華東理工大學

模式识别大作业

题　　目　出租清单查询

学　　院　信息科学与工程

专　　业　控制科学与工程

组　　员　卢宇晟、杨丹、王承天

指导教师　赵海涛

完成日期：　**2019** 年 **11** 月 **30** 日

模式识别作业报告——出租清单查询

组员：卢宇晟、杨丹、王承天

经过半个学期的对模式识别课程的学习，赵老师已经给我们展示了模式识别、机器学习等问题的处理思路，并为我们展示了经典的模式识别算法解决问题的流程和详细的推导过程。然而，“纸上得来终觉浅，绝知此事要躬行”，仅仅停留在享受推导出结果的乐趣是不实用的，只有将所应用于实际问题才能进一步巩固自己的知识。

模式识别在实际问题中的应用多种多样，当下一些新颖的算法也层出不穷，在模拟的数据集上也取得了值得关注的结果，但是从实际数据出发解决实际问题仍然是一个现实需求。因此我们小组参照这个思路进行选题，选择了一个模式识别相关的 kaggle 比赛——出租清单查询比赛。当然，在将理论与实际相结合的过程中难免会遇到阻碍。具体的在该选题下则体现为运用以 lightgbm 为代表的传统机器学习算法时复杂的特征工程和算法调用等。

在确定了选题方向后，经过小组讨论进行了明确的分工。在本文中，我们先对实际问题做了简要介绍，之后我们将从特征工程角度入手，展示 2017 年 kaggle 出租清单查询比赛优胜方案在传统模型的特征工程方法，并展示其单模型的预测结果。其中比赛的详细信息可以在 https://www.kaggle.com/c/two-sigma-connect-rental-listing -inquiries 中找到。

一、 背景介绍

1.1 比赛背景介绍

在出租房屋中，相比浏览各式各样大同小异的出租清单，人们更愿意通过打电话或者其他形式来找到适合自己的出租房屋。RentHop 公司通过使用数据对租赁清单按质量进行排序，使得对出租房屋的搜索更加智能。尽管寻找完美的出租房屋已经足够困难，但想通过编程的方式结构化并利用所有房地产数据更是难上加难，因此希望选手在本次比赛中从一系列原始数据中提取出相关特征并对出租清单按质量进行多分类。

1.2 解题思路

与其他比赛相比，出租清单查询的难点在于非结构化数据的利用方法，也可称为特征工程。采用传统模型预测常用的特征工程思路为尽可能多地提取有用的特征，并将之放入模型进行训练。其中，由于数据量巨大，过拟合风险会低一些；其次，数据泄漏是特征工程会出现的一个常见问题，会使得预测精度更高但是在

真实场景中这些特征并不可用，本文从比赛角度先不关注这些，仅复现冠军方法的特征工程思路。

二、解决方案

2.1 数据结构分析

bathrooms	bedrooms	building_id	created	description	display_address
1	1	8579a0b0d54db803821a35a4a615e97a	2016年6月16日	& G Trains !<br	145 Borinquen Place
1	2	b8e75fc949a6cd8225b455648a951712	2016年6月1日	br /> I op	East 44th
1	2	cd759a988b8f23924b5a2058d5ab2b49	2016年6月14日	odern bathroom!	East 56th Street
1.5	3	53a5b119ba8f7b61d4e010512e0dfc85	2016年6月24日	/dryer in the ap	Metropolitan Avenue
1	0	bfb9405149bfff42a92980b594c28234	2016年6月28日	uarantors Allowe	East 34th Street

features	latitude	listing_id	longitude	manager_id	photos	price	street_address	interest_level
'Dishwasher', '	40.7108	7170325	-73.9539	0843d78c784171a1.	jpg', 'http	2400	45 Borinquen Pla	medium
uilding', 'Dish	40.7513	7092344	-73.9722	77af4f40004820b4.	jpg', 'http	3800	230 East 44th	low
ing', 'Laundry i	40.7575	7158677	-73.9625	7b766204f08e613c.	jpg', 'http	3495	5 East 56th Stre	medium
[]	40.7145	7211212	-73.9425	2d0489da1b5f2c42	11212_c1785	3000	Metropolitan Av	medium
'Fitness Center'	40.7439	7225292	-73.9743	88fbb5234d8a1e88	jpg', 'http	2795	0 East 34th Stre	low

