|  |
| --- |
| **机器学习** |
| 减少商业航空死亡人数 |
| 主 研 人：赵翰宇  参 研 人： |
| 审 核 人：  方 向：工业算法案例研究  版 本 号：A |
|  |

**版本更新**

|  |  |  |  |
| --- | --- | --- | --- |
| **日期** | **版本** | **更新描述** | **作者** |
| 2019/3/29 | A | 初稿 | 赵翰宇 |

**选题表格**

|  |  |  |  |
| --- | --- | --- | --- |
| 时间 | 竞赛名 | 竞赛背景描述（50字以内） | 类型（分类/回归） |
| 2019/3/18 | Reducing Commercial Aviation Fatalities(Kaggle) | 建立一个模型来检测机组人员生理数据来检测人员的突发状况，飞行员可以在进入令人不安的状态时收到警报，防止事故发生并挽救生命。 | 分类 |

目录

[1. 背景描述 4](#_Toc25575_WPSOffice_Level1)

[1.1 竞赛赛题描述 4](#_Toc5940_WPSOffice_Level2)

[1.2 评估指标描述 4](#_Toc12357_WPSOffice_Level2)

[2. 数据来源及描述性统计分析 5](#_Toc5940_WPSOffice_Level1)

[2.1 大赛数据来源 5](#_Toc22683_WPSOffice_Level2)

[2.2 数据的描述性统计 5](#_Toc16174_WPSOffice_Level2)

[2.2.1 数据基本情况描述： 5](#_Toc5940_WPSOffice_Level3)

[2.2.2 数据字段介绍： 6](#_Toc12357_WPSOffice_Level3)

[2.2.3 数据描述性统计 7](#_Toc22683_WPSOffice_Level3)

[3. 优秀算法思路 8](#_Toc12357_WPSOffice_Level1)

[3.1 方案一 8](#_Toc32293_WPSOffice_Level2)

[3.1.1 方案一特征提取 8](#_Toc16174_WPSOffice_Level3)

[3.1.2 方案一模型设计与建立 8](#_Toc32293_WPSOffice_Level3)

[3.1.3 方案一结果、排名等 8](#_Toc29922_WPSOffice_Level3)

[3.1.4 方案一算法流程图 8](#_Toc4787_WPSOffice_Level3)

[3.2 方案二 8](#_Toc29922_WPSOffice_Level2)

[3.2.1 方案二特征提取 8](#_Toc16174_WPSOffice_Level3)

[3.2.2 方案二模型设计与建立 9](#_Toc32293_WPSOffice_Level3)

[3.2.3 方案二结果、排名等 9](#_Toc29922_WPSOffice_Level3)

[3.2.4 方案二算法流程图 9](#_Toc4787_WPSOffice_Level3)

[4. 算法比较 10](#_Toc22683_WPSOffice_Level1)

[表4-1 算法比较 10](#_Toc7253_WPSOffice_Level2)

[5. 总结与展望 10](#_Toc16174_WPSOffice_Level1)

[5.1 总结 10](#_Toc23198_WPSOffice_Level2)

[5.2 建模思路 10](#_Toc16619_WPSOffice_Level2)

1. 背景描述

大多数与飞行相关的死亡事故源于“飞机状态意识”的丧失。也就是说，飞行员可能会分心，困倦或处于其他危险的认知状态时，注意管理效率低下。减少飞机死亡率只是Booz Allen Hamilton 100多年来为商业，政府和军事领导者解决的复杂问题之一 。他们在没有路线图的情况下提供原创解决方案。现在，你可以帮助他们找到答案，拯救生命，改变世界。

**1.1 竞赛赛题描述**

建立一个模型来检测机组人员生理数据中的问题。你将使用在测试情况下从实际飞行员获取的数据，并且你的模型应该能够实时运行计算以监控飞行员的认知状态。在你的帮助下，飞行员可以在进入不正常的状态时收到警报，防止事故发生并挽救生命。

**1.2 评估指标描述**

评估的指标为预测概率和实际观察的数据之间的多类对数损失。对数损失，又称逻辑损失或交叉熵损失。这是（多项式）逻辑回归中使用的损失函数及其扩展，例如神经网络，定义为给定概率分类器预测的真实标签的负对数似然。仅记录两个或多个标签的日志丢失。对于在{0,1}中的真实标签yt和yt = 1的估计概率yp的单个样本，对数损失为

该公式可用Python中sklearn.metrics.log\_loss来实现。

提交的概率不需要总和为1，因为它们在被评分之前被重新调整（每行除以行总和）。为了避免对数函数的极值，将预测概率替换为

1. 数据来源及描述性统计分析

**2.1 大赛数据来源**

在这个数据集中，将获得来自18名飞行员的真实生理数据，这些飞行员会​​受到各种分散注意力的事件。基准训练集包括在飞行模拟器外部的非飞行环境中收集的一组受控实验。测试装置（缩写LOFT为“面向线路的飞行训练”）包括飞行模拟器中的完整飞行（起飞，飞行和着陆）。

