomiser 3 迭代二叉树 3 代）是一个由 Ross Quinlan 发明的用于决策树的算法。可以归纳为以下几点：

1. 使用所有没有使用的属性并计算与之相关的样本熵值
2. 选取其中熵值最小的属性
3. 生成包含该属性的节点

ID3 算法对数据的要求：

1) 所有属性必须为离散量；
2) 所有的训练例的所有属性必须有一个明确的值；
3) 相同的因素必须得到相同的结论且训练例必须唯一。

37.（多选）某电商推出一款新的产品，希望这个产品能大卖，让你给这个主题取个名字，如果你是数据分析师，以下哪些指标可以用来判断。

A.成交总量： 代表产品销售的收入

B.独立用户数： 代表购买产品的用户，说明产品的覆盖面

C.评价数（好评数）： 反馈用户对产品口碑

D.购买时间：代表产品的销售与时间的相关性

答案：A,B,C

解析：

本题目的关键点在于【产品大卖】，所以需要找出相关的指标。

商品大卖无非包含：潜在市场、意向用户、销售收入、利润等。

简单说，衡量商品是不是大卖的指标都是【分好坏】的指标。

像购买时间，并没有好坏之分，只是周期变动的一般趋势，衡量不出是不是大卖。

38.在满足实体完整性约束的条件下（ ）。

A.一个关系中应该有一个或多个候选关键字

B.一个关系中只能有一个候选关键字

C.一个关系中必须有多关键字个候选

D.一个关系中可以没有候选关键字

答案：A

解析：

实体完整性约束是指一个关系具有某种唯一性标识，其中主关键字为唯一标识，而主关键字中的属性不能为空。所以候选关键字可以有一个或者多个，答案选择 A。

数据完整性分为以下四类：

1) 实体完整性：规定表的每一行在表中是惟一的实体。

2) 域完整性：是指表中的列必须满足某种特定的数据类型约束，其中约束又

包括取值范围、精度等规定。

3) 参照完整性：是指两个表的主关键字和外关键字的数据应一致，保证了表之间的数据的一致性，防止了数据丢失或无意义的数据在数据库中扩散。

4) 用户定义的完整性：不同的关系数据库系统根据其应用环境的不同，往往还需要一些特殊的约束条件。用户定义的完整性即是针对某个特定关系数据库的约束条件，它反映某一具体应用必须满足的语义要求。

39.在其他条件不变的前提下，以下哪种做法容易引起机器学习中的过拟合问题（ ）

A.增加训练集量

B.减少神经网络隐藏层节点数

C.删除稀疏的特征

D.SVM 算法中使用高斯核/RBF 核代替线性核

答案：D

解析：

一般认为，增加隐层数可以降低网络误差（也有文献认为不一定能有效降低），提高精度，但也使网络复杂化，从而增加了网络的训练时间和出现“过拟合”的倾向， svm 高斯核函数比线性核函数模型更复杂，容易过拟合

40.有个苦逼的上班族，他每天忘记定闹钟的概率为 0.2，上班堵车的概率为 0.5，如果他既没定闹钟上班又堵车那他迟到的概率为 1.0，如果他定了闹钟但是上班堵车那他迟到的概率为 0.8，如果他没定闹钟但是上班不堵车他迟到的概率为 0.9，如

果他既定了闹钟上班又不堵车那他迟到的概率为 0.0，那么求出他在 60 天里上班迟到的期望。

A.30.6

B.40.1

C.25.8

D.36.8

答案：A

解析：

每天迟到的概率 P=1*0.2*0.5+0.9*0.2*0.5+0.8*0.8*0.5+0=0.51

60 天里上班迟到的期望为 :E(x1+x2+…+x60)=E(x1)+…+E(x60)=60*0.51=30.6

41.人患癌症的概率为 1/1000。假设有一台癌症诊断仪 S1，通过对它以往的诊断记录的分析，如果患者确实患有癌症它的确诊率为 90%,如果患者没有癌症，被诊断成癌症的概率是 10%。某人在被诊断为癌症后，他真正患癌症的概率为()

A.9/1000

B.1/1000

C.1/112

D.9/10

答案：C

解析：

分为真的有癌症真的检查出来了:1/1000*9/10

假的有癌症但是检查错误了:999/1000*1/10

所以概率为:(1/1000*9/10)/((1/1000*9/10)+(999/1000*1/10))=/1/112

42.有三个黑气球，其中只有一个黑气球中有金币，你可以任意选择任何一个气球，而主持人在剩下的气球中打破一个气球，然后告诉你里边没有金币:你还有一次机会，既可以坚持选择，也可以换另外一个未打破的气球。如果你选择换的话获得金币的概率为()

A.1/3

B.1/2

C.2/3

D.0

答案：C

解析：

如果你第一次选择有金币的气球（1/3 的概率），那么你换了之后肯定得不到金币，所以这种情况下得到金币的概率是 1/3*0=0。如果你第一次选择没有金币的气球（2/3 的概率），那么你换了之后，剩下的那个没有破的气球里面就是金币，所以这种情况下得到金币的概率是 2/3*1=2/3。总概率 0+2/3=2/3。

43.小易有一个长度为 n 的整数序列,a_1,...,a_n。然后考虑在一个空序列 b 上进行 n

次以下操作:

1、将 a_i 放入 b 序列的末尾

2、逆置 b 序列

小易需要你计算输出操作 n 次之后的 b 序列。

解析：

别被迷惑了，其实不需要逆序，直接从后向前间隔一个输出，然后没有输出的顺序输出即可

```
def findNum(nums, n):
    for i in range(n-1, -1, -2):
        print (nums[i])
    if n&1 == 0:
        for i in range(0, n, 2):
            print (nums[i])
    else:
        for i in range(1, n ,2):
            print (nums[i])
```

