

分类器实现汇总作业

问题

Based on the MNIST dataset, design and implement classifiers including: least squares with regularization, Fisher discriminant analysis (with kernels), Perceptron (with kernels), logistic regression, SVM (with kernels), MLP-NN with two different error functions.

数据集

MNIST 数据集由 60000 张训练图片和 10000 张测试图片组成, 为手写 0-9 的灰度图像, 分辨率 28*28。由于需要控制运行时间, 因此将训练和测试集都缩小到 1000。

one-vs-one 和 one-vs-all

由于很多算法只能进行二类划分, 在进行多类划分时有两种策略。one-vs-one 训练 $n*(n-1)/2$ 个分类器, 对于一个元素, 在每个分类器中都能得到一个类别, 最后将投票次数最多的类别作为其最终类别。one-vs-all 训练 n 个分类器, 用于分类某个元素和其他所有元素。如果一个元素在某个分类器被判定为真那么就认为这个元素属于该类。在 one-vs-one 中, 由于元素会在很多和自己类别无关的分类器中判定, 在 one-vs-all 中则是可能在多个分类器中被认为属于一类; 或是被所有分类器拒绝, 因此两种方法都各有优劣。

least squares with regularization

最小二乘法使用正则化, 采用了 L2 正则化, 即岭回归。使用 sklearn 中的岭回归函数。同时使用最小二乘法找到平面仅能做二类划分, 测试后使用 one-vs-one 方法来做多类划分。最终分类准确率 77.8%

Kernel LDA

Kernel LDA 是在普通的 LDA 上运用核方法, 使得 LDA 能够处理非线性问题。使用核方法后, 对于每次询问都需要和训练数据集中每个元素进行核函数计算, 因此相比传统 LDA 测试开销很大。核函数采用了 RBF 核。测试后使用 one-vs-one 的方法做多类划分。最终分类准确率 29.1%

Kernel Perceptron

核函数感知器使用感知器的对偶问题，这样使用核方法不会带来不同的时间复杂度。同样使用 RBF 核，由于训练速度过慢，因此采用训练分类器较少的 one-vs-all 方法。最终分类准确率 72.6%

Logistic Regression

Logistic Regression 和最小二乘法相比由于使用了 Logistic 函数使得距离分类平面很远，也就是分得很对的点对分类平面的影响减小，因此可以取得比最小二乘法更好的效果。使用 sklearn 中的 LogisticRegression，最终分类准确率 81.9%

SVM

使用支持向量机来进行分类，使用 sklearn 中的 SVM，并采用默认参数。由于直接使用原始数据无法运作，将输入的灰度图像根据阈值 128 转换为 01 黑白图像作为输入。最终分类准确率 78.3%

MLP

使用多层感知器，即神经网络进行分类。将图片拉成一条向量作为输入，隐层有 80 个神经元，输出 10 个神经元。三层之间采用全连接，并使用 ReLU 作为中间层的激活函数。输出 10 维表示 10 个类别。训练的 batch size 取 100，并进行 10000 个 epoch 的训练。

首先采用交叉熵函数作为损失函数，最终分类准确率 83.3%

然后采用 KLDivLoss，需要在输出结果上加上一层 softmax。最终分类准确率 84.1%

代码说明

代码使用 Python3 编写，Kernel LDA 代码使用了 <https://github.com/bhaktipriya/Kernel-PCA-and-LDA/blob/master/klda.py> 的代码。Ridge, Logistic Regression, SVM 使用 sklearn, MLP 使用 pytorch, Kernel Perceptron, one-vs-one, one-vs-all, 数据处理, RBF 自行编写。

代码共有 main.py 和 klda.py 两个文件，执行 python3 ./main.py 即可。

文件说明

共有如下文件/文件夹：

mnist: mnist 数据集文件

main.py/klda.py: 代码文件。

output.txt: main.py 执行的输出。

report.docx/report.pdf: 报告文档及其 pdf。