图 1 原始数据示例

首先我们分析原始数据的结构，图 1 展示了训练数据的前 5 行。我们可以发现，在原始数据中包含 15 种原始特征信息，除了一些常规的数值特征外，还有一些描述信息、特征等文本信息、图像信息、缺失信息等，而我们最终所要预测的是不同出租清单的感兴趣程度（interest level），分为 high、medium、low 三种结果，该问题实际上是一个三分类问题。要想最终能够得到较为理想的预测结果，需要尽可能多的挖掘原始数据中的特征，具体步骤在后文中具体阐述。

2.2 数据预处理 1——根据经理 id 提取有用信息

从原始数据入手，我们首先关注经理 ID（manager_id）这一特征，虽然 manager_id 本身没有多余的信息，但是我们可以根据 manager_id 挖掘出经理相关的基础特征信息，并将其扩展到原始特征数据中。特别的，本文挖掘经理推荐的房屋的 interest_level 的比例信息作为额外的特征。首先根据训练数据统计每个经理推荐房屋的 interest_level 的统计信息，之后将训练集与该统计信息进行匹配作为额外的基础特征。其中不同经理的统计信息示例如图 2 所示，每一个出租清单（listing_id）额外的经理相关的特征示例如图 3 所示。

```
building_level

{'a10db4590843d78c784171a107bdacb4': [59, 47, 6],
 '955db33477af4f40004820b4aed804a0': [9, 16, 1],
 'c8b10a317b766204f08e613cef4ce7a0': [71, 19, 3],
 '5ba989232d0489da1b5f2c45f6688adc': [67, 23, 0],
 '2c3b41f588fbb5234d8a1e885a436cfa': [47, 8, 0],
 '30a7951cfc21240e8c46b2d079d292e5': [56, 77, 36],
```

图 2 不同经理的统计信息示例

listing_id	manager_level_low	manager_level_medium	manager_level_high
7170325	0.530120482	0.397590361	0.072289157
7092344	0.352941176	0.588235294	0.058823529
7158677	0.769230769	0.205128205	0.025641026
7211212	0.720588235	0.279411765	0
7225292	0.826086957	0.173913043	0

图 3 经理相关的特征示例

2.2 数据预处理 2——挖掘其他有用信息

本次过程和前一节的过程类似都是对数据进行预处理。不过这次预处理不需要利用训练集的统计信息，只需要根据数据分析可能有用的特征就可以了。本文首先将训练数据和测试数据合并在一起，其中测试数据标签 interest_level 置为'nnnn'用来区分训练集和测试集。

对数据预处理 1 中得到的特征数据进行进一步处理，将照片数量（photos）、特征（features）数量、与中心点的距离（认为离中心点越近越接近市中心），描述的词（discription）长度，创建时间（created）的月、日、小时作为基础特征，并以天为单位对创建时间进行分类，并计算每个卧室（bedrooms）的单价，每个浴室（bathrooms）的单价，每个房间的单价，卧室浴室房间数差，房间总数、卧室占房间数百分比作为基础特征。其中，计算单价时由于除数可能为 1，对分母均进行加 1 的平滑处理。

之后，我们将一些统计特征放入我们的模型中。这里的一部分特征会发生数据泄漏，不过在本次复盘中先忽略这个问题。这里，统计信息首先包括相同描述信息的数量、相同经理的数量、相同房屋的数量、相同街道的数量、相同卧室数的数量、相同浴室数的数量、相同日期的数量。

