线性与非线性分类器求解MNIST分类问题

姓名：卢鸿浩

1. **问题一（线性分类器）**

**1.1问题描述**

图像分类问题：基于手写数字数据库 MNIST，应用最小二乘模型，对数字 0 和其他数字进行分类。MNIST 数据库来自美国国家标准与技术研究所，包含 70000张28×28的手写数字灰度图像。图像数据已经被转化为28 × 28 = 784 维的向量形式存储，例如，[0.0 0.0 ⋯ 0.0 0.380 0.376 0.301 0.462 ⋯ 0.239 ⋯ 0.0 0.0]，标签以 10维向量形式存储，例如，[0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0]。MNIST 数据库将数据分成 2 组，一组是包含 60000 张图像的训练集，另一组是包含 10000 张图像的测试集。将 𝒙 设置为 m+1 维的向量，第 1 维是常量 1，其余 m 维是至少在 600 张图片中像素值不为0的像素。如果图像为数字 0，则取 𝑦 = 1，否则取 𝑦 = −1。图1.0显示了部分数据集示例。数据库的下载网址为http://yann.lecun.com/exdb/mnist/。

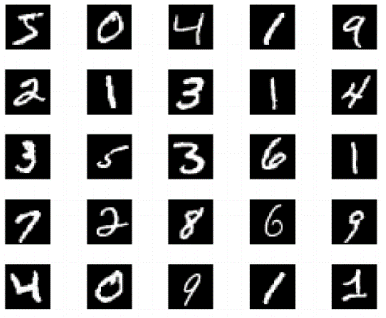


图1.0 来自MNIST数据集的25张手写数字图像

按照如下方法构造 5000个随机特征：生成随机矩阵 𝑹 ∈ ，𝑹𝑖𝑗随机地取值 ±1，取 max{0，(𝑹𝒙)𝑗}，𝑗 = 1， ⋯，5000，作为新特征，与原来的m+1个特征一起，构成5000+m+1个特征。

对于未加入5000个随机特征和加入5000个随机特征之后的分类效果分别给出了实验结果。实验结果表明：添加 5000 个随机特征以后，最小二乘分类精度显著提高

**1.2 未加入人工构造的随机特征**

取训练集的前600张图片，得出至少在 600 张图片中像素值不为0的像素的个数为587。那么特征个数为588。

如表1.1所示，给出了对标签为+1的分类错误率，对标签为-1的分类错误率，总的分类错误率。

学习到的参数存放在：

“p\_classifier1\_1\_11\_28\_18\_52.txt”

表1.1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **训练集** | | | **测试集** | | |
|  | **y=+1** | **y=-1** | **total** | **y=+1** | **y=-1** | **total** |
| **正确识别的个数** | 5162 | 53898 | 59060 | 864 | 8977 | 9843 |
| **总个数** | 5923 | 54077 | 60000 | 980 | 9020 | 10000 |
| **正确率（%）** | 87.15 | 99.67 | **98.43** | 88.16 | 99.52 | **98.41** |
| **错误率（%）** | 12.85 | 0.33 | **1.57** | 11.633 | 0.477 | **1.59** |

