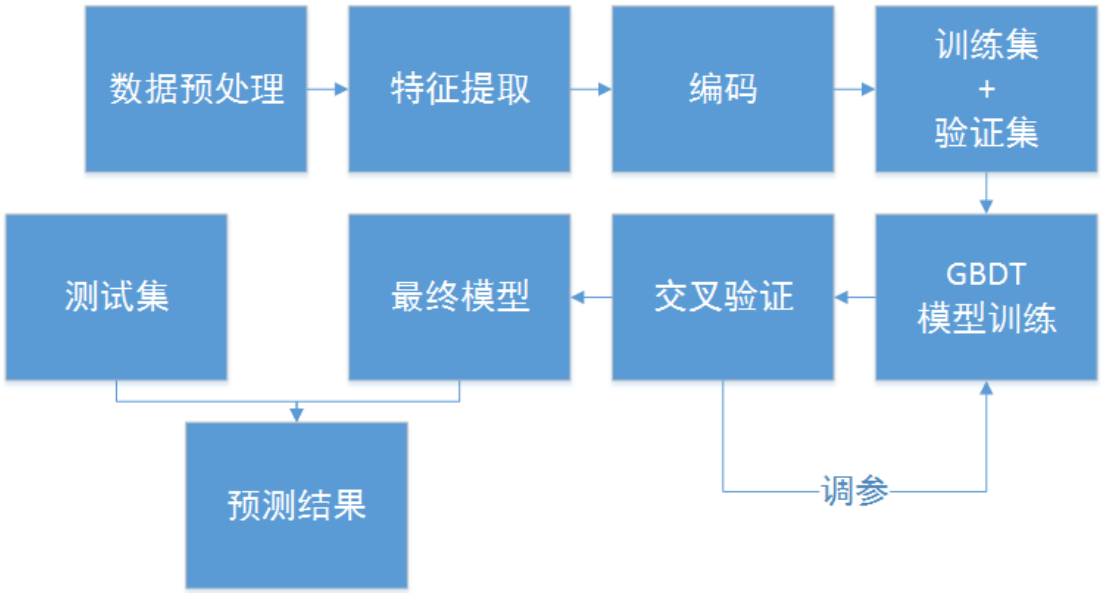
携程机票航班延误预测大赛算法说明文档

摸石头过河团队

目录

算法思路

本算法首先对数据进行预处理并提取相关的特征，然后对数据进行编码，选用GDBT 模型对数据进行训练，调整参数选出较优模型对测试集进行预测，总体算法如下：



数据预处理

本次大赛使用的数据集包括历史航班动态起降数据、历史城市天气表、机场城市对应表以及历史机场特情表等全部为主办方提供数据，未使用外部数据。数据预览如下：

表1 航班动态历史数据表

出发机场	到达机场	航班编号	计划起飞时间	计划到达时间	飞机编号
YIW	CGO	CZ6661	1.47E+09	1.47E+09	1426
XIY	TNA	CZ6959	1.47E+09	1.47E+09	2337
CTU	URC	CA4191	1.47E+09	1.47E+09	1135
PVG	NNG	FM9385	1.47E+09	1.47E+09	2660
SZX	CTU	3U8702	1.47E+09	1.47E+09	922
TAO	URC	SC8826	1.47E+09	1.47E+09	2307

表2 城市天气表

城市名	天气	当天最低温度	当天最高温度	日期
西宁	雷阵雨转晴	10	24	2016/7/1
西宁	多云转晴	9	29	2016/7/2
西宁	多云转晴	10	30	2016/7/3
西宁	晴转多云	14	32	2016/7/4
西宁	多云	16	31	2016/7/5
西宁	晴转多云	15	31	2016/7/6
西宁	阵雨	17	28	2016/7/7
西宁	小雨	16	28	2016/7/8

表3 机场城市对应表

机场编码	城市名称
AHJ	阿坝
AYN	安阳
HSC	韶关
HCJ	河池
XNT	邢台

表4 机场特情表

特情机场	收集时间	开始时间	结束时间	特情内容
CAN	2017-05-31 19:23:56Z	2017-06-01 08:30:00Z	2017-06-01 20:00:00Z	广州区域部分航路延误黄色预警提示：6月1日广州区域内部分航路受雷雨天气影响，预计08:30-20:00通行能力下降30%左右。【空中交通网】
PEK	2017-06-01 19:22:39Z	2017-06-02 06:00:00Z	2017-06-02 12:00:00Z	北京终端区航班延误黄色预警提示：6月2日北京终端区预计06:00-12:00受雷雨天气影响，通行能力下降30%左右。【空中交通网】
CAN	2017-06-01 19:28:01Z	2017-06-02 08:00:00Z	2017-06-02 20:00:00Z	广州终端区航班延误黄色预警提示：6月2日广州终端区预计08:00-20:00受雷雨天气影响，通行能力下降20%左右。【空中交通网】
CAN	2017-06-01 19:28:01Z	2017-06-02 08:00:00Z	2017-06-02 20:00:00Z	广州终端区航班延误黄色预警提示：6月2日广州终端区预计08:00-20:00受雷雨天气影响，通行能力下降20%左右。【空中交通网】

