

期中大作业题目及要求



负责教师：胡俊峰

2023/04/19



主题：基于观影数据集的数据分析与挖掘

数据环境（可任选其中一个）：

- 1、movielens数据集
- 2、豆瓣电影数据

movielens数据集

- 基础数据集，包含rating，user，movies三个表。6000用户，3700部电影。
 - 有用户年龄-职业-性别标签。有电影名称-类型-发布时间-主演明星等信息
- 扩充数据集：MovieLens 20M Dataset 约2万部电影。有简单的intro信息。评分数据本次作业暂不使用。

1 movies_info

movie_id		name	genre	release_time	intro	directors	stars
0	1	Toy Story (1995)	Animation Adventure Comedy	22 November 1995 (USA)	A cowboy doll is profoundly threatened and jea...	John Lasseter	Tom Hanks Tim Allen Don Rickles
1	2	Jumanji (1995)	Adventure Comedy Family	15 December 1995 (USA)	When two kids find and play a magical board ga...	Joe Johnston	Robin Williams Kirsten Dunst Bonnie Hunt
2	3	Grumpier Old Men (1995)	Comedy Romance	22 December 1995 (USA)	John and Max resolve to save their beloved bai...	Howard Deutch	Walter Matthau Jack Lemmon Ann-Margret
3	4	Waiting to Exhale (1995)	Comedy Drama Romance	22 December 1995 (USA)	Based on Terry McMillan's novel, this film fol...	Forest Whitaker	Whitney Houston Angela Bassett Loretta Devine
4	5	Father of the Bride Part II (1995)	Comedy Family Romance	8 December 1995 (USA)	George Banks must deal not only with the pregn...	Charles Shyer	Steve Martin Diane Keaton Martin Short
...

作业要求：（单人组）

Task1：在movielens 1M的数据集上，统计分析观影的性别偏好。

需要完成：

- 综合观影信息、评分信息，设计合理方案分别筛选出前20部比较流行的（rating > 300）男性/女性 偏好电影。
- 针对不同类型的电影（genres），统计分析男/女偏好程度（需要做归一化），通过双色直方图对比显示。