图1.1和图1.2是实验截图。

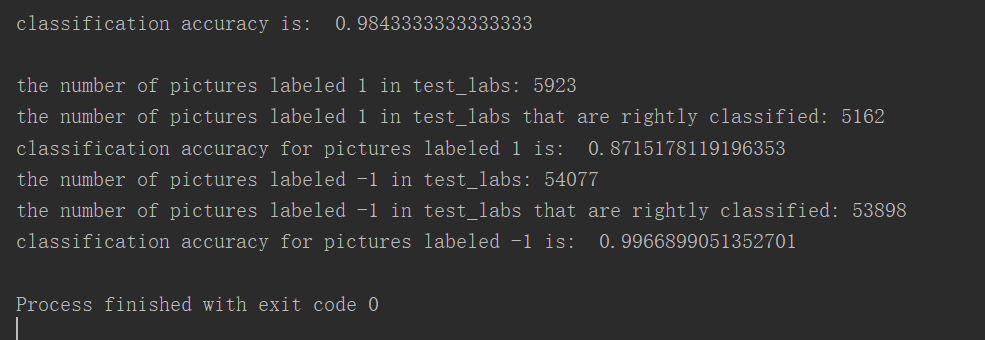


图1.1 线性分类器的训练集结果1

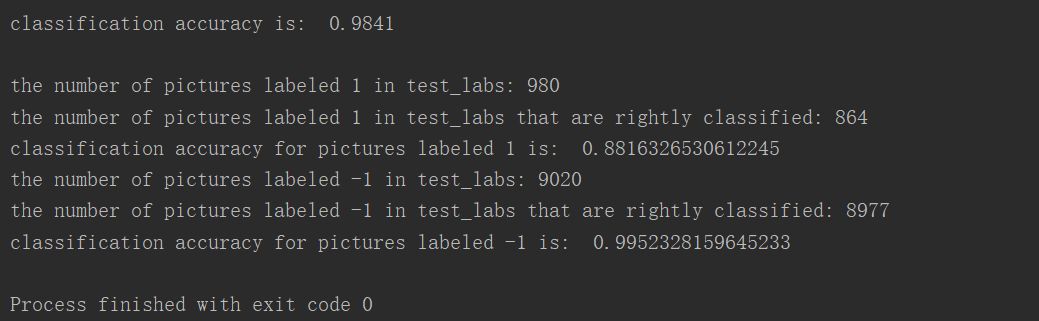


图1.2 线性分类器的测试集结果1

**1.2 加入人工构造的随机特征**

加入5000个随机特征后的训练结果如表1.2所示，相应的结果截图为图1.3和图1.4。

学习到的参数存放在：

“p\_classifier1\_2Matrix\_11\_28\_14\_59.txt”

“p\_classifier1\_2\_11\_28\_14\_59.txt”

表1.2

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **训练集** | | | **测试集** | | |
|  | **y=+1** | **y=-1** | **total** | **y=+1** | **y=-1** | **total** |
| **正确识别的个数** | 5819 | 54061 | 59875 | 967 | 9010 | 9972 |
| **总个数** | 5923 | 54077 | 60000 | 980 | 9020 | 10000 |
| **正确率（%）** | 98.28 | 99.97 | **99.80** | 98.67 | 99.88 | **99.77** |
| **错误率（%）** | 1.72 | 0.03 | **0.20** | 1.33 | 0.12 | **0.23** |

