数据分析及实践 _Assignment5

Xiaoma

2023 年 5 月 22 日

1 实验要求

在实验三实现的数据分析的基础上使用 **PISA2018** 数据集，对 **REPEAT** 列进行分类。

- 实现至少一种分类算法（例如：决策树、KNN、朴素贝叶斯、感知机和集成算法等）
- 参考实验三中的特征工程，测试算法在 **PISA2018** 数据集上的预测性能，并撰写实验报告
- 实验报告需记录最终的方案

具体要求：

- 代码实现可以使用现有的机器学习库，也可以自行编写实现算法
- 预测任务与实验三一致，以 **ACC,F1_score** 为评价指标
- 使用 5 折交叉验证的方法测试模型性能

2 实验环境

VSCode + Python3.9.13

3 实验步骤

3.1 数据预处理

首先根据实验要求，去掉相关性最高的 5 个特征，已知实验三的实验结果，抽取剩余与 $REPEAT$ 最相关的 4 个特征以及 **REPEAT** 列，去除缺失 **REPEAT** 的样本，然后观察其余缺失值，通过观察可以发现有缺失值的特征 **HOMEPOS** 的缺失数量相对于数据总体很小，故为了保证数据准确性，将缺失值直接删去。

3.1.1 数据去噪

分别使用了 3σ 准则与 **EllipticEnvelope** 方法进行数据去噪，但通过后面对实验结果的观察可以发现这两种去噪方法对模型性能的影响几乎没有差别。

3.2 模型训练

使用 **sklearn.model_selection** 包中的 **cross_val_score** 方法直接进行 n 折交叉验证。

划分特征数据与 **REPEAT** 标签，分别使用 **XGBClassifier, MLPClassifier,DecisionTreeClassifier,KNeighborsClassifier** 模型来进行训练，经过反复调参，最终得到的最佳性能分别为

模型	5 折交叉 ACC	5 折交叉 F1-score
XGBClassifier	0.8203	0.1473
DecisionTreeClassifier	0.8240	0.0741
MLPClassifier	0.7651	0.0
KNeighborsClassifier	0.7481	0.2283

根据实验结果可以推测数据引入的特征量过少，根据实验三得到的相关度系数，我们选择对不同的模型使用不同的特征组合：

- **XGBClassifier**
  ISCEDL,AGE,HOMEPOS,COBN_F,ENTUSE,PROGN,
  COBN_M,IC001Q09TA
- **DecisionTreeClassifier**
  COBN_F,ST011D17TA,ENTUSE,HOMEPOS
  IC008Q02TA,IC001Q11TA,MISCED_D,SCREADDIFF
- **MLPClassifier**
  ISCEDL,AGE,HOMEPOS,COBN_F,ENTUSE,PROGN,
  COBN_M,IC001Q09TA
- **KNeighborsClassifier**
  ISCEDL,AGE,HOMEPOS,COBN_F,ENTUSE,PROGN,
  COBN_M,IC001Q09TA

通过反复调参，最终得到的最佳性能分别为

模型	5 折交叉 ACC	5 折交叉 F1-score
XGBClassifier	0.8391	0.2050
DecisionTreeClassifier	0.8302	0.1706
MLPClassifier	0.8301	0.0
KNeighborsClassifier	0.7881	0.1732

取四个模型所有特征的并集，通过反复调参，最终得到的最佳性能分别为

模型	5 折交叉 ACC	5 折交叉 F1-score
XGBClassifier	0.8375	0.2195
DecisionTreeClassifier	0.8316	0.1401
MLPClassifier	0.8301	0.0
KNeighborsClassifier	0.7999	0.1859

由结果可知，模型性能已几乎不变。

考虑将所有特征全部使用，最终得到的最佳性能分别为

模型	5 折交叉 ACC
XGBClassifier	0.6869
DecisionTreeClassifier	0.8109
MLPClassifier	0.8013
KNeighborsClassifier	0.3872

通过观察可知，使用除了最初去除的 5 个特征以外的全部特征进行训练，得到的模型性能要比使用子集差，则我们选择 **COBN_F,OCOD2, ISCEDL,AGE, HOMEPOS,ENTUSE,PROGN,COBN_M,IC001Q09TA, ST011D17TA, IC008Q02TA,IC001Q11TA,MISCED_D,SCREADDIFF** 作为最终选择的训练特征。

4 总结

- 通过观察实验结果可以发现，在训练过程中，加入与预测标签不相关的特征对预测任务无积极效果。
- 实际上训练数据中 **REPEAT** 为 1 的值的数据只有整体的 0.168，故实际上训练的模型与随机猜测相差无几，通过，观察F1-score也验证了这一事实。

- 使用 **XGBClassifier** 作为最终的分类器模型，随机的将数据集拆分为 5 份并使用其中一份作为验证集，最终的性能稳定在 **ACC = 0.78,F1-score = 0.38** 上下。

观察使用的特征含义：

特征	含义
COBN_F	Country of Birth National Categories- Mother
COBN_M	Country of Birth National Categories- Father
OCOD_2	ISCO-08 Occupation code - Father
AGE	AGE
HOMEPOS	Home possessions (WLE)
ENTUSE	ICT use outside of school (leisure) (WLE)
PROGN	Unique national study programme code
IC001Q09TA	Available for you to use at home: Printer
ST011D17TA	In your home: <Country-specific wealth item 1>
IC008Q02TA	Use digital devices outside of school: Playing collaborative online games.
IC001Q11TA	Available for you to use at home: <ebook reader>, e.g. <Amazon Kindle>
MISCED_D	Mothers Education - alternate definition (ISCED)
SCREADDIFF	Self-concept of reading: Perception of difficulty (WLE)

可以发现复读还与家庭的经济状况，父母的职业与教育程度，个人是否沉迷与网络有关。