Neural network theory and Applications

Homework Assignment 1

汪旭鸿

017032910027

March 26, 2018

Problem 1

在任务 1 中，我采用了 one-vs-reset 策略，利用二分类的 svm 分类器，解决一个三类的表情分类问题。其中，one-vs-reset 策略是自己编写的，svm 算法使用的是机器学习库 sklearn 中已经封装好的函数，**为了节省运算时间我采用了线性核**，并没有尝试 RBF 核和多项式核，我使用的底层 svm 库是 liblinear 中的 linearsvm。

1. one-vs-reset 策略：

(1) 标签分割：

在这个表情分类任务中，共有三类标签｛-1,0,1｝。现在我将训练集标签 train_label 数据复制成三组，每组分别进行处理，。

第一组，将标签为｛-1｝的样本标签置为 0，将其他样本标签置为 1

第二组，将标签为｛1｝的样本标签置为 0，将其他样本标签置为 1

第三组，将标签为｛0｝的样本标签置为 0，将其他样本标签置为 1

这样就实现了 one-vs-reset 策略的标签分割，形成了三组数据集。具体的实现方式是 np.where(train_label == 1, 0, 1)，将 nparray 数据类型，满足” train_label == 1”条件的部分置 0，其他部分置 1.

(2) 模型训练

建立三个 SVM 分类器，分别对三组数据集进行交叉验证训练，得到三个较优的分类器 SVM_1,SVM_0,SVM_{-1}

(3) 测试&标签整合

获得测试集样本 test_data[i]，分别输入三个分类器，得到每个分类器输出 0 的置信率。如果模型 SVM_1 输出的置信率最大，那么将测试集的预测 test_pred[i]置为 1；如果模型 SVM_0 输出的置信率最大，那么将测试集的预测 test_pred[i]置为 0；如果模型 SVM_{-1} 输出的置信率最大，那么将测试集的预测 test_pred[i]置为-1。i=13588，即这个过程执行 13588 次。

(4) 模型评估

分别计算训练集和测试集的 accuracy、precision、recall、f1-score，全面评估模型的性能。

2. 实验结果分析：

(1) 交叉验证

参数 C=$[2\times10^{-8},5\times10^{-8},7\times10^{-8},1\times10^{-7},3\times10^{-7},5\times10^{-7},7\times10^{-7},1\times10^{-6}]$

如图 1，linearsvm 分类器在 one-vs-reset 的策略下，惩罚系数在10^{-7}左右时，模型的准确率较高(达到 60%左右)，方差较小，同时具有较小的过拟合与欠拟合倾向。