历史航班动态数据预处理：

1. utc 时间转换为标准时间格式，如机场特情表中的时间格式，提取出航班的计

划起飞日期 、计划到达日期、计划起飞时刻、计划到达时刻、航班月份等特征。

2. 计算起飞延误时间和到达延误时间
   起飞延误时间=实际起飞时间-计划起飞时间
   到达延误时间=实际到达数据-计划到达时间
3. 取消航班全部设置为延误 10 小时
4. 计算计划飞行时间
   计划飞行时间 = 计划到达时间 – 计划起飞时间
5. 前序航班到达延误时间
   前序航班的定义为：同一架飞机，当前航班的前一个航班即为当前航班的前序航班。比如，同一架飞机连续飞两个航班 A：南京--北京，B：北京--西安，则 A 为 B 的前序航班。首先对根据飞机编号对数据集进行分组，每个分组按航班的计划起飞时间排序，将前序航班的到达延误时间匹配到当前航班作为一个特征。
6. 起飞间隔
   起飞间隔定义为当前航班的计划起飞时间与前序航班的时间到达时间差值。
7. 航空公司
   根据航班号提取出航空公司的代号
8. 航班性质
   根据航班号的特征，把航班分为 3 种：尾号为字母的是补飞的，3 位数字国内航班，4 位数字国外航班。

城市天气表数据预处理：

1. 气温
   气温划分为 3 个取值，大于 40 度为高温，小于-10 度为低温，其他为一般
2. 天气情况
   天气情况（小雨、阴天等）根据城市天气表，首先统计两年时间内所有天气情况在各地区总共出现的频率，出现频率小于 50 的天气情况统一划归为‘other’。

机场城市对应表预处理

不作预处理

机场特情表预处理

1. 根据开始时间提取出特情开始日期、开始时刻
2. 根据结束时间提取出特情结束日期、结束时刻

合并表格

对处理后的 4 个表格进行合并。

合并天气城市表和机场城市对应表

1. 对机场城市对应表去重
2. 左连接操作。左表：天气城市表；右表：机场城市对应表。left_on=[‘城市

名’]，right_on=[‘城市名称’]。
3. 去除无关字段，命名新表为‘天气机场表’

合并航班动态表和天气机场表

匹配出发机场天气

1. 根据['日期','机场编码'] 对天气机场表去重
2. 左连接操作。左表：航班动态表；右表：天气机场表。['出发机场','计划起飞日期'],right_on=['机场编码','日期']。
3. 去除无关字段

匹配到达机场天气

1. 根据['日期','机场编码'] 对天气机场表去重
2. 左连接操作。左表：航班动态表；右表：天气机场表。['到达机场','计划到达日期'],right_on=['机场编码','日期']。
3. 去除无关字段