Task2：在movielens 1M的数据集上，通过观影及评分信息，预测观众的年龄-性别

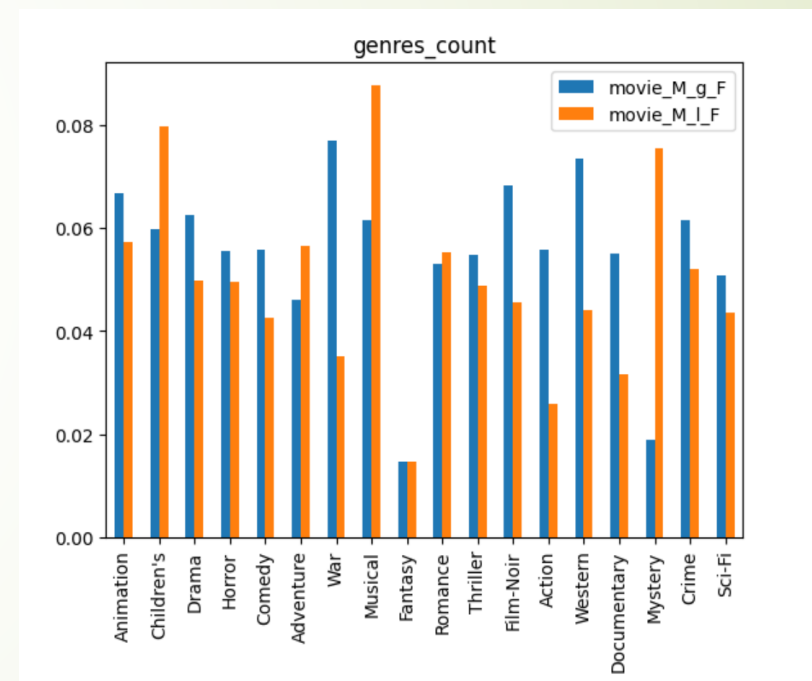
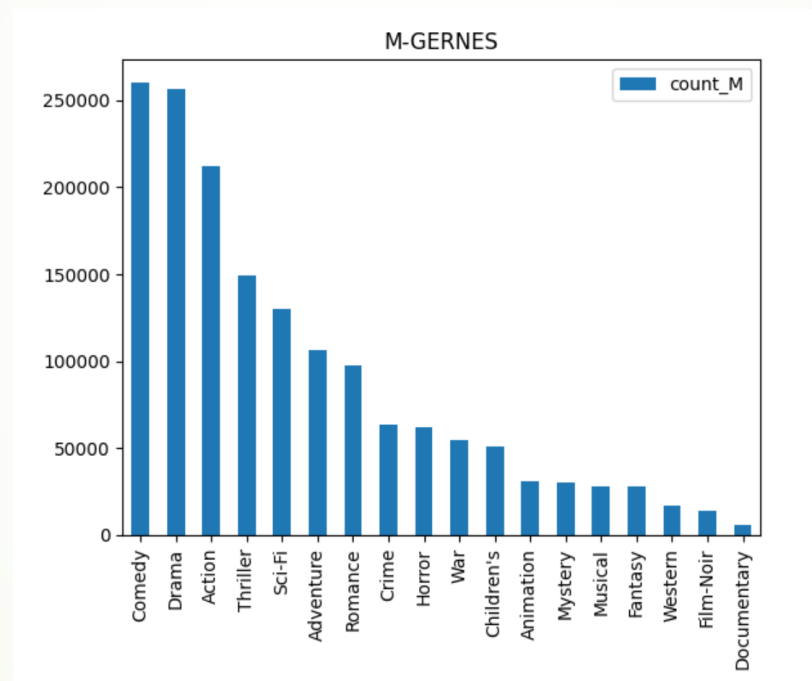
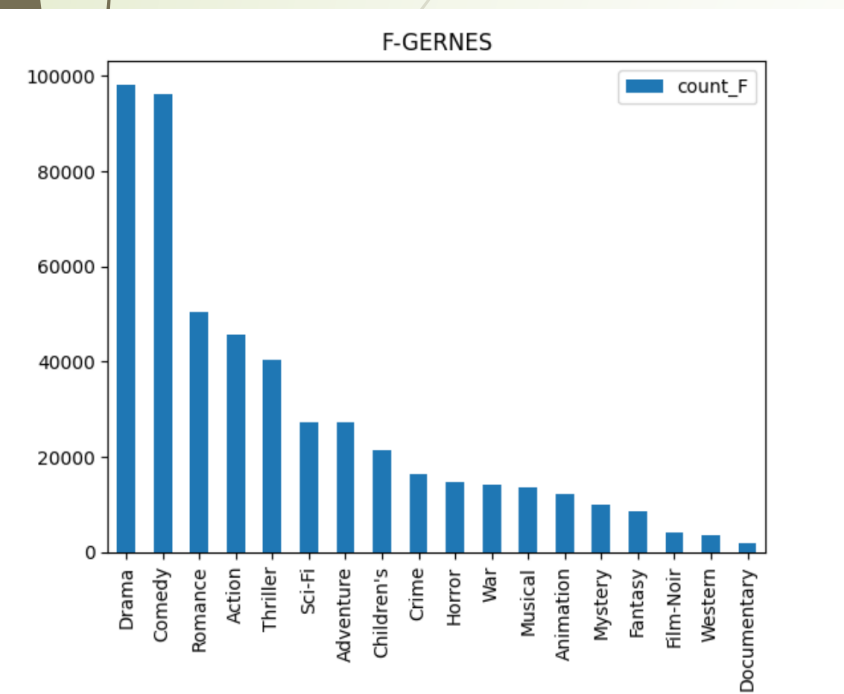
需要完成：

- 拆分训练集-测试集（20%评测），实现评测方案（准确率-召回率）
- 实现分类器模型，对观影数超过100的用户进行预测。调整模型及参数。包括并不限于特征降维来获得较好的效果。（提示：在用户年龄预测问题中，由于年龄段本身是具有一序关系的。常规的模型优化方法不一定会有明显的效果，有兴趣的同学可以看一下ordinal regression模型。有余力可以尝试，不算分。）

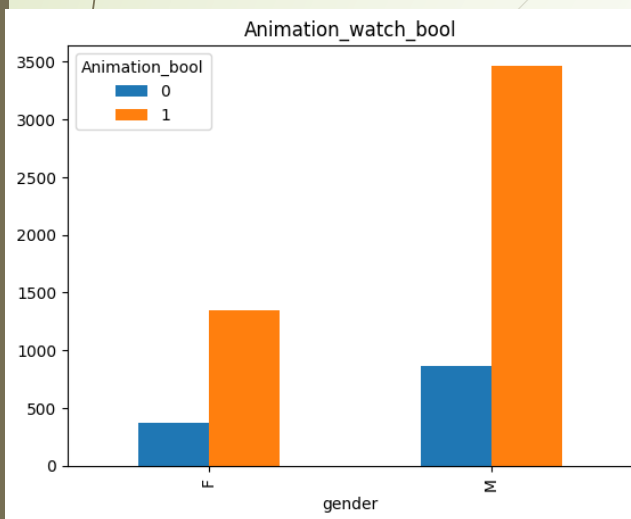
Task3：在movielens 1M的数据集上，通过观影-评分及电影简介等信息，对观影>100的用户实现用户画像。