当惩罚系数 C 小于6×10^{-8}时，模型的准确率呈现极具地下降趋势，具有非常严重的欠拟合倾向；

当惩罚系数 C 大于7×10^{-7}时，训练集的准确率越来越高，测试集的准确率却不再上升，训练与测试的准确率曲线呈现发散趋势，具有过拟合的特征。

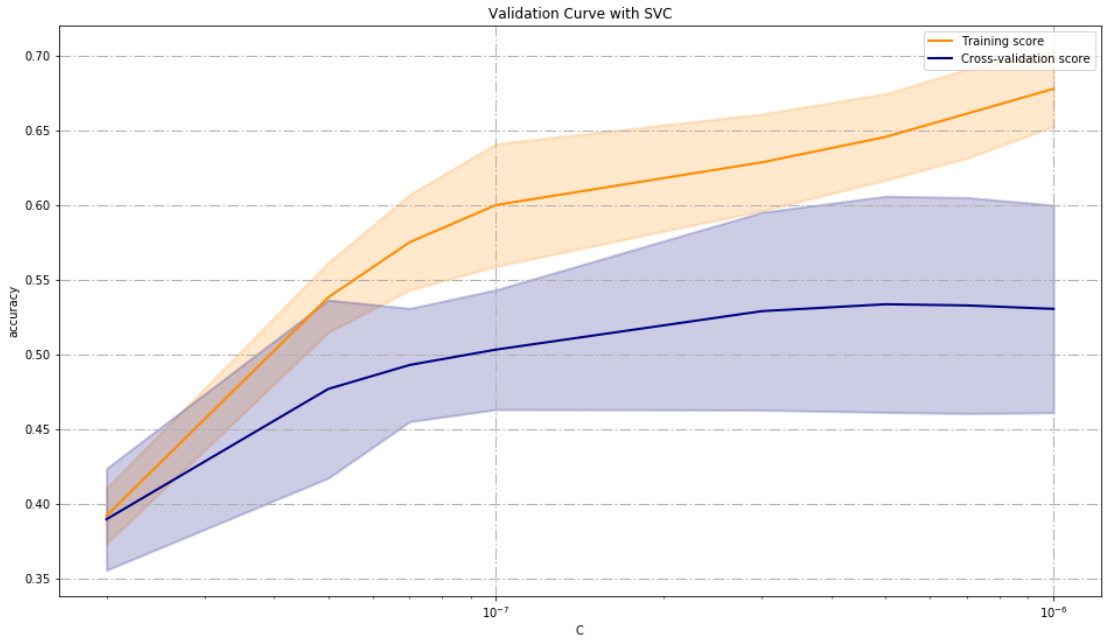

图 1 交叉验证曲线，纵坐标为 accuracy，横坐标为惩罚系数 C 的线性 SVM

(2) 参数选优:

根据交叉验证曲线，设定惩罚系数为2×10^{-7}，该 linearSVC 分类器最后的表现是，在训练集上取得了 58.18%的准确率，在测试集上取得了 60.30%的准确率。

(3) 分类报告：

Train	precision	recall	f1-score	support
class -1	0.63	0.47	0.54	12320
class 0	0.54	0.59	0.55	12144
class 1	0.64	0.74	0.69	12903
avg / total	0.61	0.60	0.60	37367

Test	precision	recall	f1-score	support
class -1	0.68	0.14	0.23	4480
class 0	0.54	0.74	0.62	4416
class 1	0.61	0.86	0.71	4692
avg / total	0.61	0.58	0.52	13588