之后，我们进一步分析潜在的有用的特征。首先我们将经理的活跃天数作为经理的活跃程度；之后统计不同经理负责推荐的房子数量和每个活跃天平均处理房屋的数量，并将每个经理活动的经度差和纬度差的乘积作为经理的活动范围，并记录活动范围内单位面积平均推荐房屋数量；统计相同的房子被多少经理拥有，经理当天发了多少个信息，不同经纬度的类别（每 1 平方千米为一个类）。将新得到的基础特征信息扩展到原始特征数据中，最后得到基础的 39 个特征。前 15 种特征如图 1 所示，新增基础特征信息示例图如图 4 所示。

photo_num	feature_num	distance	num_description_words	created_month	created_day	created_hour	time	price_bed
12	7	6	77	6	16	5	77	1200
6	6	6	131	6	1	5	62	1266.666667
6	6	7	119	6	14	15	75	1165
5	0	7	95	6	24	7	85	750

price_bath	price_bath_bed	bed_bath_dif	bed_bath_per	room_sum	bed_all_per	display_count	manager_count	building_count
1200	800	0	1	2	0.5	5	294	8
1900	950	1	2	3	0.666666667	12	64	110
1747.5	873.75	1	2	3	0.666666667	248	265	158
1200	545.4545455	1.5	2	4.5	0.666666667	54	235	5

street_count	bedrooms_count	bathrooms_count	day_count	manager_active	manager_building	manager_building_post_rt	build_day
8	39608	99086	2671	58	58	0.197278912	1
12	37114	99086	2253	30	44	0.6875	1.466666667
106	37114	99086	1491	43	56	0.211320755	1.302325581
5	18149	1642	1864	36	64	0.272340426	1.777777778

manager_place	midu	building_manager	day_manager	day_manager_rt	building_mean	jwd_class
275.0872	0.210842235	4	24	0.008985399	2400	7195
63.3086	0.695008261	35	7	0.003106968	3758	7597
44.408	1.261034048	69	17	0.011401744	3250	7596
284.9424	0.224606798	1	16	0.008583691	3000	7194

图 4 预处理 2 基础特征示例

2.3 数据预处理 3——进一步进行特征工程

之后，我们进一步分析潜在可能有用的特征。

我们将 listing_id 与时间的斜率作为额外的可能有用的特征，并将本记录价格对应经理开价平均值的差作为开价程度，相应的比值作为开价比例，并进一步提取每个经理开价程度比值均值。

根据经纬度类别统计区域内有多少不同经理竞争，一个经理经营多少区域，该区域内的均价，该区域内的房屋数。

统计每个经理平均描述字数、平均描述的特征数量；根据卧室和浴室数量提取不同房型的均价，并将其与当前房子价格差、总的平均房价价格差作为地理位置的另一维度的信息。

根据经纬度及房型统计该经纬度特定房型的均价，经理的房子出价和均价的差值和比值、同经纬度不同房型的均价的差值和比值，同房型不同经纬度下的均价的差值和比值。

之后，我们进一步提取该经理在该地区出价比的平均值，该房屋在该地区出价比的平均值，该经纬度类别在该市出价比的平均值，该经理拥有的房子的价格均值，该经理拥有的房子地区的均值，该经理拥有的房子均值与该经理拥有的房子地区均值的比值。之后对数据做了一个 kmeans 聚类提取一维无监督特征及不同类别的比率，价格均值。

之后提取总时间，统计经理平均每天工作多少小时，根据每个经理总共在多少点上发了信息，经理总共发出价格总和，经理总共发出卧室总和，经理总共挣了多少钱（将房屋价格差作为经理的收入），经理平均每个卧室挣多少钱，经理平均每天挣多少钱，经理收入和交易量的比值作为投资回报比，经理平均每天的发布额，经纬度附近的房子中价格比这个低的个数及比值，经理开价与附近房子价格低的比值均值，经理的活动范围与市区的平均聚类，经理发帖时间的方差，经理发帖时间的稳定性。之后记录 4 个不同均价和距离的比值关系（采用 5 作为平滑系数）。统计每个经理出现频率最多的前 20 个特征。

