Classiffication@affNIST dataset

姓名：张毅　　学号：21721190　　日期：2018-5

1　问题描述

在 AffNIST 数据集上训练模型进行分类任务，识别 0-9 十个数字。

数据集：AffNIST（https://www.cs.toronto.edu/ tijmen/affNIST/），训练数据与测试数据默认已经划分好，试验中采用默认的划分方式。

2　方法及原理

实验中采用了比较常用的 KNN 和 SVM 对图片进行分类，代码已经上传到 github 仓库 (yizhangzc/course), 本次实验与第二次实验代码均放置在 classification 文件夹下，运行方式及运行环境见 README.md。

KNN：如果一个样本在特征空间中的 k 个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性。

SVM：基于结构风险最小化理论之上在特征空间中构建最优超平面，使得学习器得到全局最优化。对于线性不可分的情况，通过使用非线性映射算法将低维输入空间线性不可分的样本转化为高维特征空间使其线性可分。

3　实验结果

试验直接在原始数据上进行，每个训练样本作为 1600 维的向量输入模型进行训练。由于数据集太大，训练时间非常长，所以分别只使用了训练与测试集中的部分样本 (每个.mat 文件中均取 1000 个样本)。最终结果为：

KNN:	准确率：68.8%	f1 值：67.5%
SVM:	准确率：82.2%	f1 值：82.0%

```
result: accuracy: 0.668  f1_score: 0.6746713379800807
confusion_matrix:
[[ 80   8   0   1   0   2   4   1   1   1]
 [  0 114   0   0   0   0   0   0   0   0]
 [  5  20  59   2   1   1   1  12   2   0]
 [  0   4   3  77   0   1   2   4   8   2]
 [  0  31   1   0  51   0   4   1   1   9]
 [  5  12   0  12   2  52   2   2   0   2]
 [  0  15   0   0   3   3  74   0   1   0]
 [  0  35   0   0   1   0   0  64   0   3]
 [  2  17   1   8   1   3   4   3  50   8]
 [  2  19   0   0  21   0   0  11   1  47]]
```

(a) KNN 混淆矩阵

```
result: accuracy: 0.822  f1_score: 0.8195913193186936
confusion_matrix:
[[ 87   0   0   0   2   2   3   0   3   1]
 [  0 112   1   0   0   0   0   0   1   0]
 [  3   1  82   3   1   2   3   2   5   1]
 [  3   0   1  80   0   5   0   4   5   3]
 [  5   0   1   0  78   0   5   0   0   9]
 [  4   1   0   7   0  72   0   2   3   0]
 [  3   2   0   0   3   3  82   1   1   1]
 [  0   6   3   1   2   0   0  83   1   7]
 [  3   1   0   5   6   4   3   0  68   7]
 [  2   3   0   2  10   0   0   3   3  78]]
```

(b) SVM 混淆矩阵

4　总结

（1）实验在 1600 维的向量上进行，未能成功尝试先提取特征在使用 KNN 或 SVM 进行分类比较遗憾。（2）SVM 调参至关重要，使用 sklearn 包中的默认参数效果极差（大约 20%）, 通过实验，得到效果比较好的参数是,kernel: 高斯，gamma:0.03,C:30。（3）KNN 中 K 对模型的影响相对较小，最终取值 5。