(4) ROC 曲线

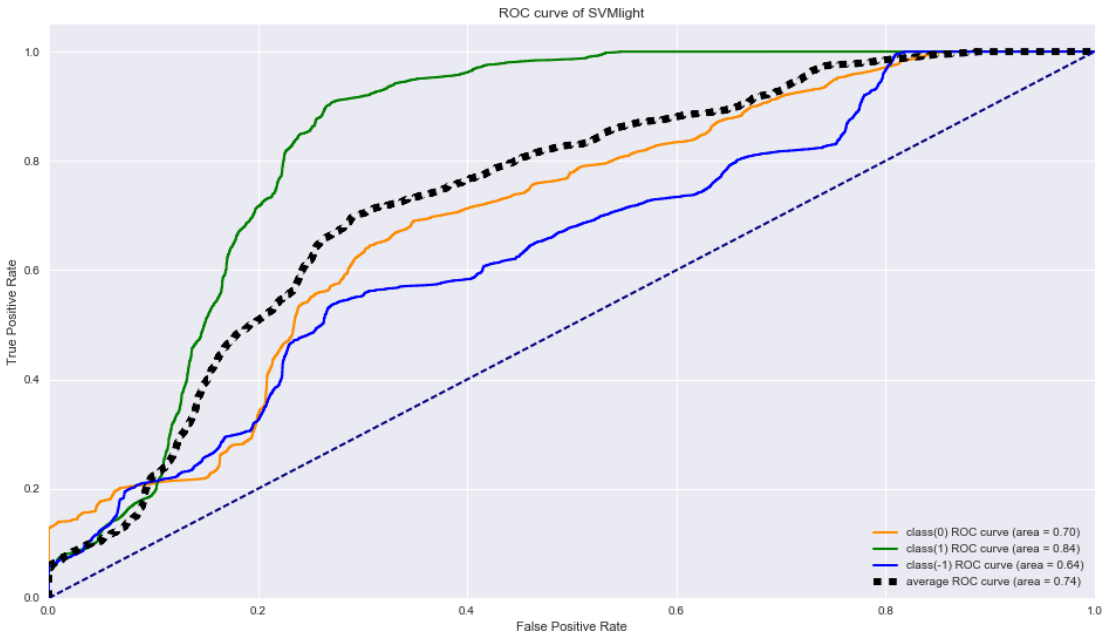

图 2 线性 SVM ROC 曲线 平均 ROC area=0.74

根据图 2 可知，这三类的平均 ROC area 为 0.74。其中，黄色线是 class(0)，绿色线是 class(1)，蓝色线是 class(-1)的 ROC 曲线，可以看出，本文训练的基于 one-vs-reset 策略的 SVM 分类器对 class(1)的数据识别结果较突出，对 class(-1)的识别结果很差，这一结论从分类报告中也可以得出。

Problem 2

在任务 1 中，我采用了 one-vs-reset 的策略，将一个三类问题分为了 3 个两类问题，但是这 3 个两类问题都存在一个样本不均衡的问题，负样本数量大概是正样本数量的两倍。为了解决这个样本不平衡问题，我采用了如下的 Min-max-modular 策略。

1. 基于先验知识的 class based min-max-modular 策略

在任务 2 中，我将采用了基于先验知识的 Min-max-modular 策略，分析每个二分类问题，发现其中的负样本数量大约是正样本的两倍。

(1) 训练集样本分割过程

为了解决样本不平衡问题，将其中的正样本分为两份，将其中的负样本分为四份，将这四份样本两两组合，生成 8 份正负混合样本，分别记为$N_{ij},\{i \in 1,2\ , j \in 1,2,3,4\ \}$，i 表示正样本，j 表示负样本。注意，在具体编程实现时，仅仅是对样本的索引号进行了操作，并没有对样本本身进行操作。

(2) 训练过程

针对每份训练样本N_{ij}，分别训练一个 linearSVM 分类器，共 8 个二分类器，记为M_{ij}。由于在任务 1 种调试过 linearSVM 的超参数——惩罚系数 C，所以在这个训练过程中，不需要再进行超参数调节了。这样就完成了训练过程，接着进行测试过程。

(3) 测试过程

如图 3，在测试过程中，将某个测试样本并行输入这 8 个 linearSVM 分类器，得到 8 个不同的 softmax 概率P_{ij}，将这 8 个分类器的概率输出按照 min-max 的规则

$$P = max(min\sum_{j=1}^{4} P_{1j}\ , min\sum_{j=1}^{4} P_{2j})$$

得到一个最终的概率输出 P，P的意义就是该测试样本属于正样本的概率。

(4) 整合过程

min-max-modular 策略解决的是二分类问题。而我们面临的是三类问题，当我们使用 min-max-modular 策略分别解决了三个二分类问题之后，将结果使用 one-vs-reset 策略进行整合，就可以获得这个三分类问题的最后输出结果。

