多分类线性支持向量机

杨朝辉　PB17071433

2020-04-14

算法简介

SVM 二分类分类器问题

引入松弛变量ξ_i后的软间隔最大化线性支持向量机求解的原优化问题

$$\min_{w,b,\xi} \ \frac{1}{2}\|w\|^2+C\sum_{i}^{N}\xi_i$$
$$s.t. \ \ y_i(w\cdot x_i+b)\geq 1-\xi_i, \ \ \xi\geq 0, \ \ i=1,2,\cdots,N$$

是一个受约束凸优化问题，求得参数解w^*, b^*，得到分类决策函数

$$f(x)=\text{sign}(w^*\cdot x+b^*)$$

对其求解可以转化为对其如下对偶问题的求解

$$\min_{\alpha} \ \frac{1}{2}\sum_{i,j}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{N}\alpha_i$$
$$s.t. \ \sum_{i=1}^{N}\alpha_i y_i=0, \ 0\leq\alpha_i\leq C, \ i=1,2,\cdots,N$$

则$w^*=\sum_{i}^{N}\alpha_i^* y_i x_i$，$b^*=y_j-w^*\cdot x_j$，其中对于下表 j 的$x_j$为支持向量（相应的$0<\alpha_j<C$）。

序列最小最优化算法（SMO）求解参数

为方便将算法中$x_i\cdot x_j$简记为$K(x_i,x_j)=K_{ij}$（当引入核函数时可以有$K_{ij}\neq x_i\cdot x_j$，不过本实验中数据属性列较多，未使用核技巧）所替代。以下对偶问题

$$\min_{\alpha} \frac{1}{2}\sum_{i,j}^{N}\alpha_i\alpha_j y_i y_j K_{ij} - \sum_{i=1}^{N}\alpha_i$$

$$s.t.\ \sum_{i=1}^{N}\alpha_i y_i = 0,\ 0 \le \alpha_i \le C,\ i = 1,2,\cdots,N$$

中由于样本数N通常很大，对N个α_i值的求解采用 SMO 算法，即将原问题分为子问题进行求解，具体包括求两个变量的二次规划方法和选择变量的启发式方法，具体描述如下：

（1） 设置迭代次数k，初始值$\{\alpha_i\} = \{0\}$，$b = 0$，每次迭代中进行以下运算

（2） 检验各个样本点及对应α_i是否满足如下 KKT 条件

$$\alpha_i = 0 \Leftrightarrow y_i g(x_i) \ge 1$$
$$0 < \alpha_i < C \Leftrightarrow y_i g(x_i) = 1$$
$$\alpha_i = C \Leftarrow y_i g(x_i) \le 1$$

并选取不满足 KKT 条件最严重的点，相应参数作为第一个变量α_1，而后随机选取与 i 不同的下标 j，相应参数作为α_2

（3） 根据相应解析过程对参数α_1, α_2进行优化，并修正b

（4） 直至经过k轮的迭代，对偶问题的参数近似求解完成

二分类拓展为多分类

二分类学习器拓展到多分类通常使用 OVO,OVR 等策略,此处考虑到 OVR 方法会引入正、负例分布不均造成的所学的模型的预测偏差，同时所用数据的标签类别数（类别数n为 6）并不算大,故采用 OVO 策略将二分类拓展到多分类,即获得$C_n^2 = n(n-1)/2$个分类超平面，并借助他们对未知数据集进行预测。

借助$C_n^2 = n(n-1)/2$分类函数对模型预测时，分别在每一个样本上进行$C_n^2 = n(n-1)/2$次的函数计算，并根据预测值采用多数表决的方法确定其最终所属类别。

实验条件

采用 python 语言及其相关的 numpy，sklearn，matplotlib 等支持，方便数据的预处

理、数值计算及混淆矩阵的可视化。构造分类器 SVF 对象，对数据训练集拟合，并通过在训练集上的交叉验证和参数调节，最终部署并于测试集中展现模型的分类效果。

数据简介

采用面部表情识别数据，共计 1002 个样本，标签为离散值 {1，2，3，4，5，6 }，每个样本有 25 个属性值 {f1，f2， ... f24， age}。进行训练之前借助 sklearn 库中的 StandScaler 和 train_test_split 方法对数据做初始的标准归一化和训练集:测试集以 7:3 比例的划分。

评估参数

该分类器将 5 折交叉验证所得的混淆矩阵和分类准确率作为评估参数，例如如下为正则化参数 C 选取为 100 时的真实值和分类预测混淆矩阵以及准确率。

Confusion Matrix:

[[83 9 3 2 10 8]

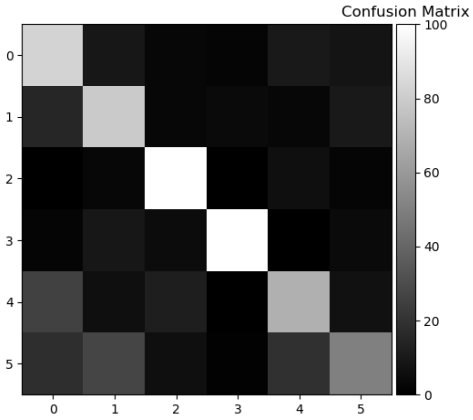
 [15 79 3 4 3 10]

 [0 3 100 0 6 2]

 [2 9 5 100 0 4]

 [26 6 12 0 69 7]

 [18 27 6 1 19 50]]



Accuracy:

[0.68794326, 0.72857143, 0.70714286, 0.70714286, 0.66428571]

交叉验证方法

在训练集采用 5 折交叉验证方法。具体来说，由于自定义的支持向量机分类器 SVF 继承了 sklearn 中的 BaseEstimator，可以方便地调用 sklearn 中的 cross_val_score 和 cross_val_predit 进行交叉验证。

结果与分析

测试集结果

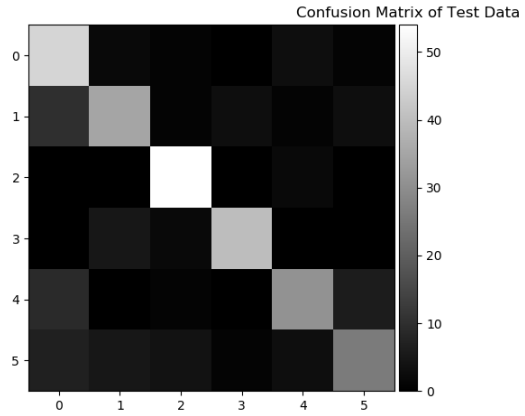
在原始数据预留的 30%比例的测试集上进行了分类预测，得到混淆矩阵如下，

以准确率$\eta = \sum_i I(f(x_i) = y_i)/N$来表征模型的泛化能力，$N$ 为测试集样　本数，得到

$$\eta = 0.77$$

混淆矩阵：

```
[[45  2  1  0  3  1]
 [10 35  1  3  1  3]
 [ 0  0 54  0  2  0]
 [ 0  5  2 40  0  0]
 [ 9  0  1  0 31  6]
 [ 7  5  4  1  3 26]]
```



根据展示结果发现模型在测试集上的表现明显好于训练集上交叉验证的表现，说明学习得到的模型在该训练集叫下时的泛化能力仍比较强。

模型限制

该表情分类数据的规模较小，样本数仅为 1000 左右，而类别数未 6，则每一类的样本数只有几百，这还远不能达到规模化机器学习的需求，模型交叉验证上的准确率也只有 0.7 左右。受限于数据集大小，事实上，即便使用 sklearn 内置 LinearSVC 对数据进行分析，经过交叉验证在训练集上也只得到 0.78左右的准确率。