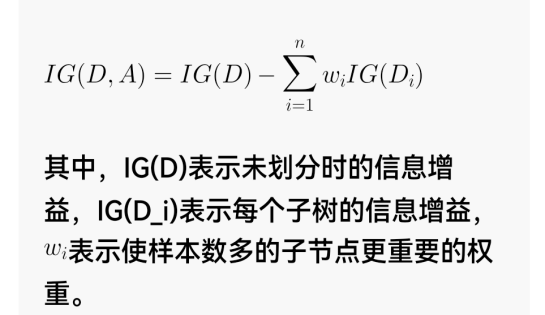
**计算**信息增益

**信息增益可以用划分前后集合熵的差值来衡量，熵可以表示样本集合的不确定性，熵越大，样本的不确定性就越大。**

**这应该是道计算题**

****

何谓数据仓库？

数据仓库是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合，用于支持管理决策的制定。存储在数据仓库中的数据通常具有一致性、完整性、稳定性等特点，这些数据来源于多个异构的数据源，经过抽取、清洗、转换、装载等一系列操作。

ID3算法的基本思想？

是一种贪心+DFS算法，先找出最有判别力的因素，然后把数据分成多个子集，每个子集选择最有判别力的因素递归划分，一直进行到所有子集仅包含同一类型的数据为止。最后得到一棵决策树，可以用它来对新的样例进行分类。

在线联机分析处理？

是用于数据分析和决策支持的技术。它允许用户快速、灵活地对大量数据进行多维分析和查询。  
 OLAP 系统通常基于多维数据模型构建，将数据组织成多个维度和度量值。用户可以通过旋转、切片、钻取等操作，从不同的角度和层次对数据进行分析和查询，以获取所需的信息。

数据规范化？

按照一定的规则和标准对数据进行整理和格式化，以便数据的存储、管理、分析、使用。

数据归约？

在保持数据完整性和可用性的前提下，通过减少数据的数量和维度，提高数据处理和分析的效率和性能的过程。

聚类

将数据集划分为若干个不同的子集，使得每个子集内具有相似的特征或属性，不同子集之间的数据具有较大差异。

数据仓库的最终目的是为用户和业务部门提供决策支持。对

大数据的分析理念是指在数据规模上强调相对数据而不是绝对数据 错

在数据生命周期管理实践中，执行方法是指数据的管理和维护。 对

在抽取样本时，总体中每个对象被抽中为调查样本的概率可能会有差异。 错

分类规则的质量可用规则的覆盖率和 准确率 表示。

规则集的两个重要性质是互斥规则和 穷举原则 。

在训练有监督的学习的机器学习模型的时候，会将数据划分为 训练集、测试集、验证集

回归分析 是确定两种或两种以上变量间相互依赖关系的一种统计分析方法

常用的降维算法有 主成分分析（PCA）、因子分析和独立成分分析。

聚类算法根据产生簇的机制不同，主要分成 划分聚类 、层次聚类和密度聚类。

OLAP的中文意思是指 联机分析处理

分类规则的质量可用规则的覆盖率和\_\_\_\_\_示。

规则集的两个重要性质是互斥规则和 \_\_\_\_\_。

在训练有监督的学习的机器学习模型的时候，会将数据划分为\_\_\_\_\_

\_\_\_\_\_是确定两种或两种以上变量间相互依赖关系的一种统计分析方法

常用的降维算法有\_\_\_\_\_、因子分析和独立成分分析。

聚类算法根据产生簇的机制不同，主要分成\_\_\_\_\_ 、层次聚类和密度聚类。

OLAP的中文意思是指\_\_\_\_\_。

何谓数据仓库？

ID3算法的基本思想？

在线联机分析处理？

数据规范化？

数据归约？

聚类