Programming report

李逸思 自动化系 2016310707

1） 最近邻分类算法不需要使用训练集进行训练，故训练时间复杂度为 0；设训练样本数目为 n，测试样本数目为 m，则最近邻分类算法分类时间复杂度为 O（mn），空间复杂度为 O（n）。

使用不同规模的训练样本，采用欧氏距离度量样本间的相似性，得到最近邻分类器的分类正确率变化如表 1.1

表 1.1 训练样本规模对最近邻分类器正确率的影响

训练样本数量	正确率
50	65.51%
100	72.53%
500	83.91%
1000	88.42%
2000	91.12%
5000	93.31%

由表可知，训练集越大最近邻分类器正确率越高，但计算速度也越慢。

2） 采用欧氏距离度量样本间的相似性，训练样本数量取 500，分别取 k=2,3,4,5,6,7，得到 k 近邻分类分类器的正确率变化如表 1.2

表 1.2 不同 k 值对 k 近邻分类器正确率的影响

k 值	正确率
2	81.88%
3	83.44%
4	83.73%
5	83.23%
6	83.38%
7	82.59%

由表可知，k 过大或过小都不能使正确率最优，训练样本数为 500 时，k=4 时正确率最高。

3） 训练样本数取 500，取 k=4，分别采曼哈顿距离、欧式距离、切比雪夫距离这三种闵式距离度量样本间的相似性，得到 k 近邻分类分类器的正确率变化如表 1.3

表 1.3 不同距离度量对 k 近邻分类器正确率的影响

距离度量	正确率
Manhattan Distance	81.31%
Euclidean Distance	83.76%
Chebyshev Distance	44.24%

由表可知，对于此数据集，采用欧氏距离距离度量正确率最高，曼哈顿距离度量效果与欧式距离相近，切比雪夫距离度量正确率较低。

4） 原数据集中，每个样本对应 784 个特征值，但这些特征值中可能存在冗余信息从而对分类造成干扰，若能通过一定的线性变换去除冗余信息，则有望提高正确率。

记训练数据集 train_x 为矩阵 A，矩阵 A 经式（1.1）的正交变换后得到矩阵 B，取

向量 m 为矩阵 M 每列的平均值，则数据经过式（1.2）的线性变换后，冗余性降低。

$$A*M=B \tag{1.1}$$

$$x^{'}=m*x \tag{1.2}$$

仿真验证，训练样本数取 500，取 k=4，采用切比雪夫距离度量样本间的相似性，不做线性变换前分类正确率为 46.3%，线性变换后分类正确率为 55.76%，说明线性变换提高了分类正确率。

5） 用切线距离替代闵式距离， 记训练集中一张图片为 x’，具体计算方法描述如下：

Step1 x’还原为矩阵后，进行平移变换（或其他变换），并将得到新的图形对应的矩阵拉伸得到向量 x

Step2 切向量 T=x-x’

Step3 最小化式（1.3）中目标函数得到切线距离 D(x,x’)

$$D(\mathrm{x},\mathrm{x}^{'})=\min_{a}\|(\mathrm{x}^{'}+\mathrm{aT})-\mathrm{x}\| \tag{1.3}$$

Step4 找到与训练样本中与测试样本 x 最接近的 k 个近邻

Step5 根据近邻类别确定 x 类别

使用切线距离度量对 MNIST 分类器仿真测试，训练样本数取 500，取 k=4 时，分类正确率为 85.21%，比欧式距离有提高。