1. 如何确保训练的模型没有过拟合？并阐述原因。

为了确保训练模型的过拟合，这里有几个方案可以解决

方案1：清洗数据。即去除掉可能有问题的数据（减少了训练样本）；

方案2：添加正则项（正则项的作用主要是通过参数来控制模型的复杂度，常用的有L1和L2正则项）；

方案3：采用“Early Stopping”方法。即所谓的提前停止方法。一般的做法是，在训练的过程中，记录到目前为止最好的validation accuracy，当连续10次Epoch（或者更多次）没达到最佳accuracy时，则可以认为accuracy不再提高了。此时便可以停止迭代了（Early Stopping）。

方案4：添加Dropout层。即通过修改网络本身的结构来实现的。（在训练开始时，随机得删除一些（可以设定为一半，也可以为1/3，1/4等）隐藏层神经元，即认为这些神经元不存在，同时保持输入层与输出层神经元的个数不变，这样就当于生成了很多个不同的简单网络。）

方案5：增加训练集样本的个数。即给予模型提供更多的数据，这些数据中可能会包含更多的可能性，更容易让模型学习到通用的特征表示。

1. 回归决策树算法和分类决策树算法有什么区别与联系？

区别：

1. 分类树主要用于将数据集分类到响应变量所对应的不同类别里，通常响应变量对应两类0 or 1. 如果目标变量对应了2个以上的类别，则需要使用分类树的一个扩展版C4.5（很popular）。然而对于一个二分类问题，常常使用标准的CART算法。不难看出分类树主要用于响应变量天然对应分类的情况。

回归树主要用于响应变量是数值的或者连续的，例如预测商品的价格，其适用于预测一些非分类的问题。

1. 标准分类树的思想是根据数据的相似性(homogeneity)来进行数据的分类。举一个简单的例子就是：对于标准的非纯度计算，一般会基于一个可计算的模型，比如entropy 或者Gini index通常用来量化分类树的均匀性。

用于回归树里的目标变量是连续的，我们通常用自变量拟合一个回归模型。然后对于每个自变量，数据被几个分割点分离。在每个分割点，最小化预测值和真实值的误差和 （SSE）得到回归模型的分类方法。

联系：

1. 他们都是用于决策的算法。
2. 什么是贝叶斯定理，它是如何使用在机器学习中的？

简单的来说，贝叶斯定理是用来预测概率的，也就是：某个条件下，一件事发生的概率有多大？为了把这个问题说明的更加清楚，还需要涵盖逆概的场景。也就是我们在不清楚我们不知道某一些情况，当我们通过一些实践后，预测下一次发生某件事情的情况。

贝叶斯在机器学习中解决的一类很重要的问题就是分类问题。很具体的一个例子，去年上海市刚决定举行垃圾分类的时候，很多人还是没有办法分清楚什么是湿垃圾，什么是干垃圾。那么有一些人开发了app来通过图像识别自动判断这个属于什么类型的垃圾。而上面这个例子中所用到的一个基本原理就是贝叶斯定理。

1. 手推逻辑回归（求导、梯度更新）

A picture containing diagram

Description automatically generated

1. 手推SVM公式（原问题，对偶问题，参数求解）。

Diagram

Description automatically generated

A close-up of a document

Description automatically generated with low confidence