

线性与非线性分类器求解 MNIST 分类 问题

姓名：卢鸿浩

一. 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 $28 \times 28 = 784$ 维的向量形式存储，例如， $[0.0 \ 0.0 \ \dots \ 0.0 \ 0.380 \ 0.376 \ 0.301 \ 0.462 \ \dots \ 0.239 \ \dots \ 0.0 \ 0.0]$ ，标签以 10 维向量形式存储，例如， $[0.0 \ 0.0 \ 0.0 \ 0.0 \ 0.0 \ 0.0 \ 0.0 \ 1.0 \ 0.0 \ 0.0 \ 0.0]$ 。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 \mathbf{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/>。

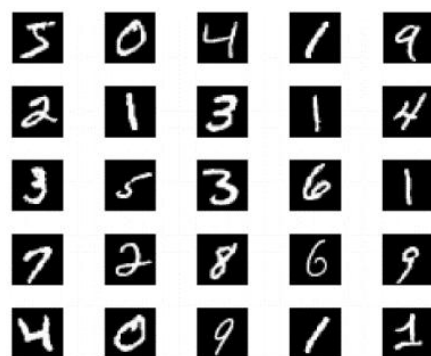


图 1.0 来自 MNIST 数据集的 25 张手写数字图像

按照如下方法构造 5000 个随机特征：生成随机矩阵 $\mathbf{R} \in R^{5000 \times m}$ ， R_{ij} 随机地取值 ± 1 ，取 $\max\{0, (\mathbf{R}\mathbf{x})_j\}$ ， $j = 1, \dots, 5000$ ，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。

```
classification accuracy is: 0.9843333333333333

the number of pictures labeled 1 in test_labs: 5923
the number of pictures labeled 1 in test_labs that are rightly classified: 5162
classification accuracy for pictures labeled 1 is: 0.8715178119196353
the number of pictures labeled -1 in test_labs: 54077
the number of pictures labeled -1 in test_labs that are rightly classified: 53898
classification accuracy for pictures labeled -1 is: 0.9966899051352701

Process finished with exit code 0
```

图 1.1 线性分类器的训练集结果 1

```
classification accuracy is: 0.9841

the number of pictures labeled 1 in test_labs: 980
the number of pictures labeled 1 in test_labs that are rightly classified: 864
classification accuracy for pictures labeled 1 is: 0.8816326530612245
the number of pictures labeled -1 in test_labs: 9020
the number of pictures labeled -1 in test_labs that are rightly classified: 8977
classification accuracy for pictures labeled -1 is: 0.9952328159645233

Process finished with exit code 0
```

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

```

classification accuracy is: 0.9980333333333333

the number of pictures labeled 1 in test_labs: 5923
the number of pictures labeled 1 in test_labs that are rightly classified: 5821
classification accuracy for pictures labeled 1 is: 0.9827789971298329
the number of pictures labeled -1 in test_labs: 54077
the number of pictures labeled -1 in test_labs that are rightly classified: 54061
classification accuracy for pictures labeled -1 is: 0.9997041255986834

Process finished with exit code 0

```

图 1.3 线性分类器的训练集结果 2

```

classification accuracy is: 0.9977

the number of pictures labeled 1 in test_labs: 980
the number of pictures labeled 1 in test_labs that are rightly classified: 967
classification accuracy for pictures labeled 1 is: 0.986734693877551
the number of pictures labeled -1 in test_labs: 9020
the number of pictures labeled -1 in test_labs that are rightly classified: 9010
classification accuracy for pictures labeled -1 is: 0.9988913525498891

Process finished with exit code 0

```

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。

二. 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 a 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 x_0 , 构成线性模型

$$\tilde{f}(a) = x_0 + a^T x_{1:n}$$

通过求解如下非线性最小二乘问题确定参数 x_0 和 $x_{1:n}$,

$$\min \sum_{i=1}^n \left[\phi \left(x_0 + \left(a^{(i)} \right)^T x_{1:n} \right) - y^{(i)} \right]^2 + \lambda_2 \|x_{1:n}\|^2$$

其中, ϕ 为 sigmoid 函数, λ_2 为正则化参数。

sigmoid 函数取为:

$$\phi(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}} = 1 - \frac{2}{e^{2u} + 1}$$

对于该非线性最小二乘, Jacobian 矩阵是

$$Df(\hat{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}_{x=\hat{x}} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} \left. \frac{d\phi(u)}{du} \right|_{u=u^{(1)}} \\ \vdots \\ \left. \frac{d\phi(u)}{du} \right|_{u=u^{(n)}} \end{bmatrix}$$

其中, $u^{(i)} = x_0 + \left(a^{(i)} \right)^T \hat{x}_{1:n}$ 。

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 $\|x(k+1) - x(k)\|^2$ 的目标, 使得最优解 $x(k+1)$ 在 $x(k)$ 附近寻找, 从而保证 $f(x; x(k)) \approx f(x)$ 。如果将正则化项加入 LM 算法中, 则目标函数为:

$$\min \left\| \hat{f}(x; x^{(k)}) \right\|^2 + \lambda_1 \|x - x^{(k)}\|^2 + \lambda_2 \|x_{2:n}\|^2$$

可进一步转化为:

$$\min \left\| \begin{bmatrix} Df(x^{(k)}) \\ \sqrt{\lambda_1} I \\ \sqrt{\lambda_2} I_0 \end{bmatrix} x - \begin{bmatrix} Df(x^{(k)})x^{(k)} - f(x^{(k)}) \\ \sqrt{\lambda_1} x^{(k)} \\ 0 \end{bmatrix} \right\|^2$$