接着我们统计每个特征出现的 interest_level 的高低比值作为不同特征的分值并统计分值，统计不同经理的房屋出现零值的次数和比率，记录经理所处经纬度的中位数，每个经理每个房屋的房子数和比率，房间总面积（认为浴室的面积是

卧室的一半），街区与市中心距离，该区域内的 listing 的数量，平均每个房子租出去的房间数量，所处位置和地点的斜率，一公里内房屋 ID 为 0 的数量，listing_id 所处位置与随机六个点的位置的距离、距离最小值和相应的位置，每个清单的图片的平均大小，价格与特征、价格与照片数的比值。

最后，并将字符串进行离散化操作，用 TFIDF 统计描述特征，得到最终特征。

2.4 多分类预测——进一步进行特征工程

之后，我们将所提取的特征放入 xgboost 模型进行多分类。xgboost 是两年前在大数据情况下利用传统机器学习分类回归效果最好的分类器。近年更好的类似原理的分类/回归器是 lightgbm。本文采用 10 折交叉验证，之后对训练集进行训练，训练时根据验证集的损失情况采取早停策略。最后采用 10 个模型的平均结果作为最后的预测结果。

high	medium	low	listing_id
0.117403	0.738788	0.143809	7142618
0.013724	0.011171	0.975106	7210040
0.004128	0.06528	0.930592	7174566
0.269792	0.573105	0.157104	7191391
0.004567	0.229545	0.765888	7171695
0.005227	0.207999	0.786774	7225206
0.000295	0.002427	0.997278	7200075
0.185815	0.541253	0.272932	7145074
0.011258	0.07312	0.915622	7193645
0.000854	0.007954	0.991192	7147703
0.006537	0.090875	0.902587	7182218
0.014612	0.315721	0.669667	7132136

图 5 最后的预测结果

其中，该模型在公开得分 0.50379（冠军混合方案得分 0.49212，得分越低越好）。

三、小组分工

程序设计及编写：卢宇晟、杨丹、王承天
程序调试：卢宇晟、杨丹、王承天
报告撰写：卢宇晟、杨丹
报告修改：王承天

四、作业总结

根据模式识别课上所学内容以及小组成员的意见，选择了 kaggle 比赛——出租清单查询作为研究题目。在确定了选题方向后，经过小组讨论进行了明确的分

工。在本次作业中中，我们先对实际问题做了简要介绍，之后我们从特征工程角度入手，展示 2017 年 kaggle 出租清单查询比赛优胜方案在传统模型的特征工程方法，并展示其单模型的预测结果。

这次大作业让我们真正结合理论与实际，利用了模式识别的相关知识解决了一个实际问题，同时也初步接触了大数据。也学习到了在大数据情况下 python 常用的聚合函数的使用方法，对计算效率有了初步的印象。虽然只是进行了简单的尝试，但是也能够让我们学到许多经验，并且培养了我们对机器学习这一方向的兴趣。感谢赵老师在这一学期的非常认真的教学，并且在作业过程中给了我们大家很多指导。

附：操作说明

1.运行 script.py 提取经理相关特征.

2.运行 feature_tt.py 提取基础特征

3.运行 feature_tt_long.py 提取 4 个需要运行约 4 小时的特征，也可以在 https://github.com/plantsgo/Rental-Listing-Inquiries 找到并下载 timeout.csv 以跳过这一步。

4.运行 xgb.py 得到最后的结果。

5.更多相关信息参见 https://github. com/plantsgo/Rental-Listing-Inquiries。

注：由于数据较大，源数据没有上传，需从 https://www.kaggle.com/c/two-sigma-connect-rental-listing-inquiries 自行下载放入相应的文件夹。