模式识别

基础实验报告

MINST 数据集探究

学院：计算机科学与技术

学号：1170300909

姓名：武磊

指导老师：金野

日期：2020.4.14

一、课题来源

概述：图像分类是模式识别中的常见任务。本次项目希望在机器学习的经典数据集上运用一些模式识别的经典方法。通过实践的方式来辨别各种方法使用的场景和优劣。

二、主要研究内容，方案

主要内容：在选取的经典数据集—MINST 手写数据集上运用经典的模式识别方法。

主要方法：KNN，SVM，PCA，Naïve Bayes

数据集：MINST 数据集

开发工具：Sklearn，Pycharm，Jupiter Notebook

三、人员安排

个人完成

四、研究方案，可行性分析

研究方案：完成对 MINST 数据集中十个数字 0-9 的多分类。

可行性分析：这是经典的问题，实现完全没有问题，主要是借此理解常用的模式识别方法。

五、实验分析及结论

下面根据实验采用的方法依次给出实验探究的结果。

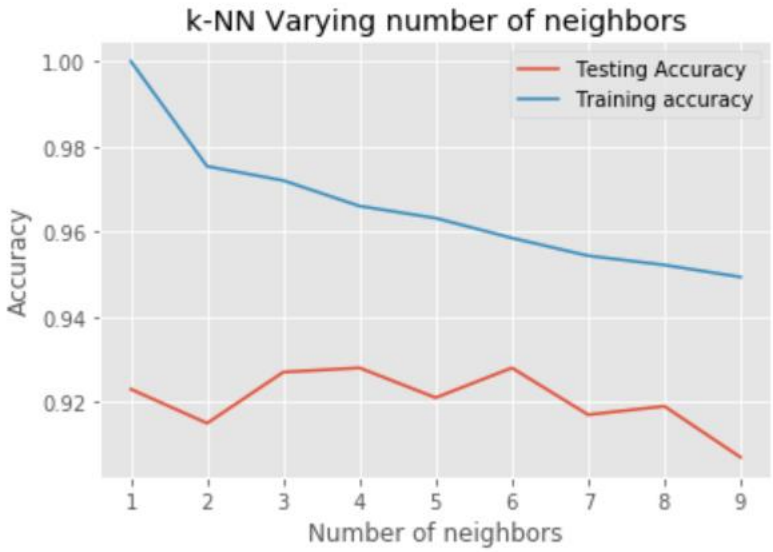
除了上述的实验方法，还尝试了一些有趣的 embedding 方法。

5.1 实验分析

5.1.1 KNN

Knn 是最朴素的分类器之一，思想原理很简单。方法中主要的超参数是 k，表示能够影响样本类别的 k 个最近点。由于 Knn 的计算复杂度较高，我们选取了 7000 个数据用于训练和测试。

参数 k 的选取:



通过测试和训练的正确度，我们可以看到随着 k 的增大，样本的训练集正确率一直下降，而测试集略有上升，之后再下降。这意味着 k 的适度增加，模型的鲁棒性能够有提升，但是若 k 过大，噪声点对样本分类的影响增加，模型性能下降。

Sklearn 中也有自动化调参的工具，如 GridSearchCV，通过交叉验证来对模型的性能进行评价，然后返回性能最高的参数 k 和对应的评分。

```
{'n_neighbors': 4} 0.9258333333333333
```

在 knn 实践过程中，还对分类器模型的评价准则有了深刻理解。一般我们通过混淆矩阵计算查准率，召回率，AOC 曲线来判别分类器的分类性能。

5.1.2 SVM-PCA

SVM 是深度学习之前最广泛使用的分类器，通过计算支持向量，得到基于最大化间隔的最优分类器。对于多分类问题，我们采取 kernelSVM 在高维空间中达到线性可分的目的。我们选取了 5000 个数据用于训练，4000 用于测试。

初步尝试使用 rbf 核的 kernel SVM 对 MINST 进行分类。

	precision	recall	f1-score	support
0	0.98	0.94	0.96	370
1	0.98	0.97	0.98	450
2	0.68	0.98	0.80	418
3	0.88	0.93	0.90	408
4	0.91	0.93	0.92	418
5	0.95	0.90	0.93	372
6	0.99	0.81	0.89	378
7	0.97	0.85	0.91	411
8	0.92	0.83	0.87	384
9	0.91	0.86	0.88	391
micro avg	0.90	0.90	0.90	4000
macro avg	0.92	0.90	0.90	4000
weighted avg	0.92	0.90	0.90	4000

```
Wall time: 1min 7s
```