其中, I_0 代表 $\begin{bmatrix} I_{(n-1) \times (n-1)} & 0 \\ 0 & 0 \end{bmatrix}$, $I_{(n-1) \times (n-1)}$ 是 $(n-1) \times (n-1)$ 阶的单位矩阵; I 代表 $n \times n$

阶的矩阵。

更新公式为:

$$x^{(k+1)} = x^{(k)} - \left((Df^{(k)})^{-T} Df^{(k)} + \lambda_1 I + \lambda_2 I_0 \right)^{-1} \left((Df^{(k)})^{-T} f^{(k)} + \lambda_2 I_0 x_k \right)$$

其中, $Df^{(k)}$ 代表 $Df(k)$ 。更新公式中, λ_1 取 0, 就变成 GN 算法的更新公式, λ_2 取 0, 则代表略去正则化项。

求解非线性最小二乘问题的 Levenberg-Marquardt 算法:

Algorithm 18.3 LEVENBERG-MARQUARDT ALGORITHM FOR NONLINEAR LEAST SQUARES

given a differentiable function $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$, an initial point $x^{(1)}$, an initial trust parameter $\lambda^{(1)} > 0$.

For $k = 1, 2, \dots, k^{\max}$

1. *Form affine approximation at current iterate.* Evaluate the Jacobian $Df(x^{(k)})$ and define

$$\hat{f}(x; x^{(k)}) = f(x^{(k)}) + Df(x^{(k)})(x - x^{(k)}).$$

2. *Compute tentative iterate.* Set $x^{(k+1)}$ as minimizer of

$$\|\hat{f}(x; x^{(k)})\|^2 + \lambda^{(k)} \|x - x^{(k)}\|^2.$$

3. *Check tentative iterate.*

If $\|f(x^{(k+1)})\|^2 < \|f(x^{(k)})\|^2$, accept iterate and reduce λ : $\lambda^{(k+1)} = 0.8\lambda^{(k)}$.

Otherwise, increase λ and do not update x : $\lambda^{(k+1)} = 2\lambda^{(k)}$ and $x^{(k+1)} = x^{(k)}$.

Stopping criteria. The algorithm is stopped before the maximum number of iterations k^{\max} if either of the following two conditions hold.

- *Small residual:* $\|f(x^{(k+1)})\|^2$ is small enough. This means we have (almost) solved the equations $f(x) = 0$, and therefore (almost) minimized $\|f(x)\|^2$.
- *Small optimality condition residual:* $\|2Df(\hat{x})^T f(\hat{x})\|$ is small enough, i.e., the optimality condition (18.3) almost holds.

获得问题 (10.6) 的近优解。

应用 Levenberg-Marquardt 算法时，如果求解的一系列问题类似，可以采用热启动的方法减少算法的迭代代数。也可以采用从多个初始点出发，多次运行算法的方法来提高算法的求解效果。

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。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_05_21_20_09.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	5781	54017	59875	944	8967	9911
总个数	5923	59798	60000	980	9020	10000
正确率 (%)	97.60	99.89	99.66	96.32	99.41	99.11
错误率 (%)	2.4	0.11	0.34	3.68	0.59	0.89

```
classification accuracy is: 0.9966333333333334

the number of pictures labeled 1 in test_labs: 5923
the number of pictures labeled 1 in test_labs that are rightly classified: 5781
classification accuracy for pictures labeled 1 is: 0.9760256626709438
the number of pictures labeled -1 in test_labs: 54077
the number of pictures labeled -1 in test_labs that are rightly classified: 54017
classification accuracy for pictures labeled -1 is: 0.9988904709950626
```

图 2.1 非线性分类器的训练集结果 1


```

classification accuracy is: 0.9911

the number of pictures labeled 1 in test_labs: 980
the number of pictures labeled 1 in test_labs that are rightly classified: 944
classification accuracy for pictures labeled 1 is: 0.963265306122449
the number of pictures labeled -1 in test_labs: 9020
the number of pictures labeled -1 in test_labs that are rightly classified: 8967
classification accuracy for pictures labeled -1 is: 0.9941241685144124

Process finished with exit code 0

```

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_05_21_20_30.txt”

“p_classifier2_2Matrix_05_21_20_30.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	5785	54018	59803	944	8967	9911
总个数	5923	54077	60000	980	9020	10000
正确率 (%)	97.67	99.89	99.67	96.32	99.41	99.11
错误率 (%)	2.33	0.11	0.33	3.68	0.59	0.89

```

classification accuracy is: 0.9967166666666667

the number of pictures labeled 1 in test_labs: 5923
the number of pictures labeled 1 in test_labs that are rightly classified: 5785
classification accuracy for pictures labeled 1 is: 0.9767009961168327
the number of pictures labeled -1 in test_labs: 54077
the number of pictures labeled -1 in test_labs that are rightly classified: 54018
classification accuracy for pictures labeled -1 is: 0.9989089631451449

```

图 2.3 非线性分类器的训练集结果 2


```
classification accuracy is: 0.9911

the number of pictures labeled 1 in test_labs: 980
the number of pictures labeled 1 in test_labs that are rightly classified: 944
classification accuracy for pictures labeled 1 is: 0.963265306122449
the number of pictures labeled -1 in test_labs: 9020
the number of pictures labeled -1 in test_labs that are rightly classified: 8967
classification accuracy for pictures labeled -1 is: 0.9941241685144124
```

图 2.4 非线性分类器的测试集结果 2