香港Airbnb数据分析与机器学习建模

情景导入

现在人们出行旅游除了酒店，还会选择利用Airbnb等软件寻找房源进行短租。通常，人们在选择短租房源的时候希望能选到一间性价比高的短租，并且会关注一个房源的评论数量。所谓性价比高，就是短租房源的位置、房间的设施、房间类型、价格等综合来讲综合比较好的。

本项目中，我们将会详细的剖析来自官方的Airbnb数据，对不同的数据进行分析及可视化，并且在这个基础上利用机器学习技术训练出适当的模型，使得之后能根据自己指定的一系列特征值（即房源的各种标准，如房源位置、房间类型等）来预测出相应的价格。这样既能节省找房的时间又能提高找到适合自己房源的效率。此外，我们还会利用机器学习技术对各个年月的评论数进行分析建模，并且能做到预测未来各个月份的评论数量。

学习目标

- 学习使用Numpy、Pandas库对表格数据进行处理
- 学习使用Matplotlib、Seaborn库对处理好的数据进行数据可视化
- 学习使用sklearn库训练模型并进行预测

项目步骤

1. 导入相关库

首先，我们需要将项目用到的Python第三方库和内置模块一次性导入到InnoLab中：

```
import os
from pathlib import Path
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error, r2_score
```

其中，这些库包含处理文件路径用到的`os`、`pathlib`库，数据处理用到的`numpy`、`pandas`，数据可视化用到的`matplotlib`、`seaborn`，机器学习工具模块`sklearn`。

2. 对csv表格数据进行分析及建模

项目中，对实现各个功能的代码段都进行了封装，在学习时可以直接运行`main.py`来得到各个功能函数的运行结果。如果要进一步学习，也可以在InnoLab中打开每一个Python脚本文件，详细学习每一个功能函数的内部实现，以全面掌握整个项目用到的技术点。

这里，我们主要说明一下调用`main.py`实现的一系列功能和产出结果：

功能一：加载calendar.py对房源在不同日期的价格做数据分析及可视化

实现代码如下：

```
# load calendar.csv and print the first 5 lines
calendar = load_calendar()
print(calendar.head())

# format price and date in calendar.read_csv
df1 = format_calendar_price()
print(f"formated calendar price dataframe:\n{df1}")
df2 = format_calendar_date()
print(f"formated calendar date dataframe:\n{df2}")

# get mean price sorted by month or weekday
month_mean_price = get_mean_price(sorted=month)
weekday_mean_price = get_mean_price(sorted=weekday)
print(f"mean price sorted by month: {month_mean_price}")
print(f"mean price sorted by weekday: {weekday_mean_price}")

# given specified price to filter houses
df3 = filter_price(500)
print(f"price lower than HK$500 houses:\n{df3}")

# plot mean price sorted by month or weekday
plot_mean_price()
```

功能二：加载listing_detailed.csv对房源价格的影响因素做数据分析及可视化

实现代码如下：

```
# load listings_detailed.csv and print the first 5 lines
listings_detailed = load_listing()
print(listings_detailed.head())

# add minimum cost column to the listings_detailedngs dataframe
df4 = add_min_cost_column()
print(f"add minimun cost column to the listings_detailed dataframe:\n{df4}")

# add room type column to the listings_detailed dataframe
df5 = add_room_type_column()
print(f"add room type column to the listings_detailed dataframe:\n{df5}")

# quickly preprocess listings_detaied dataframe and select certain colomns
```

```
listings_detailed_df = preprocess()

# plot room type distribution
plot_room_type()

# count room type and location distribution, and plot room location
distribution
df6 = count_room_distribution()
print(f"room type and location distribution dataframe:\n{df6}")
plot_room_distribution()
```

功能三：加载reviews_detailed.csv对不同年月用户评论数量做数据分析及可视化

实现代码如下：

```
reviews = load_reviews()
print(reviews.head())

# plot reviews sorted by date
plot_reviews_by_date()

# count each month reviews numbers in all years, and plot
df7 = count_year_month_reviews()
print(f"year-month reviews dataframe:\n{df7}")
plot_year_month_reviews()
```

功能四：对listing_detailed.csv中房源价格影响因素进行特征处理，并利用sklearn训练模型及进行预测

实现代码如下：

```
# process features data, split training and testing data, and fit the model
predict_price_df = predict_house_prices()
print(f"predict testing dataseta houses prices and compare them with the true
target prices:\n{predict_price_df}")
```

功能五：根据reviews_detailed.csv中的过往年月用户评论数训练模型并预测未来月份的用户评论数

实现代码如下：

```
# predict latest months reviews number
predict_reviews_df = predict_reviews_numbers()
print(f"predict the future months reviews numbers:\n{predict_reviews_df}")

# plot original and pridicted reviews numbers
plot_final_reviews()
```