第一次作业

1. 什么是机器学习？请查找资料给出一个你觉得满意的定义或解释。

答：机器学习是人工智能的一条分支，可以通过机器学习为手段解决人工智能的问题，机器学习理论可以设计或者构建一些让机器自动“学习”的算法，机器学习所设计的算法是一类从数据中自动获取规律，并利用规律对未知数据进行预测的算法。

2. 什么是过学习？有哪些避免过学习的方法？

答：过学习，也就是过拟合，即overfitting，通常发生在模型过于复杂的情况下，如参数过多，overfitting会使得模型的预测能力变弱，增加数据的波动性，发生该现象是因为评判训练模型的标准不适用于评判该模型，因为模型的训练是基于训练集的，但是一整个模型的好坏并不完全依赖于训练集，当模型完全依赖于训练集以求更好的性能是，它在未知的数据集（测试集）上就会表现出反而更差的效果，这就是所谓的过拟合，也即过拟合。

方法：（1）委托标注公司对数据进行标注，人为的对数据的特征进行标注.。

1. 进行正则化调整。可知的是过拟合的一个重要原因就是参数过多，通过正则化的办法可以解决这个问题。定义损失函数来描述基准和现实情况的区别，但是损失函数只考虑在训练集上的经验风险，就会造成过拟合，所以通过在损失函数中加入新的描述模型复杂度的正则项，这样就可以避免模型太过复杂造成的过拟合。
2. 也可以通过单纯的增加训练样本来进行过拟合的问题解决。增多训练样本，可以避免某些特殊的特征项出现。

3. 什么是交叉验证？它有哪些具体做法？

答：交叉验证，即将样本集切割成较小的子集，通过在一个子集上分析，在其他子集上用来后续进行确认的方法，换言之，就是将一整个数据集分成了训练集和验证集。

具体做法有k-交叉验证，将训练集分为k份，每次选取一块作为验证集，其他作为训练集，每个子样本验证一次，一共进行k次交叉验证。

1. 请查找资料给出一个正则化方法的文字定义。

答：机器学习的训练过程，就是通过构造一个足够好的函数F用以在数据集进行验证，为了定义一个什么叫做足够好，引入了损失函数，用来描述基准函数和模型预测值之间的差距，因此机器学习的训练过程转化为了训练集D上loss function，也即L（F）最小化的问题，公式是，但是loss function只考虑在训练集上的经验风险，这样可能会造成过拟合，也就是因为过于的最求训练集上的最优化，反而导致了模型复杂程度的加剧，为了对抗过拟合，我们需要向损失函数中加入描述模型复杂度的正则项，将经验风险最小化转化为结构风险最小化，修改后的公式如下：



通过引入正则项的方法就可以由小避免因过拟合而造成的模型复杂化问题，避免了糟糕的模型。