同济大学软件学院 2019 年春季数据挖掘课程作业 2

一、数据集

本次作业采用的数据集主要基于用户在 Yelp 和大众点评的个人信息及行为痕迹，包括用户的头像数据、用户名、用户评论数据、用户异常行为数据、用户上传图片数据等。两类数据集具体格式如下：

1. 大众点评用户数据: hw2_data/dianping/

1) 用户头像数据

该数据集包括来自大众点评网站 11,948 位用户的头像数据，dianping_avatar.zip 文件夹包含所有的头像图片，图片名称为 userid_avatar.jpg，如 128884_avatar.jpg。

2) 用户名数据

该数据集包括来自大众点评网站 11,948 位用户的用户名。文件命及数据格式如下：

- dianping_username.csv

user_id	username
user_id1	username1
user_id2	username2

3) 用户评论数据

该数据集包括来自大众点评网站 11,948 位用户的评论数据，评论总数为 572,773 条，评论以文本形式给出。数据集格式如下，comment_id 为评论 id，user_id 为发表评论的用户的 id，shop_id 为评论对应的商家 id，comment_rank 为用户对商家的打分，content 为用户评论内容。

- dianping_review.csv

comment_id	user_id	shop_id	comment_rank	content
comment_id1	user_id2	shop_id1	comment_rank1	content1
commen_id2	user_id3	shop_id2	comment_rank2	content2

4) 异常用户数据

该数据集包括来自大众点评网站 11,948 位用户是否为行为异常（例如发表大量恶意评论，对同一家店短时间内评论态度差异较大等）用户的数据，其中 is_malicious 有 0 和 1 两种值，0 代表用户行为正常，1 代表用户行为异常。文件名以及数据格式如下：

- dianping_malicious_user.csv

index	user_id	is_malicious
0	user_id1	label1
1	user_id2	label2

5) 用户性别数据

该数据集为大众点评网站 11,948 名用户的性别数据，其中 1 代表 female，0 代表 male，文件名和数据格式如下：

- dianping_gender.csv

user_id	gender
user_id1	label1
user_id2	label2

2. Yelp 用户数据

1) 用户名数据

该数据集包括了来自 Yelp 网站的 223,699 位用户的用户名，数据集格式同大众点评用户名数据集格式。文件名如下：

- yelp_username.csv

2) 用户评论数据

该数据集包括了来自 Yelp 网站的 223,699 位用户共计 2,865,907 条评论数据，评论以文本格式给出，数据集格式如下

- yelp_review_text.csv

review_id	user_id	text
review_id1	user_id1	text1
review_id2	user_id2	text2

3) 用户消费品类数据

该数据集包括了来自 Yelp 网站的 223,699 位用户消费商家的品类信息，用户消费数据集在第一次作业的 user_business_223699.json 数据集中给出，商家品类信息在第一次作业的 business_163665.json 数据集中给出。

4) 用户发表图片数据

该数据集的具体图片数据及信息在第一次作业中的 photos_64048_user_719 及 user_photo_719.json 中已给出。

注：在现有数据前提下，如果有需要可以自行进行数据集的扩展。例如根据点评用户的 uid 适量爬取发表的图片，根据 yelp 用户的 uid 适量爬取头像等。（扩展与否不影响评分）

二、任务

分别利用上述大众点评或 Yelp 的数据集完成对用户的性别预测实验，实验任务如下：

1. 特征选取：上述数据集为用户性别预测提供了多种基本的特征。在这些基本特征的前提下，自由选取特征进行处理或者组合，形成新的特征从而进行下一步预测实验。选取基本特征时要适当考虑所选取特征的含义，例如 malicious 标签对用户性别预测是否有帮助。对选取的特征也可以进行进一步的处理，例如对用户评论进行情感分析，词量统计或者主题提取；对用户购买商品的类别进行划分等；对用户发表的图片进行处理；将用户名转化为词向量等。

2. 分类器：根据任务 1 生成的特征，分别采用以下分类器，调整参数进行实验。其中 Neural Network 为可选项，其他三项为必选项：

   A. Logistic Regression

   B. SVM

   C. Random Forest Classifier

   D. (Optional, BONUS) Neural Network (ReLu, Softmax)

   在实验过程中需要将数据集进行划分，采用训练集，验证集和测试集的划分方式进行分类实验。

3. 实验结果评估：对所用到的每个分类器，计算其在测试集上的 Accuracy, Precision, Recall, F1-Score 值。

4. 综合以上实验，分析结果并完成实验报告。

注：

1. 本次实验结果涉及到的所有数据集均可在 http://10.60.43.58:9579/hw2_data/上进行下载

2. 提供的用户数据集是随机选取的，可根据实际情况进行筛选和处理，实验中不一定要用到全量数据。

3. 作业中介绍的特征不一定要全部用到，可根据实验结果进行选择。

4. 分类器中的 **Neural Network** 实现为可选项，不完成不影响评分，完成可有额外加分。

5. 由于 Yelp 数据集没有提供标签，如果选择该数据集需要自行解决标签问题，并且有额外加分。

三、提交

提交日期：**2019-4-27 23:59:59**，提交至 Piazza。提交内容要求：

提交文件命名为学号_姓名(中文)_hw2.zip。共有两个子目录，对应两个任务，命名为 dianping，yelp，每个子目录包括以下内容：

1. 源代码文件。
2. README 文件，介绍运行环境和运行方式。
3. 实验报告文件，包括数据预处理、特征提取、特征选取，分类器参数、实验结果、相应的图表以及对实验结果的比较分析等。
4. 实验结果文件。大众点评的预测结果或 yelp 的预测结果各一个文件，均为 csv 文件格式。每行具体内容如下：

   Dianping: [user_id, predict_label, true_label]

   Yelp: [user_id, predict_label, true_label]