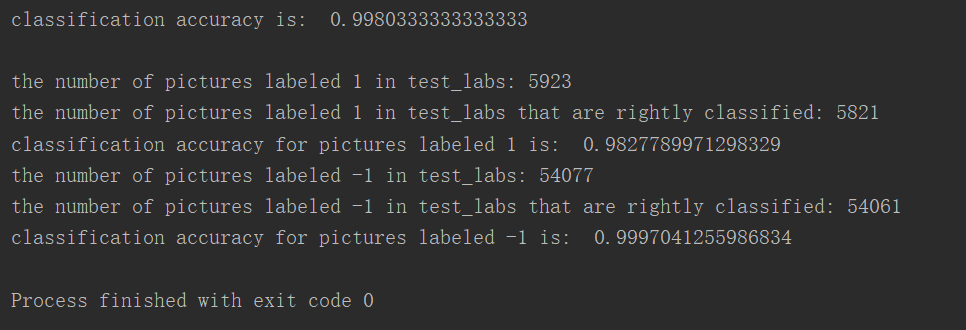


图1.3 线性分类器的训练集结果2

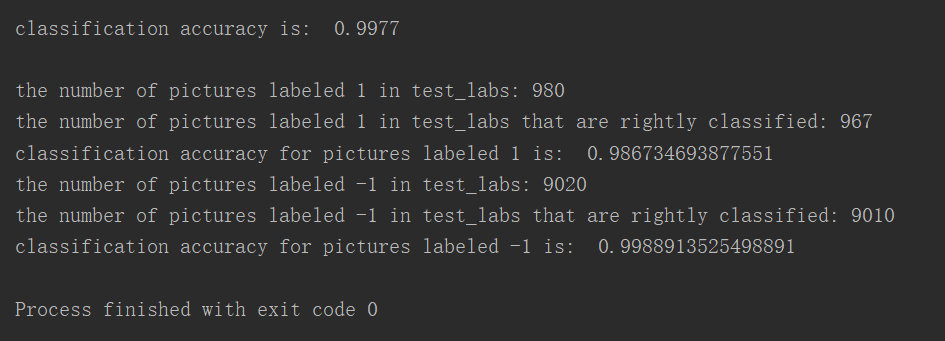


图1.4 线性分类器的测试集结果2

实验结果表明：添加 5000个随机特征以后，线性最小二乘分类的错误率由1.59%降低到0.23%，分类精度显著提高。

1. **问题二（非线性分类器）**

**1.1问题描述：**

利用非线性最小二乘分类求解的图像分类问题。将设置至少在600张图片中像素值不为0的像素,设长度为n，加上1维常量值，构成线性模型



通过求解如下非线性最小二乘问题确定参数和，



其中，𝜙 为sigmoid 函数，为正则化参数。

sigmoid 函数取为：



对于该非线性最小二乘，Jacobian矩阵是



其中，。

**1.2求解方法**

Levenberg-Marquardt算法

Levenberg-Marquardt算法是Gauss-Newton算法的延伸。在Gauss-Newton 算法的基础上，选择 𝑥(𝑘+1) 时，加入最小化的目标，使得最优解𝑥(𝑘+1)在 𝒙(𝑘)附近寻找，从而保证𝑓(𝒙; 𝒙(𝑘)) ≈ 𝑓(𝒙)。如果将正则化项加入LM算法中,则目标函数为：