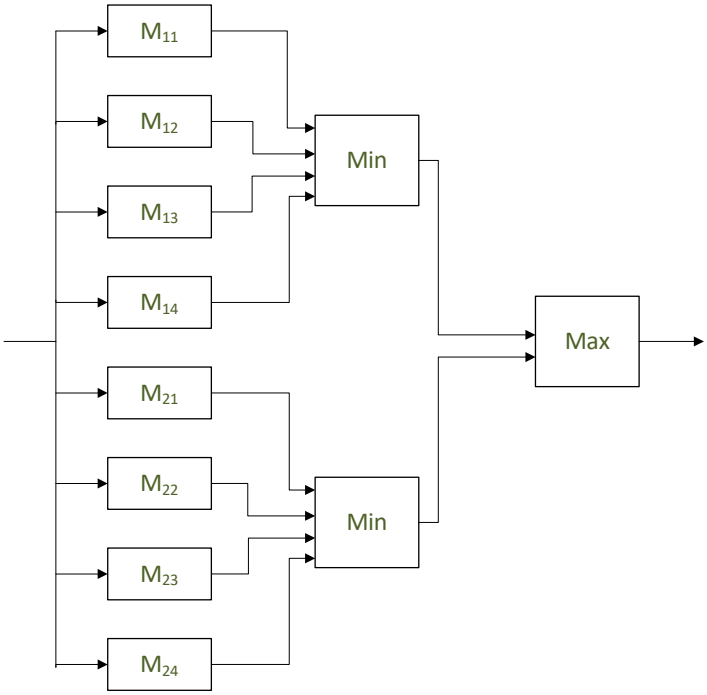

图 3 min-max-modular 结构图

2. Random min-max-modular 策略

Random min-max-modular 策略和基于先验知识的策略的**不同点在于对样本的分割方式**，基于先验知识的样本分割方式是利用了已知的样本标签信息，将样本分为正负标签均衡的 8 份；而随机的 min-max-modular 策略仅仅只是将样本随机分为 8 分，没有考虑样本不平衡的问题，根据经验来说，随机策略的 min-max-modular 策略的实现效果应该会弱于基于先验知识的 min-max-modular 策略。下面是原始 SVM、基于先验知识 M3SVM 和随机 M3SVM 的结果比较。

