赛题描述

鼠标轨迹识别当前广泛运用于多种人机验证产品中，不仅便于用户的理解记忆，而且极大增加了暴力破解难度。但攻击者可通过黑产工具产生类人轨迹批量操作以绕过检测，并在对抗过程中不断升级其伪造数据以持续绕过同样升级的检测技术。我们期望用机器学习算法来提高人机验证中各种机器行为的检出率，其中包括对抗过程中出现的新的攻击手段的检测。

比赛数据

本题目数据来源于某人机验证产品采集的鼠标轨迹，经过脱敏处理，数据分为3部分（数据量分别为3000条，10万，200万）。

赛事分为三个阶段（初赛、复赛、决赛答辩）：5月26日初赛提供3000条数据作为训练样本，供参赛者下载进行建模和模型优化，同时提供10万条正式比赛数据供下载评测，识别结果为初赛得分；复赛提供200万条比赛数据（不可下载，数据不可见，仅供评测），识别结果为复赛得分；决赛将以现场答辩会的形式进行。

【训练数据】

训练数据表名称：dsjtzs_txfz_training

字段	类型	解释
a1	bigint	编号id
a2	string	鼠标移动轨迹(x,y,t)
a3	string	目标坐标(x,y)
label	string	类别标签：1-正常轨迹，0-机器轨迹

训练样例数据：见 dsjtzs_txfz_training_sample.txt

【测试数据】

主办方：教育部高等学校计算机类专业教学指导委员会

教育部高等学校软件工程专业教学指导委员会

教育部高等学校大学计算机课程教学指导委员会

全国高等学校计算机教育研究会

报名时间：2017-04-01 08:00 - 开始

2017-06-30 10:00 - 结束

比赛时间：2017-05-26 08:00 - 开始

2017-07-21 10:00 - 结束

距离竞赛开始还剩：0 天

初赛测试表名称：dsjtzs_txfz_test1

复赛测试表名称：dsjtzs_txfz_test2

字段	类型	解释
a1	bigint	编号id
a2	string	鼠标移动轨迹(x,y,t)
a3	string	目标坐标(x,y)

测试样例数据：见 dsjtzs_txfz_test_sample.txt

测评标准

选手请将识别为机器行为的编号id提交到计算平台，需要提交的结果表，只包含一个字段：编号id。

初赛提交表名：dsjtzs_txfzjh_preliminary

复赛提交表名：dsjtzs_txfzjh _semifinal

设定Precision为P，Recall为R，白样本为正常轨迹，黑样本为机器轨迹其中：

P = 判黑的数据中真正为黑的数量/判黑的数据总量，

R = 判黑的数据中真正为黑的数量/真实黑数据总量，

比如10w条数据，其中8w条为白样本，2w条为黑样本，参赛者一共将1w条判断为黑样本（其中真正的黑样本有8000条，错将2000条白样本判黑），那么，

P=8000/10000 = 80%，

R=8000/20000=40%，

参赛队伍最终得分F = 5PR/(2P+3R)*100。最终排名按照F值评判，F值越大，代表结果越优，排名越靠前。