1. 为了测试机器学习的效果，通常有两个集合：训练集和测试集。
2. K-紧邻算法：采用测量不同特征值之间距离的方法来分类。它的工作原理是：存在一个样本数据集合，称为训练样本集，而且样本集中每个数据都存在标签。输入没有标签的新数据时，将新数据的特征与样本集数据对应的特征进行比较，然后提取样本特征值最相似的前k个分类标签，将出现次数最多的分类，作为新数据的分类。
3. K-紧邻算法的优点：精度高，对异常数据不敏感。缺点：计算的时间和空间复杂度高。
4. K近邻算法的伪代码：对于未知类型属性的每个点依次执行如下操作：1）计算已知类别数据集中的点与当前点之间的距离。2）按照距离递增排序。3）选取与当前距离点自最小的前K个点。4）确定前k个点所在类别的出现概率。5）返回前k个个点出现概率最高的类型作为当前点的预测分类。