44.小易为了向他的父母表现他已经长大独立了,他决定搬出去自己居住一段时间。

一个人生活增加了许多花费: 小易每天必须吃一个水果并且需要每天支付 x 元的房屋租金。当前小易手中已经有 f 个水果和 d 元钱,小易也能去商店购买一些水果,商店每个水果售卖 p 元。小易为了表现他独立生活的能力,希望能独立生活的时间越长越好,小易希望你来帮他计算一下他最多能独立生活多少天。

解析：

输入描述：输入包括一行,四个整数 x, f, d, p(1 ≤ x,f,d,p ≤ 2 * 10^9),以空格分割输出描述：输出一个整数, 表示小易最多能独立生活多少天。

输入例子：3 5 100 10

输出例子：11

代码实现：

```
x,f,d,p = map(int,raw_input().split())
day = (f*p+d)/(x+p)
print min(day,d/x)
```

45.监督学习和无监督学习有什么区别？

解析：

监督学习：对具有标记（分类）的训练样本进行学习，这里，所有的标记（分类）是已知的。如：决策树算法、朴素贝叶斯算法、KNN 算法。

无监督学习：对没有标记（分类）的训练样本进行学习，目的是为了发现训练集中的结构特性。这里，所有的标记（分类）是未知的。如：聚类算法。

46.写 MySQL 语句 :table1 有 url ,pv 两个字段， table2 有 url,title 两个字段，取出 url 相等的 url, title,pv 数据行 ,且筛选出 pv 大于 100 的行 ,按逆序排列。

解析：

SELECT url,title,pv FROM table1,table2 ORDER BY pv DESC

WHERE table1.url=table2.url AND pv>100

47.对 n 个数进行排序，在各自最优条件下，以下算法复杂度最低的是:

A 快速排序

B 堆排序

C 冒泡排序

D 插入排序

E 选择排序

F 归并排序

答案: C,D

解析：

这六个排序算法中，复杂度的排序为：

算法	最优复杂度
快速排序	$O(nlog_2n)$
堆排序	$O(nlog_2n)$
冒泡排序	$O(n)$
插入排序	$O(n)$
选择排序	$O(n^2)$
归并排序	$O(nlog_2n)$

复杂度的排序是：

O(1) 常数阶 < O(logn) 对数阶 < O(n) 线性阶 < O(nlogn) < O(n^2) 平方阶 < O(n^3) < O(2^n) < O(n!) < O(n^n)

值得注意的是题目的关键是“最优条件下”，而不是平均情况下，所以算法的复杂程度排序为：冒泡排序=插入排序<快速排序=堆排序=归并排序<选择排序

排序算法是算法类岗位的基础题目，复杂度的比较是一类题目，而排序算法的知识点远不止这个，你还需要掌握如何实现排序算法、区别、稳定度等等

48.如果次日用户留存率下降了 5%该怎么分析

解析：

留存率的分析在数据分析岗位应聘中是非常重要的一类题目，引入留存率是为了探究一批用户导入质量或推广渠道质量，从而进行后续运营、生产的等策略的调整。

这一类型的题目往往和实际业务结合比较紧密，建议可以从职责角度出发，如应聘的是游戏运营行业，可以从是否属于异常波动入手，如果超出正常波动范围，则可以从近几日的外部事件入手，如果在正常范围内，则应该考虑历史趋势、玩家的生命周期长度等因素

49.怎么投放可以获得最大收益，请提供思路

解析：

可以从以下角度进行思考：受众群体、不同渠道的获客成本、比对行业平均获客成本等等角度

广告投放是数据分析领域一个常见的职位，如果能够更多结合你所应聘的公司谈一些特性化的内容，相信会是一个很大的加分项！

如：以联合利华家乐广告投放 CCTV 移动传媒为例

1)受众群体 :CCTV 移动传媒的受众有接触信息频繁、和其他投放渠道形成互补性、出行群体购买行为活跃等特征，符合品牌的受众群体特征和投放需求

2) 渠道获客成本：CCTV 移动传媒有较高的公众认知度和高品牌价值、提高公众信任感的可能性，获得目标客户的成本较为合理

因此 CCTV 移动传媒对于联合利华家乐而言是一个较为理想的投放广告渠道

50.如何构建一个推荐系统

解析：

推荐算法主要有两类：基于内容的推荐算法和协同过滤推荐算法。

- 基于用户信息和物品内容的推荐算法：这一类的推荐算法主要考虑的特征是用户的基本信息和使用购买物品的历史记录
- 协同过滤算法：通过用户行为来计算出用户或者物品间的相关性，简单理解就是，用和你行为相似的用户的偏好来对你进行推荐

除了了解推荐算法之外，构造一个推荐系统还需要注意一下几个方面

1）获取数据：可能会面临用户冷启动问题，需要考虑如何尽快让用户表达兴趣

2）数据稀疏性：现实中当用户数量和物品数量较大时 ，往往数据矩阵会是一个稀疏矩阵，对于这种情况可以通过降维、对用户打分来进行解决

3）实时性：优秀的推荐系统总是在实时更新的

例如：以豆瓣 FM 的推荐系统为例，想要实现个性化的推荐可以通过以下角度思考：

- 采集用户数据：社交行为、用户听歌习惯（红心、垃圾、跳过）、用户行为数据
- 计算和每首歌近似的歌曲集合：可以通过音乐本身的标签、同好用户的音乐偏好、用户评价来分类
- 聚类用户，给用户打标签：通过以上收集到的用户数据作为特征进行分析聚类