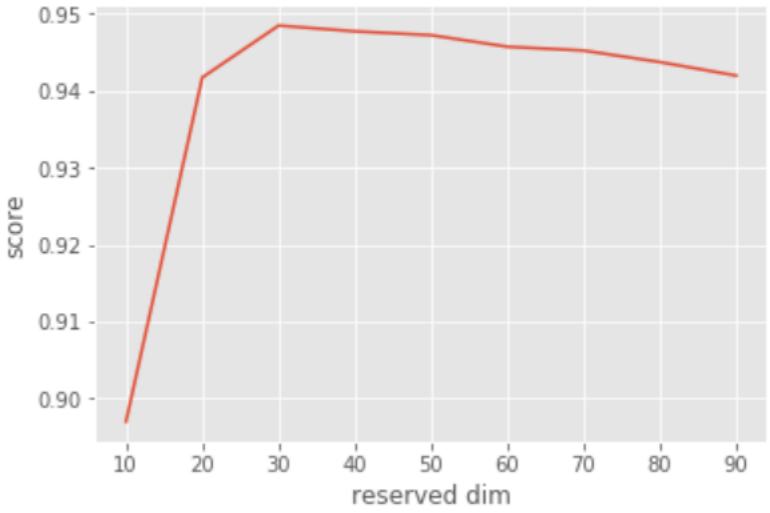
很明显的感受到 SVM 对于数据量小于 10000 的样本能够很快地完成训练。但是当使用完全样本进行训练，速度非常慢。

如何加快 SVM 的速度呢？我们考虑使用 PCA 方法。

	precision	recall	f1-score	support
0	0.96	0.98	0.97	370
1	0.97	0.99	0.98	450
2	0.96	0.96	0.96	418
3	0.93	0.94	0.93	408
4	0.94	0.95	0.94	418
5	0.94	0.95	0.95	372
6	0.96	0.95	0.95	378
7	0.97	0.92	0.95	411
8	0.93	0.92	0.93	384
9	0.91	0.92	0.91	391
micro avg	0.95	0.95	0.95	4000
macro avg	0.95	0.95	0.95	4000
weighted avg	0.95	0.95	0.95	4000

