1. 问题描述

科技发展日新月异，智能识别推陈出新。随着智能手机、掌上电脑等移动信息工具的普及，手写识别技术也进入了规模应用时代。阿拉伯数字是唯一被世界各国通用的符号，手写数字识别作为模式识别领域的一个重要问题，有着重要的理论价值。

手写数字识别是将带有手写数字的图片输入到已经训练过的机器，且机器能够很快识别图片中的手写数字，并将之作为输出打印出来。大致可以分为预处理、特征提取、分类器分类等几个过程，预处理为能获得正确的有关数字的特性作了基础，而得到有关的数字特征则保证了最后分类的有效性。

我们小组选择了SVM模型进行手写数字识别，SVM是目前分类器中效果最好的之一，其核函数可以将数据映射到高维空间中，可以对数据很好的分类。该算法是基于VC维及结构风险最小化原则，对线性分类器提出的设计准则，最开始的应用就是在手写字符的识别上，美国邮服务政局利用邮政标准的手写数字库做了测试对比，发现SVM方法比其他复杂的神经网络算法效果好。它具有通用性，可以在各类函数中用来构建新函数，具有较好的鲁棒性，在使用上不需要借助复杂方法做支撑，所以我们小组选择使用SVM进行手写数字识别。

1. 数据来源和处理

我们采用了经典的MNIST数据集，该数据集是直接从网上下载的，（数据量、数据的预处理阶段）。

数据的处理:根据找到的

1. 使用的package

我们使用了pandas、numpy、matplotilb（每个package的用处）,sklearn

1. 建模过程
2. 建模时的特征选取
3. 调参
4. 是否对原始模型有自己的更改
5. 结果分析
6. 和其他模型的对比分析
7. 决策树模型

决策树是一种类似流程图的从上到下的树形结构，它的优点是准确性高、速度快，易于理解，基本上无需对要作判别的数据作处理，即使需要也很容易。但它的缺点也很明显，当存在信息类别多且数目不统一的情况时，出错的速度在某种程度上会出现很大幅度的增加。

1. KNN近邻分类模型

KNN是一种懒惰学习（Lazy learning）方法，其所谓训练过程就是将训练数据存入空间中。然后在测试时，将待测试数据投入到数据空间中寻找近邻，通过近邻类别的投票来确定该数据的类别。

这种方法精度高，对异常值不敏感，无数据输入假定，但是计算的复杂度高，空间复杂度也高。

1. 神经网络模型

神经网络是由一定数目的神经元按照彼此之间某些相连的关系互相衔接组成，但用这样的网络来处理大尺寸的图片参数太多，计算量巨大，很难训练。神经网络因训练时间和分类时间太长，一般不用于大字符集的直接分类。

1. 运行结果
2. 改进手段

我们在测试的时候发现SVM的几个缺点：   
1）不适用于大规模数据，SVM在小规模数据中可以很好抓住特征和数据之间的非线性关系。然而数据量大的时候，在训练速度和分类精度都不是那么的恰如人意。   
2) kernel选择困难，调参麻烦