项目中期报告

李苏畅 李杨 罗伟 魏思杰

## **1.项目介绍**

我们组的项目是kaggle上的一个旅馆推荐的题目。项目目标是对已经进行过聚类处理的旅馆类别中预测用户最有可能选择的5组旅馆类别并作出推荐。数据中包含了大量的信息，包括了用户的预定时间，所使用的门户网站，用户所在的国家、城市，用户与目的地距离，用户是否使用移动设备，用户是否携带行李，用户的登入登出日期，用户包含的成人、儿童数量，用户预定的房间数，旅馆所在的大陆、国家、市场，以及用户做过几次类似的预定等数据。

项目介绍链接：

https://www.kaggle.com/c/expedia-hotel-recommendations

项目数据集链接：

https://www.kaggle.com/c/expedia-hotel-recommendations/data

## 2.预处理

图1-1 数据之间的相关性

由于训练数据集的大小达到了4.07G，并且数据集中存在大量的数据缺失，所以我们做的第一个工作是对数据进行清洗。我们首先对数据之间的关系进行分析。

如图1-1所示，数据项之间的相关性较小，所以我们采用人工的方法去除大部分的数据项，然后根据保留的数据项进行分类讨论。

我们通过讨论确定了要使用的数据项，通过提取数据项的特征，我们把要处理的情况分成了5种情况：

1. 如果存在城市和距离的都满足条件的训练数据，我们将会对这种情况的旅馆类别的得分进行累加1。
2. 如果目的地，旅馆所在国家、城市都相同，并且预定时间在2014年，我们同样对这种情况的旅馆类别的得分增加一个常数append\_1。
3. 如果存在目的地数据，那么也同样对这种情况的旅馆类别的累加常数append\_1。
4. 如果存在旅馆所在国家数据，那么累加另一个常数append\_2。
5. 否则仅仅对该类别的旅馆累加得分1。

对其他情况的数据暂时没有纳入分数计算的范围，将会在下一步的工作中进行考虑。每种类型的情况的得分计算完成后，数据预处理的工作就完成了。

## 3.计算结果并输出

首先对上一步得到的分数进行top5的计算。得到计算结果之后根据测试数据集不同程度的数据缺失情况进行分析并且计算最可能的5个旅馆类别。在计算中如果多次得到相同的旅馆类别则自动忽略，后文不再赘述。依次使用上一步骤中的5种情况的得分，将得分最高的旅馆类别作为推荐的类别，直到获得5种旅馆类别。并将结果作为测试用例的结果进行输出。

## 4.实验结果和计划

本项目的得分计算公式是。其中U是用户事件的数量，P(k)是k时的预测值，n是预测的旅馆聚类的数量。我们算法的得分是0.49480。预测的效果与期望相比还有很大的距离（排名大概900名）。我们计划在下一阶段修改分数计算方式，通过找到更合理的分数计算方法来提高准确率。同时我们也将重新对数据之间的相关性进行分析，来找到更有效的特征，并利用特征来改进算法。