Web 信息处理与应用Lab3

PB17111649 常鑫鑫 、 PB17111594 杜艺帆

实验要求

数据来源为豆瓣电影的评分记录，根据训练数据中的用户评分信息，判断用户偏好，并为测试数据中user-item 对进行评分。

方法简述

使用基于模型的协同过滤方法。主要思路为将user-item矩阵分解为物品属性矩阵Q与用户偏好矩阵P，并使得这两个矩阵的乘积与评分矩阵R尽可能接近，在得到矩阵P与Q后可得到完整的评分矩阵R，从而做出预测。

运行环境

python3.7

需安装pandas及numpy包，安装方法为：

`pip install numpy`、`pip install pandas`

项目结构非常简单，仅有一个`data`文件夹存放数据及代码文件`main.py`，直接运行即可

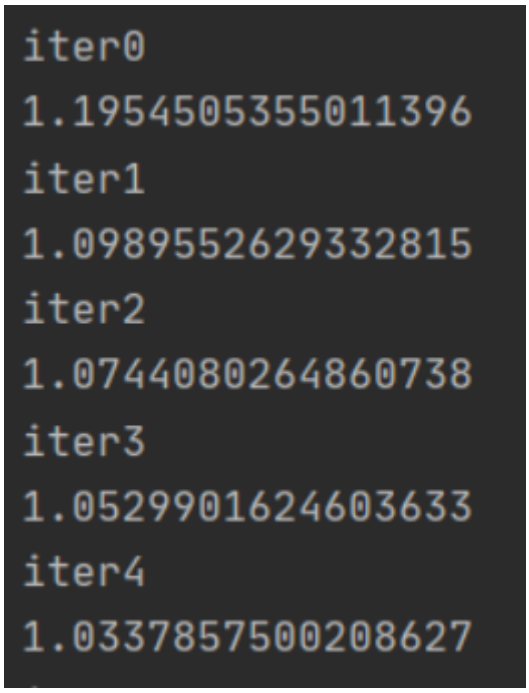
关键函数

```
def train(self):
        # 使用随机梯度下降法
        P, Q = self.init_matrix()

        current_err, last_err = 0, 1000
        err_limit = 0.001
        for i in range(self.epochs):
            print("第%d次迭代"%i)
            error_list = []
            for uid, iid, r_ui in self.dataset.itertuples(index=False):
                v_pu = P[uid]
                v_qi = Q[iid]
                err = np.float64(r_ui - np.dot(v_pu, v_qi))

                v_pu += self.alpha * (err * v_qi - self.beta * v_pu)
                v_qi += self.alpha * (err * v_pu - self.beta * v_qi)

                P[uid] = v_pu
                Q[iid] = v_qi
```

```
                error_list.append(err ** 2)
            current_err = np.sqrt(np.mean(error_list))
            print(current_err)
            if current_err > last_err or abs(current_err - last_err) <
err_limit:
                break
            last_err = current_err
            self.alpha *= 0.9

        return P, Q
```

实验中遇到的问题

- 刚开始参数设置不当，导致噪声矩阵E = R – PQ所对应的2-范数很快溢出，参考资料后设置步长α为0.02，λ = 0.01，迭代次数30
- 按照如上设置，在迭代几次后仍出现err溢出的情况，考虑可能是由于步长α仍过大导致出现发散的情况，但是若将步长α设置过小，则起始收敛速度太慢，搜索资料后得知较为实用的方法，即每次迭代将步长α变小，如乘以0.9，可以既保证起始收敛速度，又减少发散的可能性
- 按照30次迭代次数，输出每次迭代的噪声error值如下，发现刚开始噪声值逐步减小，在第16次迭代时达到极小值，之后又开始上升，说明在第16次迭代附近已接近收敛，故增加迭代终止条件：当两次噪声值差值小于某阈值（0.001）时或当前噪声值大于上一次时终止



```
iter14
0.9534542860542733
iter15
0.9526111167863904
iter16
0.9524922497652956
iter17
0.9530166512484555
iter18
0.9541138926300687
iter19
0.9557212089471223
```

```
iter27
0.9803275139416793
iter28
0.9840440000325069
iter29
0.987714048482456
932993
```

实验结果

经过数次调整参数，达到的最优结果为1.68

PB17111649-9.txt	1.681777627975588