匹配好的表格依旧命名为“航班动态表”。

合并航班动态表和机场特情表

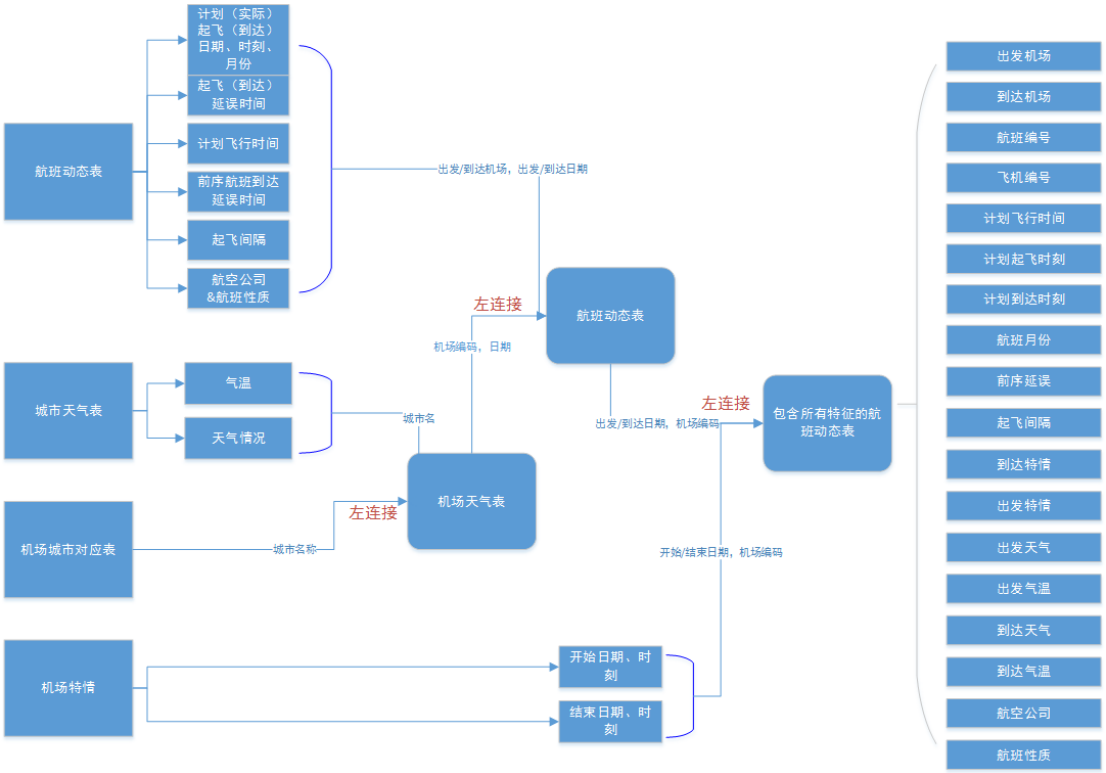
匹配出发机场特情

1. 根据['出发日期','机场编码'] 对机场特情表去重
2. 左连接操作。左表：航班动态表；右表：机场特情表。['出发机场','计划起飞日期'],right_on=['机场编码','日期']。
3. 若出发时刻在特情的开始与结束时刻之内，则有特情
4. 去除无关字段

匹配到达机场天气

1. 根据['到达日期','机场编码'] 对机场特情表去重
2. 左连接操作。左表：航班动态表；右表：机场特情表。['到达机场','计划到达日期'],right_on=['机场编码','日期']。
3. 若到达时刻在特情的开始与结束时刻之内，则有特情
4. 去除无关字段

最终生成包含所有特征的航班动态表，数据清理的大体流程如下图：



编码

在正式使用提取完特征的数据之前，需对这些数据的非数值特征类型进行编码，需要编码的特征有'出发机场'、'到达机场'、'出发天气'、'到达天气'、'航班编号'、'出发气温'、'到达气温'、'航空公司'。编码方式使用 skleam. preprocessing.LabelEncoder(),它将离散的字符分别映射为一个数值。

算法模型

本次大赛算法所选模型为 GBDT (Gradient Boosting Decision Tree)模型，直接调用了 sklearn. sklearn.ensemble, GradientBoostingClassifier()方法，并进行相关的参数调整。为了选取较优的参数，首先将编码后的数据集按照 3:1 的比例随机划分为训练集和验证集以调整参数。

调参步骤：

评判准则，验证集的 auc 得分。

1. 学习器个数（n_estimators）

   保持其它默认参数不变，考虑到样本量较大，故将学习器的个数设置大一些，分别将 n_estimators 设置为 100，200.，300，400，发现 auc 的得分是随着 n_estimators 的增多而变大的，但学习器个数 300 和 400 之间 auc 得分增加的已不明显，并且要兼顾到模型训练的时间开销，所以设置n_estimators=300.

2. loss function(loss)

   保持学习器个数不变，改变 loss，观察结果，选一个较优的，本算法选择了 loss=' exponential'

3. 树的最大深度（max_depth），叶子的最小样本个数（min_samples_leaf）

   保持学习器个数和 loss function 不变，调整 max_depth 和 min_samples_leaf。max_depth 越大，树的结构越复杂，组合特征的能力越强，拟合能力越强，但同时更容易过拟合，分别设置 max_depth 为 4,6,8,10,同时为了防止过拟合，min_samples_leaf 的个数逐渐调大，范围在 50-500 之间。

4. 下采样率（subsample）

   subsample=0.8,一般情况下都比较好。

5. 学习率
   学习率默认设置为 0.1，学习率过大容易错过最优解，所以最后适当缩小学习率的值，但减小学习率，收敛速度变慢，这时候适当增加学习器个数，最终学习率设置为 0.08.

提交文件说明

1. 文件夹—原始训练集

   - flight.csv：航班动态数据
   - weather.csv： 城市天气
   - airport.csv： 机场城市对应表
   - spcial.csv： 特情

2. 文件夹—原始测试集

   - flight.csv：航班动态数据
   - weather.csv：城市天气
   - airport.csv： 机场城市对应表
   - spcial.csv： 特情

3. 文件夹—处理后训练集

- train_data1.csv：数据预处理后的最终航班动态数据

4. 文件夹—处理后测试集

- test_data1.csv：数据预处理后的最终 8 月份航班动态数据（只包含标识为 1 的数据）

5. 文件夹--附件

- weather_case.csv：处理后的所有天气情况的汇总（出现频率小于 50 的天气情况都设置为‘other’）
- colums_type.pickle：数据预处理完训练集每一列的数据类型

6. 代码文件

- preprocessing_train_data.ipynb：训练集数据预处理
- preprocessing_test_data.ipynb：测试集数据预处理
- Model_training.ipynb：模型以及训练数据

代码说明

环境说明

运行系统：win10

编程语言：Python3.6

外部程序包：pandas, numpy, matplotlib, sklearn

变量说明

见详细注释。

运行说明

数据预处理运行说明

1. jupyter 打开 preprocessing_train_data.ipynb 和 preprocessing_test_data.ipynb
2. 修改训练集和测试集文件夹所在路径
3. 逐步运行每一个代码块
4. 输出结果

模型运行说明

1. jupeter 打开 Model_training.ipynb
2. 修改 train_data 和 test_data 所在路径
3. 逐步每一个代码块
4. 输出预测结果