**Programming: KNN classification for MNIST**

陈斯杰 2016310721 自博16

主要程序为kNN.m和NCA.m。 kNN.m使用了MATLAB Parallel Toolbox做了并行加速；切向量计算使用了C语言混编来提升速率。预测结果的文件名形如TrainNum60-k1-euc.res.mat；性能分析结果文件名形如：TrainNum60-k1-euc.txt；它们都表示60个训练样本，k取1时，使用欧几里得度量的情况。

**1)训练样本量对于分类器性能的影响**

我们使用了1-闵式距离、1.5-闵式距离、欧氏距离、切线距离，分析了最近邻算法在样本数为60、600、6000、60000时的性能。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 距离  训练样本 | 1-闵式距离 | 1.5-闵式距离 | 欧氏距离 | 切线距离 |
| 60 | time=34.246463 secs, accuracy=0.6273000000 | time=35.349806 secs, accuracy=0.6395000000 | time=34.041252 secs, accuracy=0.5418000000 | time=39.174500 secs, accuracy=0.7182000000 |
| 600 | time=294.463994 secs, accuracy=0.8397000000 | time=73.956833 secs, accuracy=0.8504000000 | time=277.201242 secs, accuracy=0.8548000000 | time=136.636382 secs, accuracy=0.9086000000 |
| 6000 | time=239.880466 secs, accuracy=0.9310000000 | time=455.524762 secs, accuracy=0.9352000000 | time=227.105914 secs, accuracy=0.9392000000 | time=1349.569770 secs, accuracy=0.9630000000 |
| 60000 | NA | NA | NA | time=6523.787159 secs, accuracy=0.9797000000 |

由于60000个样本点计算量太大，我仅仅计算了基于切线距离的分类器性能。

从表中可以得出闵式距离参数值越大，分类越准确；基于切线距离的分类器是最准确的。

**2) k值对于分类器性能的影响**

我们考察6000个训练样本时，k值对于KNN分类性能的影响。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 距离  K值 | 1-闵式距离 | 1.5-闵式距离 | 欧氏距离 | 切线距离 |
| 1 | time=239.880466 secs, accuracy=0.9310000000 | time=455.524762 secs, accuracy=0.9352000000 | time=227.105914 secs, accuracy=0.9392000000 | time=1349.569770 secs, accuracy=0.9630000000 |
| 3 | time=239.880466 secs, accuracy=0.9310000000 | time=753.620202 secs, accuracy=0.9351000000 | time=662.468712 secs, accuracy=0.9372000000 | time=992.789580 secs, accuracy=0.9615000000 |
| 5 | time=240.625380 secs, accuracy=0.9281000000 | time=755.190019 secs, accuracy=0.9356000000 | time=648.749592 secs, accuracy=0.9372000000 | time=1344.821988 secs, accuracy=0.9591000000 |
| 7 |  | time=747.766152 secs, accuracy=0.9312000000 | time=648.804786 secs, accuracy=0.9379000000 | time=1358.408490 secs, accuracy=0.9594000000 |

从数据可以看到，在6000个训练样本的情况下，k值对于分类效果影响不大。

**3) 不同距离度量对于分类器性能的影响**

见上面的表

**4) 对数据做变换后进行分类**

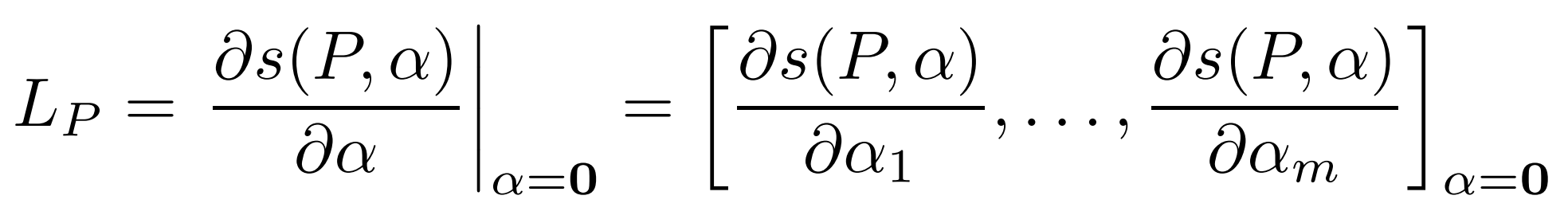
定义点之间的软距离，使用梯度下降法求出参数A。程序见NCA.m。但是由于梯度下降计算量很大，优化不佳，本小问没有做出数值结果。

**5）切线距离代替欧氏距离**

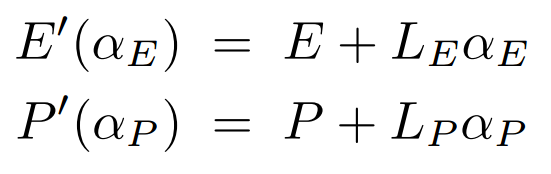
考察作用在图像P上的单参数变换s(P,α)，如果s对于α是可微的，则可以得到一个流形SP，其定义为：

。

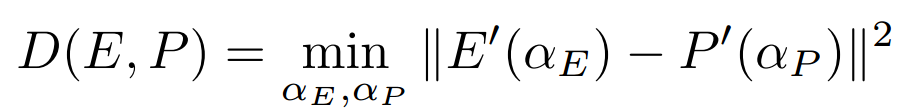
该流形的切超平面为：

，

我们使用一阶近似来表示流形上的点：



则切距离可以定义为：

。

在本题中，我考虑了九种单参数图像变换：垂直平移、水平平移、放缩、双曲变换1、双曲变换2、旋转、线宽、加性亮度调整、积性亮度调整。

为提升计算速率，实现切距离计算的程序在tangentDistanceCImpl文件夹中，使用C语言实现，通过matlab/C混编接口调用。