

作业 3

2022 年 3 月

1 支持向量机（SVM）

使用 SVM 解决一个图像的二类别分类问题。数据集是从 Fashion-MNIST 数据集中采样出的两个类别（T 恤和裤子）构成训练集和测试集。

数据集信息

训练集包含 12000 张图片，存储在 `X_train_sampled.npy` 文件中，每张图片的大小为 28×28 ；训练样本的标签存储在 `y_train_sampled.npy` 文件中。测试集包含 2000 张图片，存储在 `X_test_sampled.npy` 文件中，每张图片的大小为 28×28 ；测试样本的标签存储在 `y_test_sampled.npy` 文件中。

实验步骤

步骤 1. 提取图像的 Histogram Of Gradient (HoG) 特征（见讲座 2 的第 14 页）。对于一个样本 $x \in \mathbb{R}^{28 \times 28}$ ，提取到的 HoG 特征向量表示为 $h_x \in \mathbb{R}^{784}$ 。

提取 HoG 特征的代码见 `HoG.py`，特征提取的详细过程已在代码中注释。

步骤 2. 利用提取到 HoG 特征向量 h_x ，尝试使用不同的 SVM 分类器进行分类（你可以使用 scikit-learn 库实现 SVM，参考文档请见链接：

<https://scikit-learn.org/stable/modules/svm.html>，或者文档

Support Vector Machines—scikit-learn 1.0.pdf）。你需要实现三种 SVM (with outliers, 见讲座 2 的第 13 页) 分类器：

(1) 线性 SVM ；

(2) RBF SVM；

(3) 其他任选一种核函数的 SVM，比如多项式 SVM。你需要为核函数中的各个参数找出合适的值，例如合适的 C 的值 (ξ 的系数)。SVM 库

函数里面的 C 变量即为 ξ 的系数 (ξ 相关的公式见 lecture 2 的 13 页)。

请汇报以下结果：

1. SVM 在测试集上的分类准确率。
2. 对于 Linear SVM, 请汇报参与参数 w 的计算的支持向量有哪些? (支持向量的定义见 lecture 2 的第 7 页)
 - 2.1 一共有几个支持向量参与了参数 w 的计算?
 - 2.2 在参与了参数 w 在计算的支持向量中, 有几个正样本? 有几个负样本? 样本? 你需要把这些支持向量可视化出来, 将支持向量可视化的图片粘贴在报告中, 并附上各个支持向量所对应的权重 (即 $y_i * \alpha_i$ 的值)。

• 注意事项

1. 此次作业只会大致比较提交上来的 SVM 分类器的性能 (测试集上的分类准确率), 性能较好的会获得更高的分数。
2. 只要 SVM 分类器的性能大致在一个量级即可, 不会严格去比较谁性能高半个百分点。性能高半个百分点。
3. 提交代码备份, 对于性能异常优越(或异常差)的代码, 可能会查代码。