可进一步转化为：



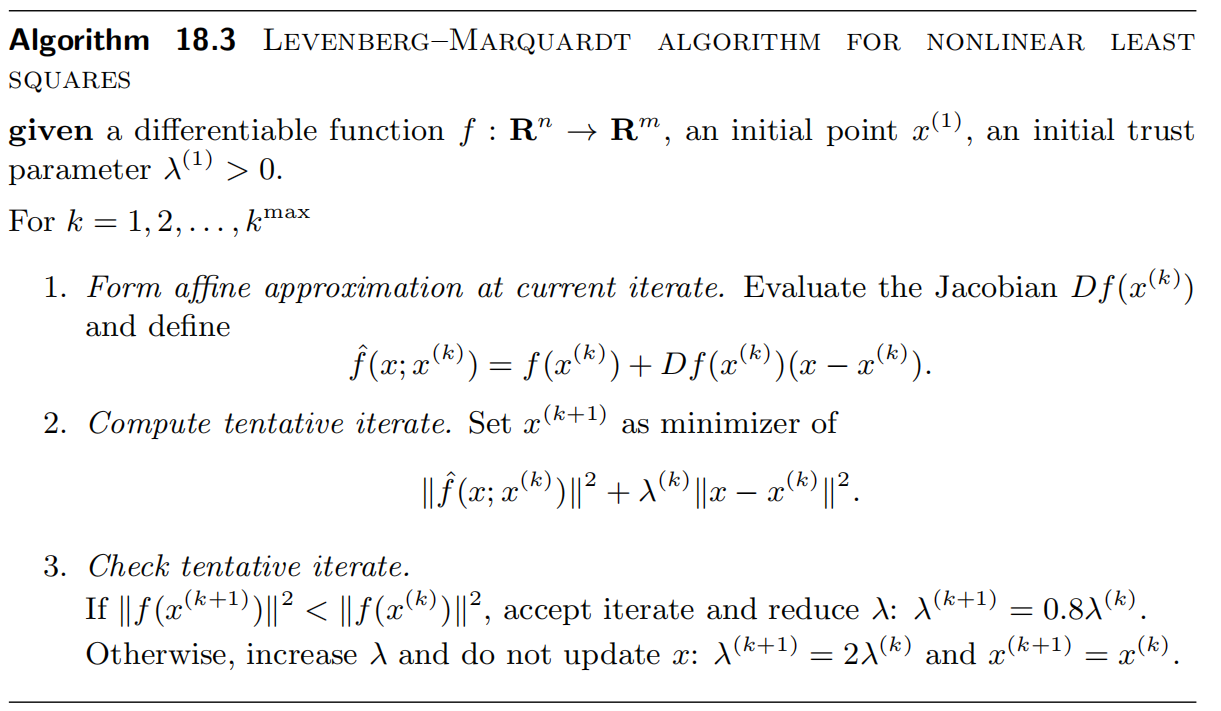
其中，代表,是阶的单位矩阵；代表阶的矩阵。

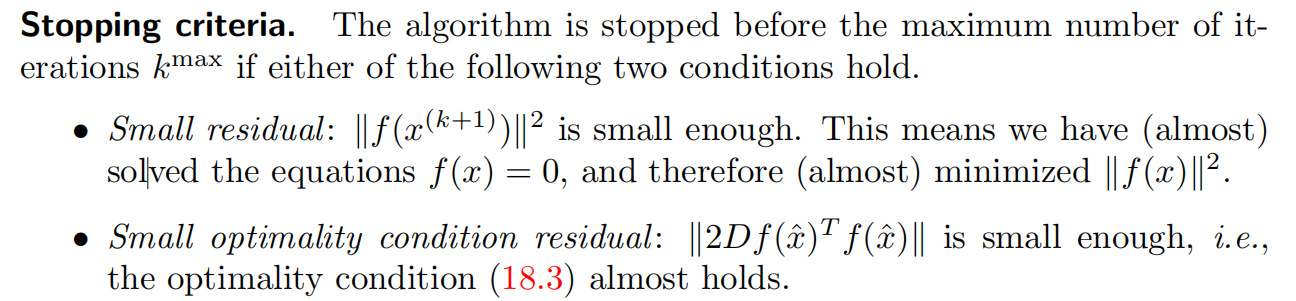
更新公式为：

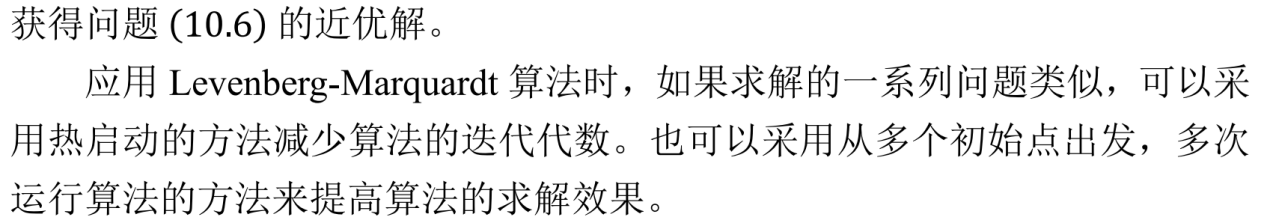


其中，代表。更新公式中，取0，就变成GN算法的更新公式，取0，则代表略去正则化项。

求解非线性最小二乘问题的Levnberg-Marquardt算法：







LM算法对初始点的选取要求比较严格，LM算法中，最优解只在初始点x(0)附近寻找，而如果初始点x(0)选取的不合适，会导致算法只能搜索到局部最优解，有可能与全局最优解差别很大。实验中，使用LM算法，随机产生初始点，多次实验的结果中，训练集的正确率从0.1~0.95之间波动。

实验中采用GN算法。

**1.2 未加入人工构造的随机特征**

表2.1是对未加入5000个随机特征时，训练结果较好的一次结果，图2.1 和图2.2是相应的训练集结果和测试集结果的截图。

学习到的参数存放在：

“p\_classifier2\_1\_11\_28\_19\_49.txt”

表2.1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **训练集** | | | **测试集** | | |
|  | **y=+1** | **y=-1** | **total** | **y=+1** | **y=-1** | **total** |
| **正确识别的个数** | 5819 | 47960 | 59875 | 855 | 7992 | 9972 |
| **总个数** | 5923 | 54077 | 60000 | 980 | 9020 | 10000 |
| **正确率（%）** | 87.57 | 88.69 | **88.58** | 87.24 | 88.60 | **88.47** |
| **错误率（%）** | 12.43 | 11.31 | **11.42** | 12.76 | 11.40 | **11.53** |

