机器学习第一次作业

1. 机器学习的基本问题有哪些？这些问题的基本含义是什么？请给每个问题举一个应用实例。

**机器学习的基本问题有：**

**1.分类：将数据划分到指定的类别中。如垃圾邮件的识别分类；**

**2.回归：预测数值型目标变量的值。如利用房屋的特征进行房价预测，疫情时期预测某地某日感染人数。**

**3.聚类：将数据样本分成不同的组别。如社交媒体用户的分类。**

**4.联想：关联规则学习，找到数据中项目之间频繁出现的模式或者关联规则。如网络流量分析。**

**5.优化：给定一组约束条件下，寻找能够最大化或最小化某个指标的解。如例如通过优化线性回归模型的参数，建造一个线性回归模型来预测房屋价格。**

1. 请简述完成一个机器学习任务的基本过程。深度学习的过程与传统机器学习过程有何区别？

**机器学习的基本过程：**

**1.收集数据**

**2.清洗数据**

**3.训练模型**

**4.获得知识**

**5.应用和检验知识**

**深度学习与传统机器学习的过程主要有以下区别：**

**1.模型组成与复杂性：传统机器学习算法通常基于简单的模型，如线性模型，决策树模型，而深度学习模型则通常由多个神经网络层组成。一般来说深度学习模型包含更多的参数。**

**2.特征表示：传统机器学习算法需要手动设计特征工程，来对原始数据进行特征提取和选择，深度学习模型能够自动学习特征表示，从原始数据中直接学习高层次的特征，无需人工设计。**

**3.计算需求：由于深度学习模型的复杂性，训练过程通常需要大量的计算资源和时间。深度学习模型需要大规模的数据集和高性能的计算设备来训练和调优。传统机器学习算法通常计算需求较低，训练速度相对较快，可以在较少的计算资源下进行训练和调优。**

1. 数据一般有哪些类型的属性？这些类型属性上可以施加哪些运算操作？

**数据一般可以分为以下几种类型的属性：**

**名词属性，可以施加相异性操作（=，≠）；**

**顺序属性，可以施加相异性（=，≠）、有序性（< >）操作；**

**区间属性，可以施加相异性（=，≠）、有序性（< >）、加法(+ -)操作；**

**比值属性，可以施加相异性（=，≠）、有序性（< >）、加法(+ -)、乘法(X /)操作。**

1. 什么是数据清洗？假设我们从网络上收集了1T的文本类型数据用于训练一个文本问答系统，那么我们一般要做哪些数据清洗工作？

**数据清洗是数据预处理的一个重要步骤，其目的是识别和纠正数据中的错误、不完整、重复或不一致的部分，以确保数据的质量和可用性。**

**我们一般需要做以下数据清洗工作：**

**1.数据格式转换及规范化、归一化处理**

**2.初步决定数据相关属性列表**

**3.处理数据缺失值（删除或者填充估计/默认/缺省值）**

**4.处理异常值**

**5.选择采样方法**

1. 评价分类问题算法结果常用的指标有哪些？请给出具体公式定义。

**通常包括以下几种：**

**1.准确率（）**

**其中， 表示真正例数， 表示真负例数， 表示假正例数， 表示假负例数。**

**2.精确率（）**

**其中， 表示真正例数， 表示假正例数**

**3.召回率（）**

**其中， 表示真正例数， 表示假负例数**

**4.F1分数（）**

**其中，表示精确率，表示召回率**