- 包括且不限于：最喜欢-最不喜欢的电影类型。输出3-5部代表性的电影反映该用户的观影偏好。（可以通过对偏好的电影集合运用图分析技术或SVD分解来实现）
- 自定义一些合理的类型概念，如，家庭主妇最爱，烧脑神剧等，对用户进行标签标记。或者综合电影风格，生成用户观影偏好的雷达图。生成用户偏好词云等。

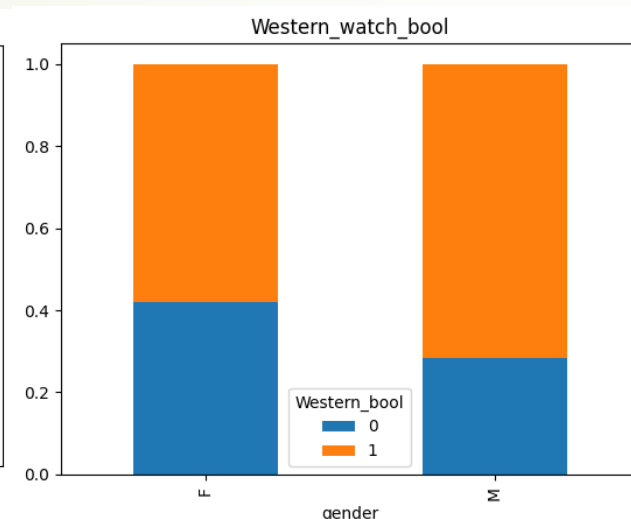
几个示例 (by 姜和丰助教)



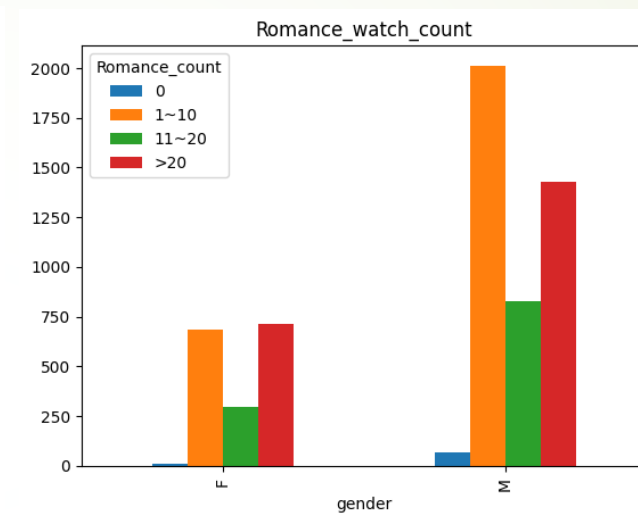
几个示例（by 姜和丰助教）



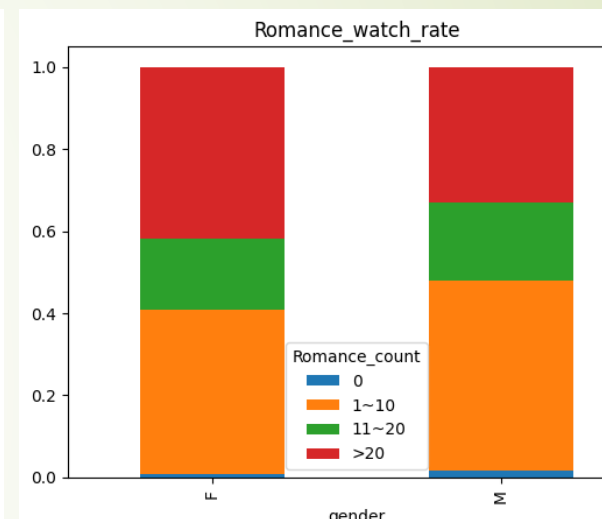
不同性别用户是否过观看、
Animation电影




不同性别用户观看Western
电影的比例



不同性别用户观看
Romance电影数量的统计



不同性别用户观看Romance
电影数量的比例



作业报告要求：

- 简述每个任务用到的模型包括关键计算公式。同时粘贴代表性的运行输出及可视化结果、运行评测结果图示。
- 如果尝试采用了不同方案进行实现，最好有效果的对比分析。
- 自定义类型概念部分最好有比较详细的思考（intuition）说明。