3. 实验结果分析

在本次的实验结果采用 ROC 曲线来衡量，ROC 曲线围成的面试越大，代表模型的性能越优越。

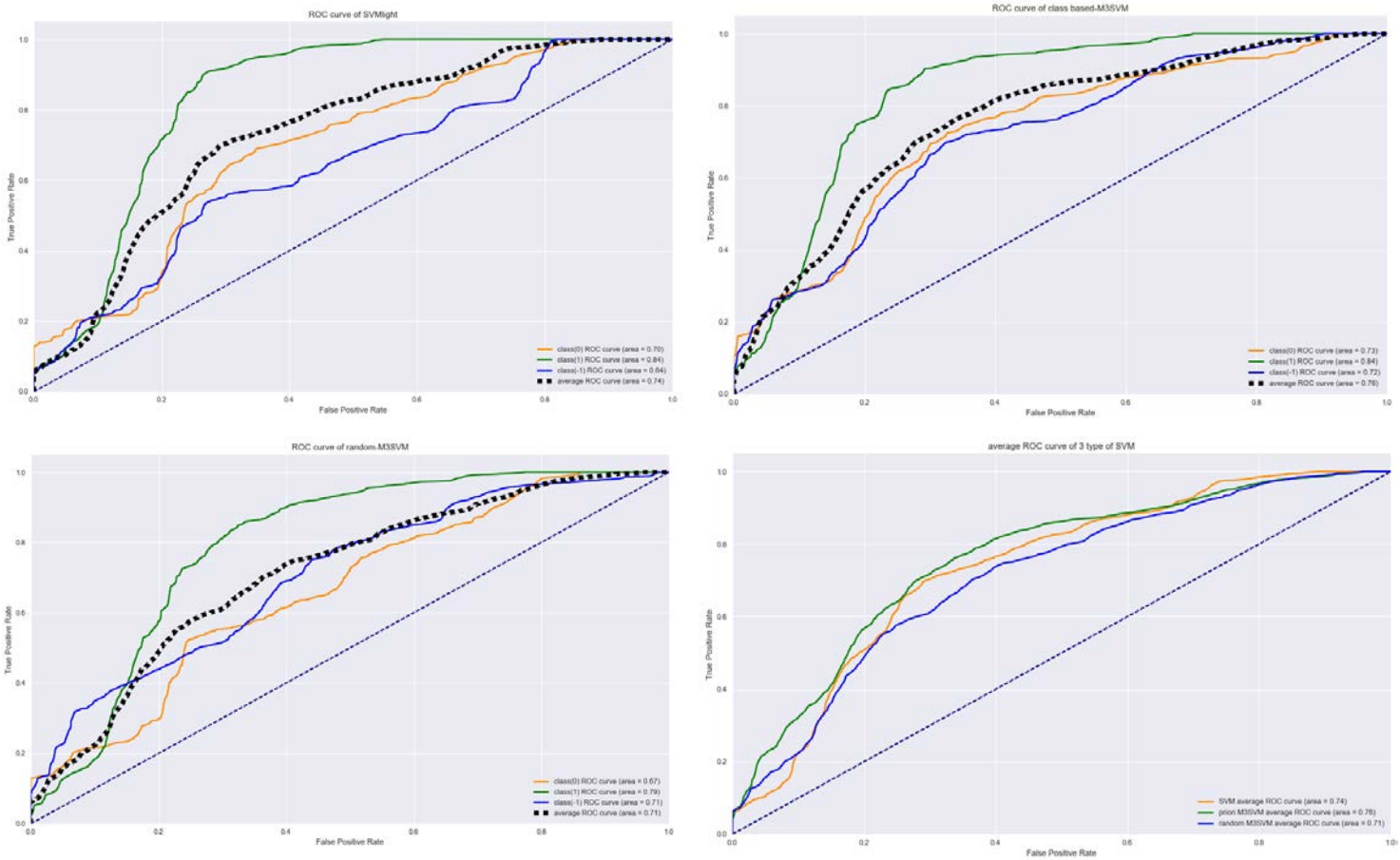

图 4 (a)朴素的 linearSVM ROC 曲线，平均 area=0.74　(b)基于先验知识的 M3SVM ROC 曲线，平均 area=0.76

(c) 随机 M3SVM ROC 曲线，平均 area=0.76　(d) 三种方法的平均 ROC 曲线对比

表 1 ROC area 统计

	CLASS(1)	CLASS(0)	CLASS(-1)	AVERAGE
SVM	0.84	0.70	0.64	0.74
先验 M3SVM	0.84	0.73	0.72	0.76
随机 M3SVM	0.79	0.67	0.71	0.71

图 4(a)的线性 SVM 分类器没有使用 min-max-modular 策略，在 class(1)的表现很突出，达到了 0.84 的 ROC area 成绩；在 class(0)中，该模型的表现基本与平均线近似；在 class(-1)的表现较差，仅仅只有 0.64 的 ROC area 成绩。根据每类样本的不同模型的表现差距很大。

而在图 4(b)使用了基于先验知识的线性 M3SVM 分类器，在 class(1)的表现依然很突出；而在 class(-1)的表现相比图 4(a)有了较大的改进，提升至 0.72 的 ROC area 成绩；class(-1)和 class(0)的分类结果基本达到了平均 ROC area 成绩。因为 class(-1)样本分类准确率得到了提升，所以模型最终达到了 0.76 的平均 ROC area 成绩。

在图 4(c)使用了随机的线性 M3SVM 分类器，相比于图 4(a)和图 4(b)，在 class(1)、class(0)和 class(-1)的得分均得到了不同成都的削弱，分别为 0.79、0.67、0.71；但另一方面，三类样本的分类得分分布更加集中，方差更小，可能会具有更小的过拟合风险。

总结，朴素的线性 SVM 分类器测试得分中规中矩，但是样本类别差异较大，具有一定的过拟合风险；基于先验知识的线性 M3SVM 分类器得分较高，同时削弱了样本类别的差异性，综合表现较优越；随机的线性 M3SVM 分类器得分较低，但因为其为样本分割加入了随机性，大幅削弱了样本类别的差异性，具有良好的减小过拟合的特性。