机器学习大作业报告

沈冠霖 段凡 杨天煜 符景洲

1.题目选择

原链接：https://tianchi.aliyun.com/competition/entrance/531825/introduction

本次一见钟情学习赛，是用机器学习算法，分析一个线下约会实验的问卷结果数据集。

数据集的内容包括实验志愿者的性别、年龄、人种、专业、地区、收入等特征，以及志愿者对配偶是否来自同一地区、同一信仰等观点的预期。

选手需要训练一个机器学习模型，去预测实验人身上一个或多个特性对其相亲成功与否的影响。也就是利用其它特征信息，预测数据集中的“match”字段的结果，1=成功，0=不成功。

总结一下，这是一个二分类问题。

2.算法选择

对于二分类问题，我们能选择若干有效的算法。

一是基于线性模型的逻辑回归，它适合处理连续数据，对于数据集里大量存在的“约会双方互相认可度”这个0-10之间的整数也处理的较为不错。

二是决策树，它适合处理分类问题，可解释性很强。决策树还可以进行bagging变成随机森林，也可以使用Adaboost算法进行boosting。

因此我们分工如下：

- 段凡：逻辑回归
- 符景洲：决策树（C4.5/CART）
- 杨天煜：随机森林
- 沈冠霖：AdaBoost

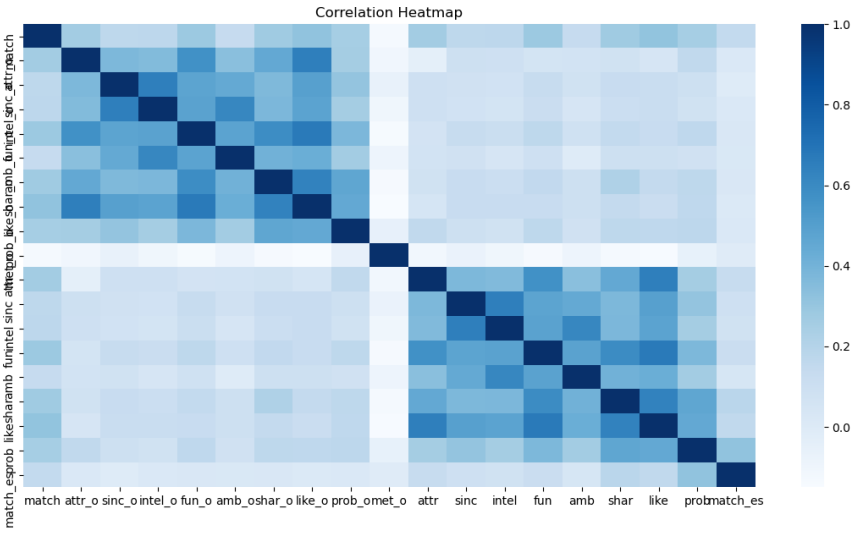
3.模型调优

3.1 特征选择

首先，我们考虑，约会对象的编号、组号等数据显然对解决问题毫无帮助，就把这些数据去除。

其次，为了防止过拟合，我们将缺失率较高的特征（缺失率大于0.7的）剔除

之后，我们计算了所有特征和结果的相关性，把相关性的绝对值大于等于0.1的特征留了下来。相关性可视化结果如下：



这样做能只保留17组特征，特征简化的很多了，而且在AdaBoost算法下达到了88.61%的训练准确率和86.14%的测试准确率，初步说明这样做有道理。

最终，我们分析特征的语义，以及根据上面的相关性结果，选择合适的特征组合。根据上方的相关性分析，我们发现，attr_o/attr等特征和结果相关性较高，而且这些特征反映的是第一次约会后双方对对方的评价，能够反映双方的互相喜欢程度，我们称这些特征为组1。我们发现，这个数据集里还有其他能够反映双方对对方评价的特征：约会后一天填写的问卷；约会后3-4周填写的问卷，我们记这些特征为组2，组3。将特征组1、2、3进行组合，就能得到7种特征组合。

3.2 数据处理

我们没有进行归一化处理。因为我们有三个算法是决策树类型的，无需进行归一化。事实上我们之前的特征选择的也都是约会双方的相互评价指标，都是0-10之间的整数，归一化也没用。

