机器学习

第8次实验：
使用协同过滤推荐算法进行电影推荐

实验目的

01

会用Python创建协同过滤推荐模型

02

会用协同过滤推荐模型给用户推荐感兴趣的电影

03

会用测试集评价推荐模型的准确率

实验内容

使用pyspark.mllib.recommendation的ALS类对海量电影数据集建立协同过滤推荐模型；用测试集验证推荐模型的准确率，要求RMSE小于1；对任意用户推荐指定部数电影

数据集：ratings.csv

变量说明：

- ✓ 第1列：用户编号
- ✓ 第2列：电影编号
- ✓ 第3列：电影评分
- ✓ 第4列：评分时间

实验内容

1

载入数据集，按照6:2:2把数据集分为训练集、验证集和测试集

```
training_RDD, validation_RDD, test_RDD =
small_data.randomSplit([6, 2, 2], seed=10)
```

2

使用训练集训练协同过滤推荐模型，使用验证集进行验证，显示最佳秩和最小误差

```
model = ALS.train(training_RDD, rank, seed=seed,
iterations=iterations, lambda_=regularization_param)
```

实验内容

3 使用最佳秩重新训练协同过滤推荐模型，使用测试集对模型进行测试，显示最小误差

```
rates_and_predictions = test_RDD.map(lambda r: ((int(r[0]), int(r[1])), float(r[2]))).join(predictions)
```

4 预测用户189对电影2598的评分，显示结果

```
predictedRating = model.predict(user_id, movie_id)
```

5 对用户385推荐10部电影，显示结果

```
topKRecs = model.recommendProducts(user_id, movie_num)
```

执行结果图例

```
 1.加载评分文件.....
 2.按照6:2:2分为训练集、验证集、测试集.....
 3.设置协同过滤推荐算法ALS（交替最小二乘法）参数....
 4.训练模型，确认最佳秩值（rank），确认最小误差.....
最佳秩值： 4
最小误差RMSE： 0.940384976800151
5.用最佳秩值重新训练模型.....
6.使用测试集对模型进行测试.....
7.计算测试集最小误差RMSE.....
测试集模型最小误差RMSE = 0.946063748424406
8.预测用户对电影的评分.....
用户编号:23 对电影:1704 的评分为:3.8206872803843406
9.向某一用户推荐10部电影.....
向用户编号:25的用户推荐10部电影:
Rating(user=25, product=83411, rating=5.325548991265791)
Rating(user=25, product=80, rating=4.950690342222783)
Rating(user=25, product=72647, rating=4.792994011243717)
Rating(user=25, product=8511, rating=4.792994011243717)
Rating(user=25, product=25764, rating=4.792994011243717)
Rating(user=25, product=31547, rating=4.792994011243717)
Rating(user=25, product=44587, rating=4.792994011243717)
Rating(user=25, product=5059, rating=4.792994011243717)
Rating(user=25, product=7074, rating=4.792994011243717)
Rating(user=25, product=3357, rating=4.763950090204592)
```