**1.1简要叙述训练数据集和测试数据集的含义。**

测试集：

机器学习学科中，学习样本三部分之一，测试集用来检验最终选择最优的模型的性能如何。

训练集：

机器学习学科中，学习样本三部分之一，训练集用于建立模型。验证集用来确定网络结构或者控制模型复杂程度的参数，而测试集则检验最终选择最优的模型的性能如何

**1.2 一个初学机器学习的朋友对房价进行预测。他在一个N=1000个房价数据的数据集上匹配了一个有533个参数的模型，该模型能解释测试数据集上99%的变化，请问该模型能很好地预测来年的房价吗？简要说明原因。**

不能，因为这个模型他设置了533个参数，参数过多，导致过拟合现象，虽然训练集拟合很好，但是把错误点也考虑了，实际误差会较大

**1.3 请简要说说一个完整机器学习项目的流程。**

（1）将问题抽象成明确的数学问题

（2）获取数据集，数据要对整个问题的情况具有一定代表性，分布合理

（3）特征预处理与特征选择，进行数据处理，数据清洗

（4）训练模型与调优模型算法

（5）模型诊断，即确定模型调优的方向思路

（6）模型融合，提高算法的效果

（7）上线运行，测试工程

**1.4试简述机器学习在互联网搜索的哪些环节起什么作用？**

（1）搜索引擎直接给出搜索的问题答案，如搜索春节多少天，直接给出距离春节天数

（2）直接进行图片、视频（等多元数据）的搜索

（3）更精准的排序（也可成为「精准营销」的部分）

（4）对用户行为进行综合分析（如历史搜索数据、点击模式、身份信息等进行结构化 信息整合）

（5）对话式智能交互搜索：如Baidu的语音搜索、利用Siri进行搜索

（6）对垃圾网站的筛选（模式识别）