基于机器学习的抖音商用广告视频识别

未明学院线上课程项目训练

1.任务与数据

1.1 任务背景

近几年短视频行业异军突起，目前短视频独立用户数已经超过 5 亿，占国内网民总数的一半左右。许多商家与广告up主也抓住时机，加速入局短视频营销，借力平台进行推广，2018年短视频KOL营销市场规模已突破10亿。但低质量的硬广十分影响用户对视频平台的使用体验，许多平台都对“营销号”进行审核并限流。人工审核视频以识别广告的方式显然难以满足短视频平台庞大的视频量，因此阿里云等机构使用人工智能等新兴科技提升识别广告的效率。

1.2 任务

一般来说，为了吸引观众的注意力，广告视频的长度、音频、文本位置和画面会有与众不同之处。这里我们将使用人工智能的方法构建一套商用广告识别系统来预测抖音短视频是否为商用广告。通过对抖音平台上视频的时长、声音频谱、视频光谱、文字分布和画面变化等特征，进行特征抽取、特征过滤等方式处理后进行建模，构建一套商用广告视频识别系统，来快速区分出投稿视频中的商用广告。具体内容包括：

a. 了解这份数据

b. 进行必要的数据清洗

c. 自由进行特征生成、特征选择、特征降维等工作

d. 建立合适的预测模型，并进行调参

e. 选用合适的方式进行模型集成，优化模型

1.3 数据

数据从5次采样、总长度为150小时的抖音视频中提取的视频镜头，并构建标准视听特征。最终包括129685份样本，收集了视频的时长、声音频谱、视频光谱、文字分布和画面变化等特征。最终的数据包含1个标签、231个特征。详细变量解释见readme文件。

2.要求

选择合适的模型、策略和算法，使用jupyter书写报告和代码。

每人提交一份**jupyter代码（必做）**，也可额外提交一份**项目报告（选做）**。