飞行员经历了被打扰的过程，旨在诱发以下三种状态之一：

大致而言，频道化注意（CA）是关注一项任务以排除所有其他任务的状态。这是通过让测试者玩一个引人入胜的基于拼图的视频游戏在基准测试中引起的。

转移注意力（DA）是指通过与决策相关的行动或思维过程转移注意力的状态。这是通过让测试者执行显示监视任务来诱导的。测试中会定期出现数学问题，在返回监控任务之前必须解决这个问题。

通过让测试者观看具有跳跃恐怖的电影剪辑来诱发惊吓/惊喜（SS）。

对于每个实验，一对飞行员（每个飞行员具有其自己的crew id）记录并且经受CA，DA或SS认知状态。训练集包含三个实验（每个状态一个），其中飞行员只经历过一个州。例如，在experiment= CA中，飞行员处于基线状态（无事件）或CA状态。测试装置包含一个完整的飞行模拟，在此期间飞行员可以体验任何状态（但一次不会超过一个）。

每个传感器以256 Hz的采样率运行。请注意，由于这是来自真人的生理数据，因此数据中会出现噪音和伪影。

这里是数据下载的超链接：https://www.kaggle.com/c/11835/download-all

**2.2 数据的描述性统计**

**2.2.1 数据基本情况描述：**

sampleSubmission.csv文件是输出样例。大小为18m×5

**train.csv大小为4.8m×28，**包含了近几乎每0.004秒记录一次的9组飞行员在4项测试中的脑电图，心电图等生理数据。

**、2.2.2 数据字段介绍：**

表2-1 sampleSubmission.csv数据表字段介绍

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **变量** | | **变量含义** | **变量类型** | **缺失率** |
| **Id** | 机组人员和时间的联合编码，是概率预测的对象 | | 连续递增整型 | 0% |
| **A** | 基准线 | | {0,1} | 0% |
| **B** | SS | | {0,1} | 0% |
| **C** | CA | | {0,1} | 0% |
| **D** | DA | | {0,1} | 0% |

表2-2 **train.csv**数据表字段介绍

|  |  |  |  |
| --- | --- | --- | --- |
|  | **变量含义** | **变量类型** | **缺失率** |
| **id** | 船员+时间组合的唯一标识符 | 离散 | 0% |
| **crew** | 一对飞行员的唯一身份证 | 离散 | 0% |
| **experiment** | 一个CA，DA，SS或LOFT的测试。前3个包括训练集。后者是测试集。 | 离散 | 0% |
| **time** | 进入实验的秒数 | 离散 | 0% |
| **seat** | 左（0）或右（1）座位的飞行员 | {0,1} | 0% |
| **eeg-** | 脑电图记录 | 离散 | 0% |
| **ecg** | 心电图信号。传感器的分辨率/位为.012215μV，范围为-100mV至+ 100mV。数据以微伏提供。 | 离散 | 0% |
| **r** | 呼吸，衡量胸部的上升和下降。传感器的分辨率/位为.2384186μV，范围为-2.0V至+ 2.0V。数据以微伏提供。 | 离散 | 0% |
| **gsr** | 电镀皮肤反应，一种皮肤电活动的量度。传感器的分辨率/位为.2384186μV，范围为-2.0V至+ 2.0V。数据以微伏提供。 | 离散 | 9% |
| **event** | 给定时间的飞行员状态：其中一个A=基线，B= SS，C= CA，D= DA | {A,B,C,D} | 0% |

表2-3 **飞行员状态**字段介绍

|  |  |
| --- | --- |
|  | **变量含义** |
| **Baseline** | 无事件发生 |
| **CA** | 关注一项任务以排除所有其他任务的状态。 |
| **DA** | 通过与决策相关的行动或思维过程转移注意力的状态 |
| **SS** | 通过让受试者观看具有跳跃恐慌的电影剪辑来诱发**惊吓/惊喜** |

**2.2.3 数据描述性统计**

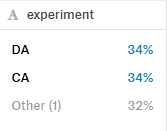
 该竞赛数据量非常大，共有28个维度，数据的缺失率低，可以得到较好的训练模型。其中脑电图数据的分布基本一致，同属于segma较小的正态分布。心电图数据的分布为segma较大的正态分布。对于实验种类，CA和DA分别占了三分之一，其余由SS和baseline构成。对于呼吸数据，离群值较多，分布不均。电镀皮肤反应分布较为分散。