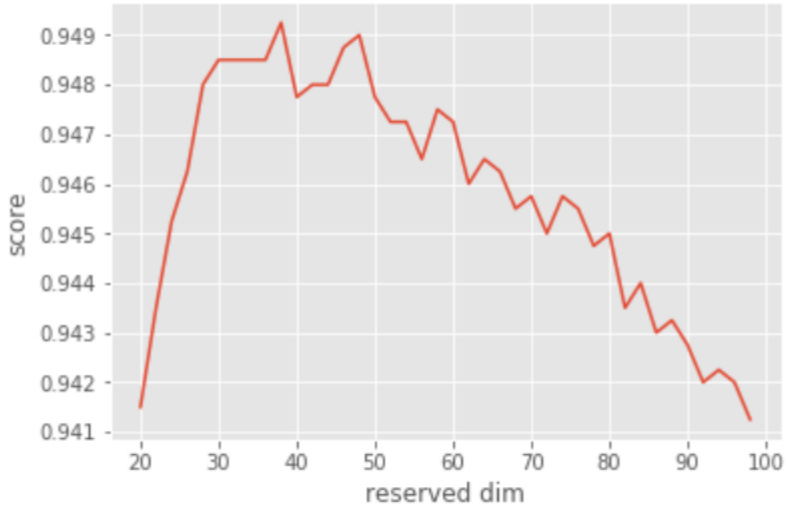
Wall time: 1.93 s

我们直接将数据使用 PCA 方法降至 30 维，之后再利用 SVM，速度和正确率都有了很大的提升。在下图中，我们看到 k = 30 时，分类器性能最佳。

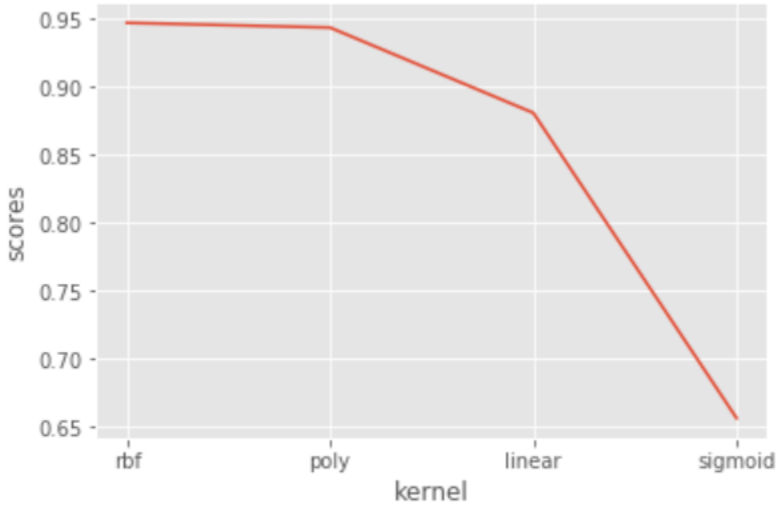


如果我们更加细化保留的维度 d，在（20-100）的区间，步长=2。我们可以看到分类的准确率会随着降维维度波动变换。这说明降维的过程具有一定的随机性。虽然总体来说一定存在一个较优的 k，能够在去除冗余信息和噪声的同时保留对于分类有价值的特征，但是这个过程是波动的。同时，我们测试如果想要保留样本 95%的信息，此时样

本的降维维数应该为 230-260，但是此时 PCA 的目的绝不是恢复原信息，而是最优化分类，加速分类过程。



我们选取 k = 30，探究不同的核函数的分类器性能。在下图中，我们看到性能最好的核是 rbf 核，这是符合预期的（一般来说 rbf 的性能最好）。



5.1.3 Naïve-Bayes

朴素贝叶斯分类器的想法是首先假设原始数据的分布类型（先验信

息），然后通过样本计算出数据特征，主要是方差和均值，然后根据类条件概率和先验概率得到后验概率，获得的最大概率所在的类别即为样本预测类别。

我们尝试了 Sklearn 中三类分类器 GaussianNB，MultinomialNB 和 BernoulliNB，先验概率初始为 1/m，m 为类别数。可以看到朴素贝叶斯的计算非常快，耗时极长的训练时间，但是效果也较有限。数据量的提升能够有效的提升朴素贝叶斯分类器的准确度。

训练数据：10000 测试：2000

```
GaussianNB score: 0.5165
GaussianNB time cost :(time 0.45s)
MultinomialNB score: 0.7735
GaussianNB time cost :(time 0.07s)
BernoulliNB score: 0.787
GaussianNB time cost :(time 0.18s)
```

训练数据：60000 测试：10000

```
GaussianNB score: 0.543
GaussianNB time cost :(time 2.69s)
MultinomialNB score: 0.8184
GaussianNB time cost :(time 0.49s)
BernoulliNB score: 0.8413
GaussianNB time cost :(time 0.83s)
```

5.1.4 Embedding

这个问题是非常好玩的。我们知道 MINST 已经算是 toy dataset，许多模式识别的分类方法都能够在 MINST 数据集上获得很好的效果，识别的准确率也能够达到 90%，甚至 99%的程度。但是如何在二维或者三维空间中理解这些二进制的图片数据在空间中分布情况呢？

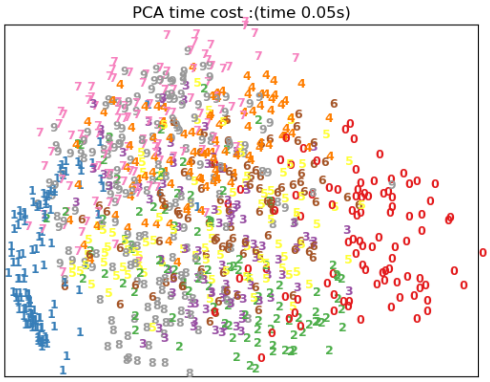
其实嵌入在我的理解中和降维差不多，是如何在利用低维度的数据表

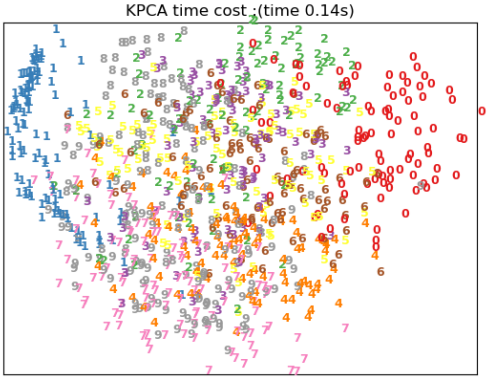
示高维度。而嵌入的好处在于如果能够让我们在二维或者三维的空间中看到数据样本是如何在一个高维空间中以一个低维流形存在的。