作业评分要求：

- 1、基本实现要求的功能 (task 1-3) 。 70%
- 2、正确使用向量化计算提升效率。清晰的大作业报告。 20%
- 3、完成实现task3要求 10%
- 4、通过数据分析自定义一些合理的类型概念，并对用户进行标注 5%
好的可视化，对所得结果进行深入的分析，得出符合直觉或可验证的结论 10%

某个方面有优秀实现被大作业讲评引用或邀请报告交流的都会有适当的加分。

电影的海报数据：

```
1 plt.figure(figsize=(14,10))
2 for i in movies_info[:6]['movie_id']:
3     plt.subplot(1, 6, i)
4     poster_i = cv2.imread('data/poster/' + str(i) + '.jpg', 1)
5     poster_i = cv2.cvtColor(poster_i, cv2.COLOR_BGR2RGB)
6     plt.imshow(poster_i)
7     plt.title(movies_info.iloc[i-1]['name'], size=9)
8     plt.xticks(())
9     plt.yticks(())
10 plt.show()
```

Toy Story (1995)



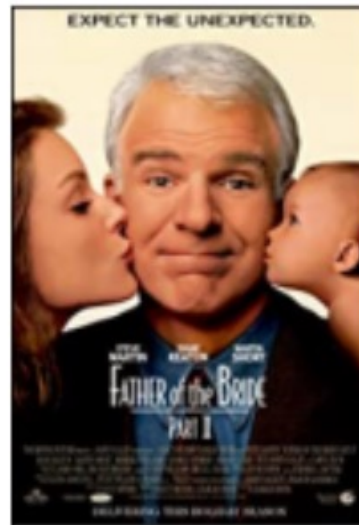
Jumanji (1995)



Grumpier Old Men (1995)



Waiting to Exhale (1995) Father of the Bride Part II (1995)



Heat (1995)



双人组作业（第二人）任务要求（待续）

- 在单人组作业的基础上，利用海报数据进行特征分析与类型化。通过且不限于：色彩、明暗分布，画面主题内容等不同角度提取分析海报的情感及主题类型。70%
- 与已有的电影标注信息（含自定义标签标注信息）进行对比分析。得出符合直觉或可验证的结论。20%
- 报告内容符合当前学术规范 5%
- 结合海报信息与用户画像，实现更好的电影推荐。5%
- 双人组作业（第二人）需要提交单独的大作业报告和运行代码文件。

超人组作业要求（可以1-3人）：

任选一即可，用深度学习模型实现的可以替代期末大作业：

- 生成一段文本描述，实现对用户的画像。
- 要求：文本表达流畅，涵盖主要信息。
- 生成一幅图片，对用户实现观影兴趣画像
- 要求：画面符合直觉，内容表达用户特点
- 评分：基本完成功能，模型设计合理，报告内容清晰90%
实现效果好10%
- 提前邮件报名：hujf@pku.edu.cn

豆瓣电影数据

- 基础数据集，包含comments，person，movies，ratings，users五个表。
- 电影聚类分析部分只使用其中的movies表。该表包含21个字段，38188个电影

1 data = pd.read_excel('./movieLens-douban/data/movies.xls')

