1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：选择k的时候采用验证集合、交叉验证的方法，每次重新划分训练集和测试机都需要重新计算一下所有点的距离，时间复杂度会很大，有没有什么其他选择k的方法？

讨论后的理解：可以事先大致确定k的范围再去验证，这样可以减少次数，降低时间复杂度，从而优化算法。

1. 提出的问题2：Bagging为什么会对不稳定学习算法的性能显著提高，而对于稳定的的分类器反而会降低准确率？

讨论后的理解：Bagging性能依赖基分类器稳定性，基分类器不稳定，Bagging有助于降低训练数据的随机波导致的误差，如果基分类器稳定，则组合分类器的误差主要为基分类器偏倚所引起的，此时Bagging对基分类器性能可能没有显著改善，甚至会降低。

1. （必填）别人提出的问题的理解（选择几个问题罗列，并给出理解）：
2. 问题3：knn在样本不平衡时会有缺陷，例如有一类样本容量很大，其他类样本容量小，导致大容量的样本在k个邻居中占多数。有无改进的办法？

自己的理解：可以对k邻近点进行加权，即距离近的权值大，距离远的权值小。

1. 问题4：为什么Bagging算法取一个自展复制的时候每个样本平均有63.2%的原始样本？

自己的理解：这是采用“自助法”减少训练样本规模不同造成的影响的方法，蚕蛹这种方法时，样本在m次采样中始终不被采到的概率是，取极限得到0.368，即通过自助采样的方法，初始数据集中约有36.8%的样本未出现在采样数据集中，所以自展复制的时候每个样本平均有63.2%的原始样本。

1. 问题5：knn每个样本都要遍历一遍数据集，这样算法时间复杂度很高。如何优化knn算法？

自己的理解：输入X可以采用BallTree或KDTree两种数据结构，优化计算效率，可以在实例化KNeighborsClassifier的时候指定。

1. （必填）读书计划
2. 本周完成的内容章节：3.10-3.11

2、下周计划：2.1-2.5

四、（选做）读书摘要及理解或伪代码的具体实现（读书摘要、伪代码的具体实现代码等可以写到这个部分）

1、读书摘要及理解（选做）

KNN：

KNN是一种**非参的**，**惰性**的算法模型。非参的意思意味着这个模型不会对数据做出任何的假设，与之相对的是线性回归（我们总会假设线性回归是一条直线），KNN建立的模型结构是根据数据来决定的。惰性是指KNN算法不需要有明确的训练数据的过程，或者说这个过程很快。

KNN算法优点：

1. 简单易用，相比其他算法，KNN算是比较简洁明了的算法。即使没有很高的数学基础也能搞清楚它的原理。
2. 模型训练时间快，KNN算法是惰性的。
3. 预测效果好。
4. 对异常值不敏感

KNN算法缺点：

1）对内存要求较高，因为该算法存储了所有训练数据

2）预测阶段可能很慢

3）对不相关的功能和数据规模敏感