以下哪个算法属于非监督学习算法？
正确答案: A (
聚类分析
逻辑回归
决策树
神经网络
二元分类器效果指标不易受正负样本比例影响的是哪个？
正确答案: D (
查全率
查准率
提升值
AUC
以下关于逻辑回归的说法不正确的是？
正确答案: C (
逻辑回归必须对缺失值做预处理；
逻辑回归要求自变量和目标变量是线性关系；
逻辑回归比决策树，更容易过度拟合；
逻辑回归只能做 2 值分类，不能直接做多值分类；
以下哪个问题不是分类问题？
正确答案: B (
用户流失模型
身高和体重关系
信用评分
营销响应
常见的 sql 优化技巧，不正确的是？
正确答案: D (
建分区表
建索引
避免笛卡尔积
单一过滤条件尽量写在两表关联之后

常用的排序算法中，平均时间复杂度为 O(n*logn)的有哪些？
正确答案: A C D (
堆排序
冒泡排序
快速排序
归并排序
你建立一个回归模型来预测人群收入水平。你在模型中选用了用到了年龄和经验两个独立变量。得到模型结果后，你发现经验的系数是负的，这好像跟直观相反。另外，你发现该系统的 t 统计量很低但是回归模型却具有很高的拟合优度。是什么原因造成这些结果？
正确答案: D (
不正确的标准误差
异方差性

序列相关性
多重共线性

决策树中属性选择的方法有？
正确答案: B C D （
信息值
信息增益
信息增益率
GINI 系数

如下哪些方法是用来提升模型的泛化能力的？
正确答案: A B C D （
ridge
Lasso
ElasticNet
Dropout
下列哪些方法可以用来对高维数据进行降维？
正确答案: A B C D （
因子分析
主成分分析
奇异值分解
线性判别分析

SQL 题目：

假设用户A和用户B关注了同一件产品相似度记为1，共同关注2件商品相似度记为2，依次类推。已知mysql表名为t_user，字段名分别为fuser（用户），fprod（关注产品），数据如下所示：

用户	关注产品
A	a
A	b
A	c
B	b
B	c
B	d
C	a
C	d
C	e
D	a
D	e

1、用sql语句统计所有用户两两间的相似度

你的答案

答案1：

```
SELECT     x.fuser,y.fuser,count(distinct b.fprod) as fsim FROM      (SELECT distinct fuser from
t_user r) x join (SELECT distinct fuser from t_user r) y on x.fuser < y.fuser join t_user a on a.Fuser
= x.fuser left join t_user b on a.fprod = b.fprod and y.fuser = b.fuser and a.fuser <> b.fuser
group by x.fuser,y.fuser
```

答案2：

```
SELECT l.fuser1,l.fuser2,IFNULL(r.result,0) as result
FROM
(
SELECT  a.fuser AS fuser1,b.fuser AS fuser2
FROM (SELECT DISTINCT fuser FROM test.t_user t) a  JOIN (SELECT DISTINCT fuser FROM
test.t_user t) b
ON a.fuser < b.fuser
) l
LEFT JOIN
(
SELECT  c.fuser  AS fuser1,d.fuser AS fuser2,COUNT(1) AS result
FROM test.t_user c  JOIN test.t_user d
ON c.fuser <  d.fuser AND c.fprod = d.fprod
GROUP BY c.fuser,d.fuser
) r
ON l.fuser1 = r.fuser1 AND l.fuser2 = r.fuser2
```