图2-1 实验种类分布

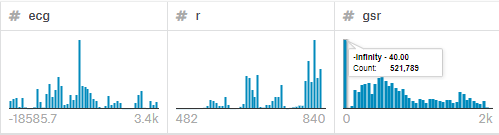
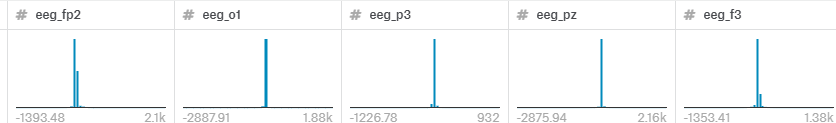
图2-2 各类脑电图分布节选

图2-1 心电图，呼吸，电镀皮肤反应分布

3. 优秀算法思路

**3.1 方案一**

**3.1.1 方案一特征工程**

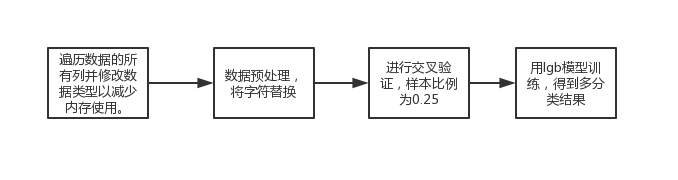
此方案首先遍历数据的所有列并修改数据类型以减少内存使用。其次开始数据预处理，将A、B、C、D替换为0、1、2、3。随后按训练集和测试集的0.25随机抽取样本数进行交叉验证。

**3.1.2 方案一模型设计与建立**

方案一采用LGB训练模型，其目标设为多分类，学习率为0.01，梯度增强2000次。其余没有做任何优化处理。

**3.1.3 方案一结果、排名等**

此方案获得了第9名，公共排行榜上多类对数损失为0.31120排名第5位，私人排行榜上多类对数损失为0.55509，排名第9位。

** 3.1.4 方案一算法流程图**

**图3-1**

**3.2方案二**

**3.2.1 方案二特征工程**

方案二先进行了数据分析，将每列数据的分布图打印出来观察，进行数据缺失值处理，去除其所在行。选用三折交叉验证。其余和上面方法相同。

**3.2.2 方案二模型设计与建立**

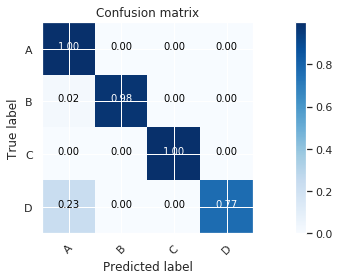
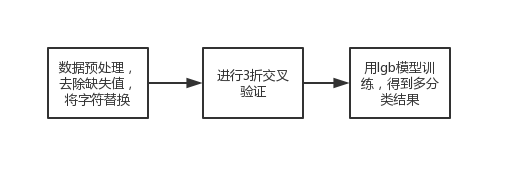
 方案同样采用LGB模型，学习率为0.1，最大深度为8层，其余参数和方案一一致。训练后输出了混淆矩阵作为分类结果展示。之后对测试集进行预测。

图3-1 多分类的混淆矩阵

**3.2.3 方案二结果、排名等**

此方案获得了第11名，公共排行榜上多类对数损失为0.41543，排名第59位，私人排行榜上多类对数损失为0.56702，排名第11位。

**3.2.4 方案二算法流程图**

**图3-2**

4. 算法比较

**表4-1 算法比较**

|  | **评估指标** | **特征工程** | **基础算法** | **基本库** |
| --- | --- | --- | --- | --- |
| **算法1** | **多类对数损失** | **替换字符串** | **梯度提升决策树改进（lgb）** | **Python的lightgbm** |
| **算法2** | **多类对数损失** | **去除缺失值，替换字符串** | **梯度提升决策树改进（lgb）** | **Python的lightgbm** |

方案1，2运用了相同的算法。唯一的差距在于学习率的设置上。方案二在特征提取上的做的工作要比方案一细致，但构建模型时参数为调整好，学习率太大了因此可能会欠拟合。方案一同时赢在了构建训练集上。参数调整的适宜是他的预测结果较好。

5. 总结与展望

**5.1 总结**

此次飞行员状态分类竞赛的关键在于处理模型参数的设置，设计的算法以lgb为最优。但在特征工程建设方面还有待提高。

**5.2 建模思路**

我的思路是，先对数据进行预处理，将字符串替换，去除缺失值所在行，其次对脑电图的特征进行PCA降维降至3维，再用PCA将这3维和其余特征一起降至3维。由于训练集数据量过大，先对数据进行聚类分为4类，随机从4类中挑选1000个数据进行训练，重复1000次。训练的模型采用多分类支持向量机。