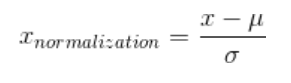
# 报告

## 算法原理

在数据进行模型训练前要先对数据进行标准化处理。使用的是Z-score处理方法，其原理如下：



其中，为样本均值，为样本标准差。

对于实验涉及到最小距离、k近邻、SVM三种方法，原理在下面进行依次介绍。

首先是最小距离分类。它是根据已知样本中各类别的中心到未知样本的距离大小，来将未知样本进行分类。判断未知样本所属哪一类的标准就是未知样本到已知类别中心的距离最小。

其次是k近邻分类。k近邻的分类过程和最小距离的过程在前期处理是一样的，它与最小距离不同之处有两点：1、k近邻计算的距离是待测试样本到每个训练样本的距离；2、在最后的判别上，k近邻是选择前k个最小距离，并通过判断这k个距离中哪个类别出现的次数最多，则当前被测试的样本的类别就出现次数最多的类别。

最后是SVM，SVM是一个有着严格数学证明的典型的二分类模型。在进行分类时是找到一个能将样本分开的超平面，且这个超平面到样本的距离最大，所以SVM也称作最大间隔分类器。其优化目标如下式所示：

SVM也可以应用到多分类问题上。处理的方式有两种：1、是一类对一类方式，当有n个类别时，这样会得到你n(n-1)/2个分类超平面。2、是一类对其余的方式，当有n个类别时，会得到n个分类超平面。

## 分类过程

程序实现分类过程如下：

一：最小距离分类过程：

（1）数据提取：确定类别n，并提取每一类所对应的已知的 样本。进行Z-score数据归一化处理。

（2）每一类的每一维都有特征集合，通过集合，可以计算出 一个均值，也就是特征中心。

（3）利用选取的距离准则（本次使用的是欧式距离），对待分 类的本进行判定，即选择聚类最小的作为当前被测试样 本的类别。

二：k近邻分类过程：

（1）数据提取：将训练数据和测试数据分别从文件中读取出 来。进行Z-score数据归一化处理。

（2）计算训练集中样本与待测样本之间的距离。

（3）按距离排序。

（4）选取与当前样本距离最小的k个邻居样本。

（5）确定此k个样本中各个类别出现的次数。

（6）出现次数最高的类别作为该样本的预测分类结果。

三：SVM分类过程：

（1）数据提取：将训练数据和测试数据分别从文件中读取出 来。进行Z-score数据归一化处理。

（2）通过sklearn函数库中的svm.SVC()来建立分类模型，并通 过fit()和训练样本来拟合模型

（3）之后通过模型对测试样本进行预测，同时还可输出每个 测试样本的分类概率。

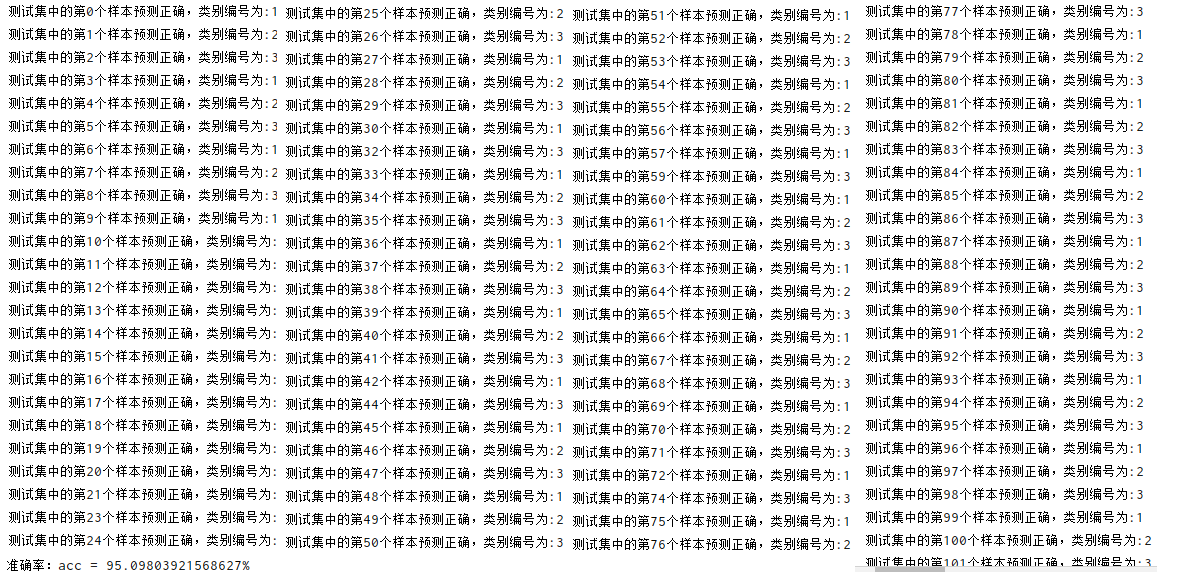
## 结果对比

一：最小距离的分类结果：



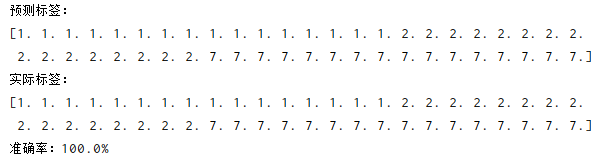
模型的分类准确率为：90.48%

二：k近邻分类结果



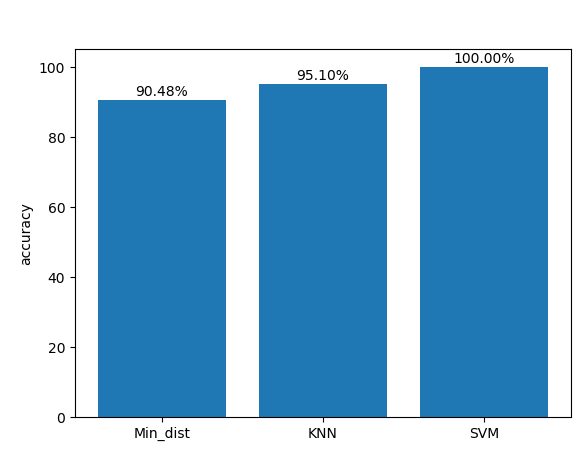
模型的准确率为：95.098%

三：SVM分类结果：



模型的准确率达到100%

四：三种模型准确率柱状图：



SVM的准确率最高，K近邻次之，最小距离最差