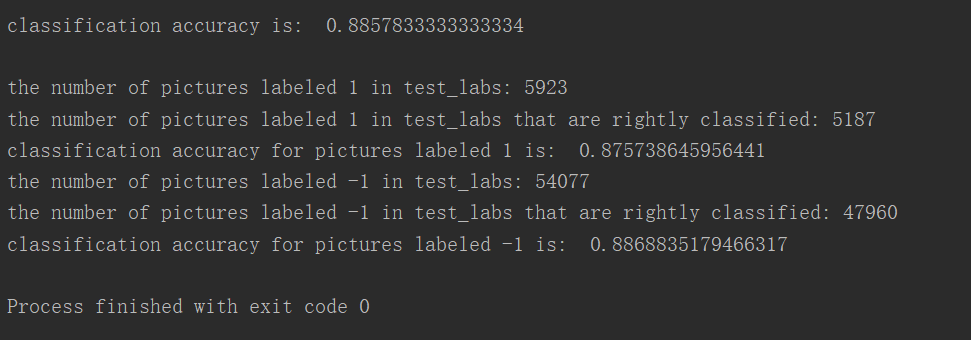


图2.1 非线性分类器的训练集结果1

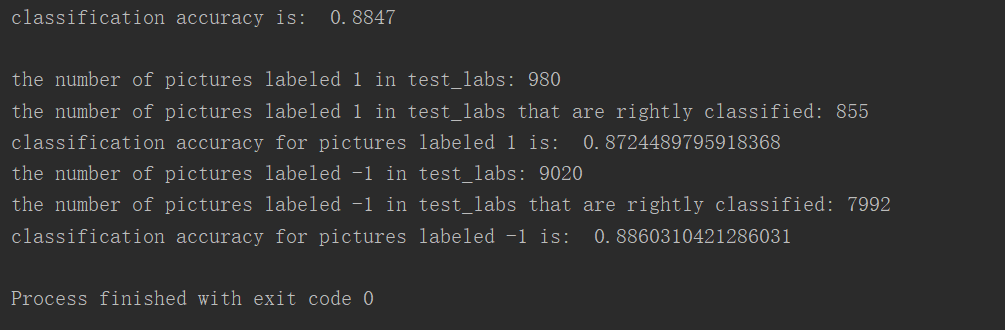


图2.2 非线性分类器的测试集结果1

**1.2 加入人工构造的随机特征**

因为对非线性分类器，如果加入过多特征，会需要大量的计算内存，且计算速度会下降，所以这里只取2000个构造的随机特征，构造方法同1.1。

表2.2是对加入2000个随机特征时，训练结果较好的一次结果，图2.3 和图2.4是相应的训练集结果和测试集结果的截图。

学习到的参数存放在：

“p\_classifier2\_2\_11\_29\_11\_22.txt”

“p\_classifier2\_2Matrix\_11\_29\_11\_22.txt”

表2.2

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **训练集** | | | **测试集** | | |
|  | **y=+1** | **y=-1** | **total** | **y=+1** | **y=-1** | **total** |
| **正确识别的个数** | 4147 | 34596 | 59875 | 678 | 5670 | 9972 |
| **总个数** | 5923 | 54077 | 60000 | 980 | 9020 | 10000 |
| **正确率（%）** | 70.02 | 63.98 | **64.57** | 69.18 | 62.86 | **63.48** |
| **错误率（%）** | 29.98 | 36.02 | **35.43** | 31.82 | 37.14 | **36.52** |

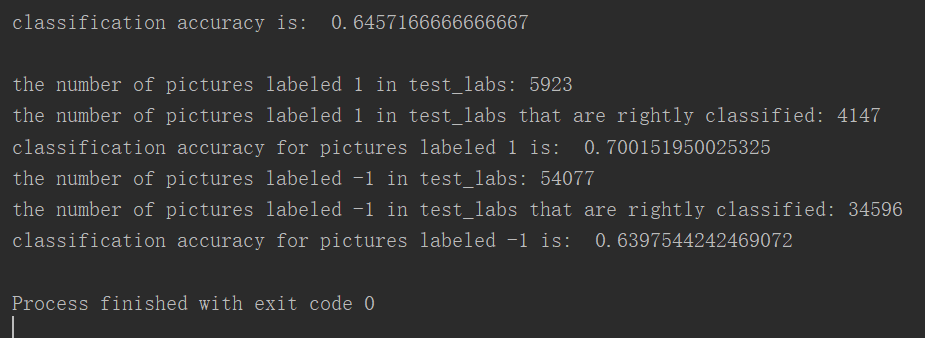


图2.3 非线性分类器的训练集结果2

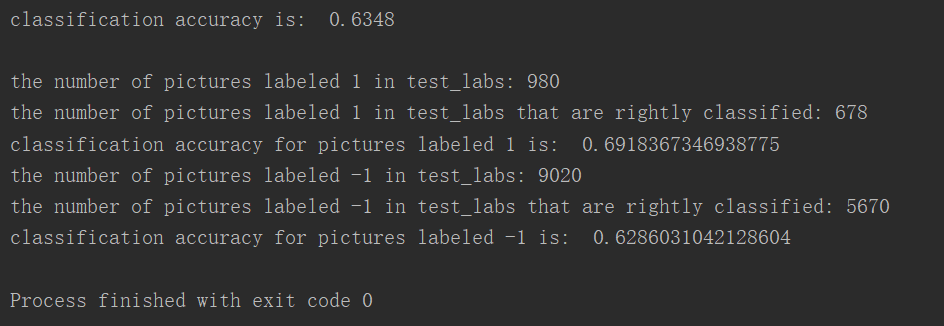


图2.4 非线性分类器的测试集结果2