2 data

✓ 9.3s

Python

	movie_id	name	alias	actors	cover	directors	douban_score	douban_votes	genres	imdb_id	...	mins	official_site	regions	release_date	slug	storyline	tags	year	actor_ids	director_ids
0	26670818	情定河州	情定临夏天使然	王博/吴佳尼/王姬/高丽雯/郭力行/尹哲/沈丹萍/罗中旭/臧金生/罗刚/屈文沛/阎青妤	NaN	尹哲	0.0	0	剧情/爱情	NaN	...	0	NaN	中国大陆	NaT	RmNQQeyzb	电影《情定临夏天使然》讲述临夏新一代青年人发奋图强、借助国家一带一路战略励志创业的故事。	甘肃/临夏/伊斯兰/中国/2016/中国大陆/烂片/宣传伊斯兰教的电影	2049	王博:1313262 王姬:1275275 高丽雯:1325661 郭力行:135...	尹哲:1326188
1	25815002	我不是李小龙	NaN	谷尚蔚/吴孟达/曾志伟/杜海涛	NaN	洪金宝	0.0	0	动作/爱情	NaN	...	0	NaN	中国大陆	NaT	EZnVfiNYf	桀骜不驯的如龙武功高强，在一场比赛中，被打成重伤，被诊今生不能再用心功夫。女友荆兰为激发他重新...	穿越/华语	2049	谷尚蔚:1330813 吴孟达:1016771 曾志伟:1002862 杜海涛:1313024	洪金宝:1055887
2	26392287	曼哈顿中国女孩	NaN	NaN	NaN	NaN	0.0	0	剧情	NaN	...	0	NaN	中国大陆 / 美国	NaT	NuUvEJnzb	平民女孩李莉只身初入曼哈顿求学，在历经迷失与困惑之后，凭借努力与善良收获了事业上的成功，同时...	NaN	2049	NaN	NaN
3	26695995	绿毛水怪	NaN	NaN	NaN	梁栋/吴国樞	0.0	0	爱情	NaN	...	0	NaN	中国大陆	NaT	rqaqyb6ea	王小波经典中篇小说《绿毛水怪》将改编电影。《绿毛水怪》是王小波早期手稿作品，以天马行空的想象...	小波/王小波/爱情/小说改编/文学改编/剧情/中国/2017	2049	NaN	梁栋: 吴国樞:
																	1932年上海虹				

作业要求（电影聚类分析）（单人组）

- Task1：数据预处理和特征向量化。在这个任务中，我们需要对原始数据进行预处理，将文本数据转换为特征向量。
- 需要完成：
 - 读取数据：从给定的数据文件中读取电影数据。
 - 清洗数据：处理缺失值、异常值等，对“简介”中的中文数据进行分词，英文数据进行清洗。
 - 使用TF-IDF对“类型”和“标签”进行向量化：将文本类型和标签数据转换为数值特征向量，并用PCA降维
 - 使用Word2Vec对分词后的“简介”进行向量化：将电影简介数据转换为数值特征向量。
- Task2：特征融合和降维。在这个任务中，我们需要将不同的特征融合在一起，并使用PCA降维。
- 需要完成：
 - 归一化数值特征：对数值特征进行归一化处理，使其在相同的取值范围内。
 - 特征融合：将不同的特征向量组合成一个整体特征向量。（建议特征：TF-IDF结果、WV结果、电影年份、豆瓣评分）
 - PCA降维：通过主成分分析（PCA）对特征向量进行降维处理。（可设置设置累积方差贡献率阈值确定留下多少主成分）
- Task3：K-means聚类。在这个任务中，我们需要使用K-means聚类方法对降维后的特征向量进行聚类。
- 需要完成：
 - 肘部法确定最佳簇数量：通过计算不同簇数量下的误差平方和，确定最佳簇数量。
 - 使用最佳簇数量进行K-means聚类：根据最佳簇数量，对降维后的特征向量进行聚类。
 - 分析聚类结果：输出每个簇的代表电影、簇中电影的数量、评分分布等，也可对结果进行可视化展示。

作业要求（电影聚类分析）

- Task 4：基于电影的 Embedding 和表中其他信息，为导演和演员生成 Embedding，并进行无监督分类（最好不要用 K-means，可以尝试层次聚类、DBSCAN、GMM 等方法）。然后使用可视化的方法，分析 2-3 个属于不同类别导演的特点
- 需要完成：
 - 计算导演和演员的 Embedding
 - 对导演、演员分别进行无监督分类
 - 分析聚类结果：输出每个类的代表人物、每个类的人数、导演电影数分布等，也可对结果进行可视化展示。
 - 使用可视化的方法，分析 2-3 个属于不同类型的导演、与 2-3 个属于不同类型的演员的特点

示例效果——电影聚类效果（by杨礼铭助教）

Cluster 0:好莱坞大片

代表电影：The Teller and the Truth

观看人数最多的 5 个电影：

	NAME	DOUBAN_VOTES
28630	源代码	525750
33563	恐怖游轮	466665
19654	催眠大师	285840
23784	搜索	268758
36152	谍影重重 3	258957

