KNN实现数字识别

createDataSet()生成测试数组  
kNNclassify(inputX, dataSet, labels, k)分类函数

inputX 输入的参数  
dataSet 训练集  
labels 训练集的标号  
k 最近邻的数目

KNN 算法原理

KNN算法则是从训练集中找到和新数据最接近的k条记录，然后根据他们的主要分类来决定新数据的类别。该算法涉及3个主要因素：训练集 ， 距离或相似的衡量 ， k值

kNN算法的指导思想是“近朱者赤，近墨者黑”，由你的邻居来推断出你的类别。

1）算距离：给定测试对象，计算它与训练集中的每个对象的距离  
    2）找邻居：圈定距离最近的k个训练对象，作为测试对象的近邻  
    3）做分类：根据这k个近邻归属的主要类别，来对测试对象分类

2、距离或相似度的衡量  
什么是合适的距离衡量？距离越近应该意味着这两个点属于一个分类的可能性越大。  
觉的距离衡量包括欧式距离、夹角余弦等。  
对于文本分类来说，使用余弦(cosine)来计算相似度就比欧式(Euclidean)距离更合适。  
  
3、类别的判定  
投票决定：少数服从多数，近邻中哪个类别的点最多就分为该类。  
加权投票法：根据距离的远近，对近邻的投票进行加权，距离越近则权重越大（权重为距离平方的倒数）