因为我们选择的特征都是0-10之间的整数，无论逻辑回归还是决策树类算法都能有效处理这种数据，不用担心离散/连续值的问题。

很多特征有缺失，为了防止过拟合我们剔除了缺失度较高的特征，并且用众数填补了缺失数据。

3.3 参数调优

首先，对于每个算法，我们都使用了grid search进行调优，找到了最优的超参数和特征组合。

比如，对于决策树算法，我们对比了C4.5和CART两种算法、是否向量化、以及树的最大深度，得到了不同结果：

C4.5：

	向量化	**无向量化**
特征选择	**85.30%**	84.69%
无特征选择	84.78%	84.45%

CART：

	向量化	**无向量化**
特征选择	**85.23%**	84.57%
无特征选择	84.48%	83.70%

树的最大深度：

树最大深度	5	6	7	8	9
CART	84.46%	84.43%	85.23%	84.10%	84.01%
C4.5	84.81%	85.47%	**85.30%**	84.34%	84.53%

我们得到，选用两种算法效果大体相当，其中C4.5算法和最大深度7效果最优。

之后，我们在最优网络超参数下，对于特征组合进行了一些分析：

逻辑回归：

	1	2	3	1+2	1+3	2+3	1+2+3
训练准确率	85.66%	83.63%	83.63%	86.11%	86.17%	83.63%	86.20%
测试准确率	85.15%	75.25%	75.25%	87.13%	**88.12%**	75.25%	88.12%

AdaBoost：

	1	2	3	1+2	1+3	2+3	1+2+3
训练准确率	86.88%	83.64%	83.63%	87.47%	87.4%	83.63%	87.51%
测试准确率	89.11%	75.25%	75.25%	85.15%	**93.07%**	75.25%	87.13%

可以看出，特征1是最基础的特征，必须含有特征1模型才能很好运行。而特征1加上特征3能够更好改进模型效果。因此，我们选择使用特征1+特征3的组合作为最终特征。

4.总结分析

4.1 结果比较

算法/指标	训练准确率	测试准确率	训练时间	测试时间
逻辑回归	86.15%	89.11%	0.091s	0.002s
决策树	85.23%	85.14%	0.007s	0.001s
随机森林	91.16%	82.18%	1.503s	0.009s

AdaBoost 算法/指标	87.40% 训练准确率	**93.07%** 测试准确率	2.082s 训练时间	0.036s 测试时间

整体效果来看，AdaBoost是最好的，达到了93.07%的准确率，而且没有出现过拟合。随机森林出现了过拟合现象；逻辑回归和决策树都有一定欠拟合现象；

速度上，四个算法训练测试都很快，其中决策树和逻辑回归尤其快。

对于这个问题，选择AdaBoost效果最好。

4.2 比赛结果

1	鲸落南北再入东西	宁波工程	1.0000	2020-10-26
1	小杰想喝奶茶	宁波工程	1.0000	2020-10-09
1	txkcpomwhczfe	郑州轻工	1.0000	2020-10-30
1	醉丶青楼	湖南文理	1.0000	2020-10-10
1	猫猫丸	山东大学	1.0000	2020-10-13
1	lvlouhss	西北工业大学	1.0000	2020-11-21
51	Gui_Yingbin	同济大学	0.9901	2020-11-29
52	天池壮壮	家里蹲	0.9703	2020-09-22
53	Minionsyh	华北电力大学(保定)	0.8515	2020-10-02
54	HeRaNO	电子科技大学	0.8020	2020-11-05
55	小白飞飞郑	中央财经大学	0.7921	2020-10-20
56	酥糖不加糖yy	西南民族大学	0.7822	2020-10-22
57	东来乡沙马特	四川大学	0.7723	2020-10-06
57	SisconCCCC	郑州轻工	0.7723	2020-10-27

截至2020年12月5日，我们以**93.07%**的准确率在排行榜上**排名第3**（这个竞赛的测试数据集有失误，结果match这一列可以直接用dec和dec_o做且运算做出来。。。所以排名靠前的准确率都是1，这显然是不合理的，我们没有用这两个特征，第3是刨除准确率为1的队伍后的结果）。

4.3 总结

这次作业是很好的用机器学习进行实战的机会。因为机器学习框架（sklearn、还有深度学习的pytorch等）的成熟，实现算法并不是一件很重要的事情，重要的是如下的事情：

- 根据实际问题选择正确的算法：分类还是回归问题？数据性质如何？大数据还是小数据？
- 对数据进行有效的处理：归一化、连续/离散值、缺失值、特征选择
- 超参数调优：统计学习的grid search，对于深度学习可能需要一些调参技巧