簇中电影的数量：1799

平均评分（忽略 0 分电影）：5.68

评分标准差（忽略 0 分电影）：1.51

评分最高的 3 个电影：

	NAME	DOUBAN_SCORE
36152	谍影重重 3	8.8
13674	斯隆女士	8.7
30723	焦土之城	8.6

各类型的电影数量：

悬疑	1799
剧情	807
惊悚	645
恐怖	220
爱情	204
动作	181

Cluster 1:华语大片（包括大陆港澳台），不如 cluster2 优秀（但评分最高的三个都能 cluster 边缘的点恰好评分高，也可能在豆瓣平台上华语电影观看人数更多，cluster 特征可能有 bias）

代表电影：斗鱼

观看人数最多的 5 个电影：

	NAME	DOUBAN_VOTES
31781	窃听风云	211151
14844	战狼	186045
23224	金蝉脱壳	168415
14968	澳门风云 2	119597
31559	线人	113040

簇中电影的数量：2496

平均评分（忽略 0 分电影）：5.20

评分标准差（忽略 0 分电影）：1.31

评分最高的 3 个电影：

	NAME	DOUBAN_SCORE
33507	WWE Smackdown 十周年精华集	8.8
32703	混乱日	8.5
16475	View from a Blue Moon	8.3

各类型的电影数量：

动作	2496
剧情	768
喜剧	430
惊悚	400
爱情	179
恐怖	171
古装	108
武侠	79
战争	74
运动	55

Cluster 3:“烂片”，烂的很标准的那种

代表电影：Studio Illegale

观看人数最多的 5 个电影：

	NAME	DOUBAN_VOTES
15964	纯洁心灵·逐梦演艺圈	84842
31212	嘻游记	20480
16117	从天“儿”降	19465
9058	麻辣学院	486
5212	上位 2	404

簇中电影的数量：1925

平均评分（忽略 0 分电影）：2.88

评分标准差（忽略 0 分电影）：0.51

评分最高的 3 个电影：

	NAME	DOUBAN_SCORE
26140	贫民英雄	3.7
16117	从天“儿”降	3.6
31212	嘻游记	3.3

各类型的电影数量：

喜剧	1925
剧情	657
歌舞	37
运动	15
战争	8
情色	5
儿童	4
历史	3
古装	2

dtype: int64

电影年份均值：2013.41

Cluster 4:恐怖片

代表电影：死亡游乐场

观看人数最多的 5 个电影：

	NAME	DOUBAN_VOTES
3682	昆池岩	119497

簇中电影的数量：102311

平均评分（忽略 0 分电影）：65731

评分标准差（忽略 0 分电影）：49614

评分最高的 3 个电影：

簇中电影的数量：3322

平均评分（忽略 0 分电影）：4.73

评分标准差（忽略 0 分电影）：1.04

评分最高的 3 个电影：

	NAME	DOUBAN_SCORE
14164	无名女尸	7.5
32782	鸡皮疙瘩 NO.6	7
7628	鬼故事	7.2

各类型的电影数量：

恐怖	3322
喜剧	331
剧情	134
情色	26

作业评分要求：

- 1、基本实现要求的功能 (task 1-3) 。 60%
- 2、清晰的大作业报告与过程结果可视化。 15%
- 3、完成task4 15%
 - 可解释性较好的无监督分类结果 10%
 - 好的可视化，对所得结果进行深入的分析，得出较好的演员特点 5%

注意：

- 为了使结果可复现，请尽量在所有需要随机性的地方都设置随机数种子。
- 可以根据自己的理解，在task1-3中使用要求之外的数据项，鼓励在过程中多进行可视化。

作业数据环境及提交作业文件要求：

- 作业所需数据会放在 data子目录打包下发。
- 同时打包的还会有一个user子目录。大家程序里需要较长时间运行生成的中间结果可以保存在这目录里。后面的单元则通过加载该数据继续运行。提交作业时data目录不需要提交。只需要提交代码文件、报告文档及user子目录就行。

作业提交要求：

- 作业提交截止时间：2023年5月4号 中午 11点。（不是晚上23点）
- 提交方式：讲作业的notebook文件，作业报告及user子目录
- 选择movielens数据集的，
 - 单人组用 学号.zip压缩提交（扩展名是.zip）。
 - 双人组作业，用 学号1-学号2.zip提交。其中学号2的为以海报图像处理为主同学的学号。
 - 双人组作业由学号1同学提交一份完整作业就行，学号2同学提交自己负责任务的pdf报告和代码文件就行。
 - 多人组作业参考双人组方案进行（需要提前邮件给hujf@pku沟通说明）
- 选做豆瓣电影数据集的，用学号.rar格式压缩提交。注意压缩包格式是rar，不是zip。