1. 算法原理

1.1最小距离

在一个n维空间中，最小距离分类法首先计算每一个已知类别X（用向量表示是[x1,x2,…,xn]）的各个维度的均值，形成一个均值M，(用向量表示[m1,m2,…,mn])。那么第二个类别依此类推，为类别Y（用向量表示是[y1,y2,…,yn]）的各个维度的均值，形成一个均值A，(用向量表示[a1,a2,…,an]).

在预测的数据中，分别计算预测数据到不同的类别的均值的距离，该预测数据属于距离最小的距离。

1.2 K近邻

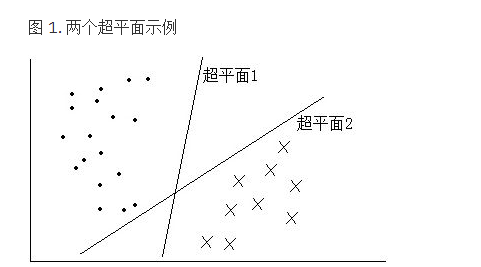
KNN是一种即可用于分类又可用于回归的机器学习算法。对于给定测试样本，基于距离度量找出训练集中与其最靠近的K个训练样本，然后基于这K个“邻居”的信息来进行预测。

在分类任务中可使用投票法，选择这K个样本中出现最多的类别标记作为预测结果；在回归任务中可使用平均法，将这K个样本的实值输出标记的平均值作为预测结果。当然还可以基于距离远近程度进行加权平均等方法。

1.3 SVM

支持向量机(SVM)是90年代中期发展起来的基于统计学习理论的一种机器学习方法，通过寻求结构化风险最小来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。

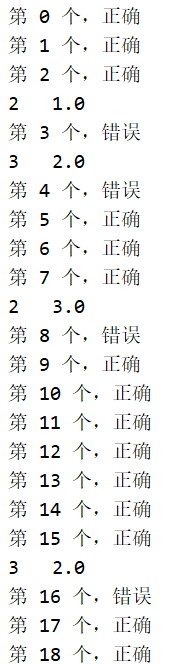
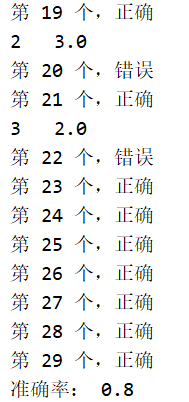
支持向量是距离分类超平面近的那些点，SVM 的思想就是使得支持向量到分类超平面的间隔最大化。距离分类超平面近的那些点到该超平面的间隔最大化代表了该超平面对两类数据的区分度强，不容易出现错分的情况。如图 1 所示，支持向量到超平面 1 的间隔大于支持向量到超平面 2 的间隔，因此超平面 1 优于超平面 2。



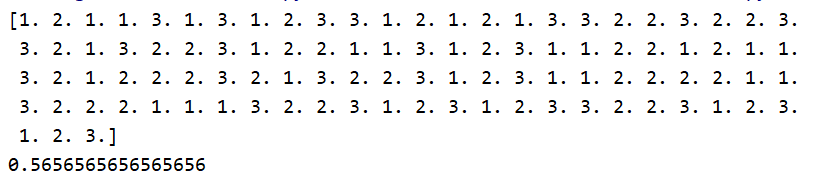
SVM 可以很好得解决二分类问题，对于多分类情况，就需要对模型进行改动。如 one-versus-rest 法，这种方法每次选择一个类别作为正样本，剩下其他类别作为负样本，假设一共有 3 个类别，这样相当于训练出了 3 个不同的 SVM。然后将测试数据分别带入 3 个 SVM 模型中，得到的 3 个结果中的最大值则为最终的分类结果。

1. 分类过程

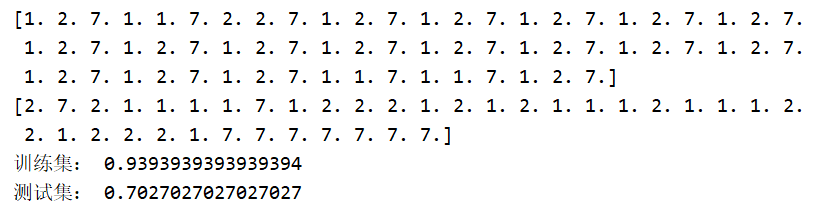
最小距离：

KNN：



SVM:



1. 结果图形化比较

可以发现：用SVM的准确率较高，最小距离和k近邻的准确率差不多。