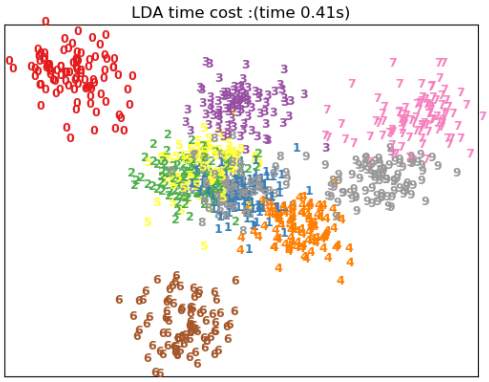
我们选取了 PCA 和 Isomap，LDA，Spectral，T-sne 这些方法来比较优劣。我们选取了 1000 个数据进行可视化显示。

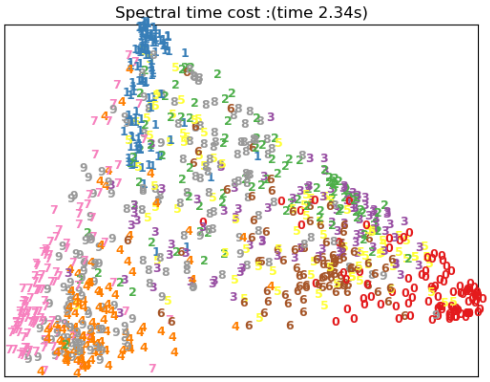
其实我们可以看到效果最好的是 LDA 和 T-SNE。

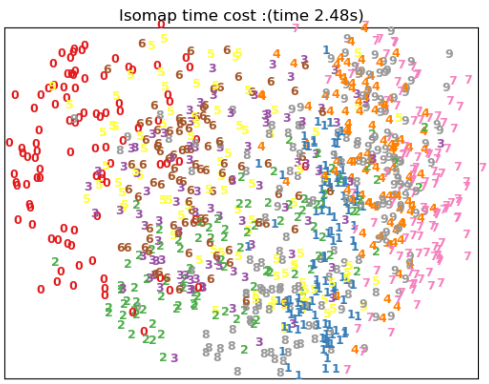
其中 LDA 和其他五种方式不同的是他还利用的 label 的信息，是在监督下获得的最优判别降维。而其他方法，Isomap 和 T-SNE 都是更为

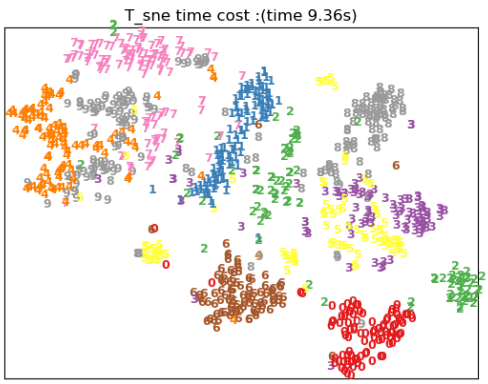












先进的方法，主要考虑了学习数据分布的局部特征。以 T-sne 为例，可以看到初值的选择会影响 embedding 的效果，一类数据点会被另一类数据点隔断。这需要合理的调整参数。

我们熟悉的 PCA,KPCA 和谱聚类虽然都有分类的比较好的地方，还是有很多样本混杂在一起。

5.2 实验总结

1、 我们在 MINST 数据集上实践了一些经典的模式识别方法，对于各种方法从模型准确度，计算时间，模型评价各个方面都有了了解，但是理解的还不算深刻。

2、 如何准确的理解数据的分布，同时根据样本的特点选取合适的方法是非常重要的。

3、 数据对于模式识别任务是重要的，但是不代表数据越多，模型就能学习的越好。对于朴素贝叶斯分类器，这确实成立，大数据量能够提升模型性能。但是对于 SVM，实验中发现，大数据量（60000 例）的模型准确度较小数据量（5000 例），模型的性能还有所下降。这对于其他模式识别方法也是同样的，需要对样本数据进行调研。

4、 相同的数据，其表示不同，得到的结果也不用。MINST 数据集可以使二进制文件格式，也可能是通过 0-255 图片形式存储。需要先明确样本的数组组成，整数，浮点数。

5、 对于如乳腺癌分类，鸢尾花，房价等含有多维度，多量纲的数据，归一化，标准化是重要的。