Part I

个人任务：在 UCI 数据集上实现分类算法或者聚类算法
每个数据集(https://archive.ics.uci.edu/ml/datasets.html)有详细的介绍
实验过程：
1、在 UCI 数据集中选择一类任务（分类或者是聚类），然后选定某个合适的数据集，要求这个数据集不能太小，否则会影响最终的得分；
2、对你选择的数据集进行简单的描述，比如特征和数据集大小；
3、对数据集进行预处理，具体的预处理的方法可以参照课件 NO. 11（**要求将数据集的最后10%划分为测试集**）；
4、对于分类任务，要求分别实现 KNN 算法和逻辑斯蒂回归（Logistic Regression）算法；对于聚类任务，要求分别实现层级聚类（Hierarchical clustering）算法和 K-means 算法。针对你选择的任务，需要对算法进行描述（可以采用伪代码的形式）；
5、模型评估：根据课件中提到的评估准则（聚类评估标准不少于 3 种），对你选择的任务中使用的两种方法进行性能分析比较，要求有相应的图表分析给出；
6、编程语言不限（**注意，如果直接调用机器学习的包，不予计分**）
实验报告要求：
1）实验报告需要给出你所选定数据集的描述，以及你对数据集所做的预处理的步骤；（10%）
2）实验报告需要你给出针对特定数据集的两个算法的描述（以**伪代码**的形式给出）；（20%）
3）由于数据集为开放选取，所以对模型的评估需要你在测试集上进行，给出相应的图表分析。在聚类任务中，对于 K-means 算法，需要你给出不同 K 值的选取以及对应的结果分析，对于层级聚类，需要你给出预设的簇的个数 K，以及对应的结果分析。（20%）
提交内容：
1）数据预处理的源码（10%）、两个算法的源码（20% + 20%），
以 python 实现分类算法为例，分别为 preprocessing.py、knn.py、logistic_regression.py
源码单独保存在一个文件夹“codes”中！
2）实验报告，保存为 pdf 格式（50%）。
以上内容置于文件夹“实验二_part1”中

Part II 小组任务(**分组要求：每组 2~3 人**)

①推荐系统（Recommendation System）

该部分的实验我们采用了经典的推荐系统数据集：MovieLens。由于 MovieLens 有很多版本，出于对计算力的考虑，本次实验我们采用 MovieLens 1M Dataset 数据集。该数据集的详细描述会在 http://grouplens.org/datasets/movielens/1m/中的 README 中给出。
任务描述：
1、预处理：下载得到的文件包括三个部分：ratings.dat, users.dat, movies.dat。其中，ratings.dat 是核心文件，其余两个文件作为补充信息给出。为了便于处理，不同的数据集可以通过关键字进行合并。另外，在进行模型训练前，需要在数据集中随机选择一部分作为测试集。
2、任务描述：
1）根据 ratings.dat 建立评分矩阵（ratings matrix），**分别**使用 Content-Based Methods 和 Collaborative Filtering Methods 来补全矩阵；

Tips：**使用这种标准方法只能获得基本分**。我们的数据集给出了附加的数据 users.dat 和 movies.dat，所以可以尝试根据这些数据提高模型的性能。这里给出一个思路仅做参考：通过用户评分过的电影来找到电影所属的流派，进而分析用户的流派喜好，根据用户的流派喜好寻找最近邻，从而做出预测。

2）测试：获取测试集上的预测结果。

3）评估：使用不少于 4 项的评价标准（例如 RMSE、MAE、Precision 和 Recall）来评估模型在训练集上的性能。同样，需要给出相应的图表分析。

4）编程语言不限（**注意，如果直接调用机器学习的包，不予计分**）

3、根据实验的结果，对这两种方法的优缺点（Pros & Cons）做出比较。

4、进一步的思考：课上在最后提到了现在的推荐系统任务中，仍然存在很多的挑战，比如多样性（Diversity）、准确性（Accuracy）以及扩展性（Scalability）。

针对多样性问题，在个性化的推荐系统中，多样性有 3 个方面的含义：个体多样性、总体多样性和时序多样性。个体多样性（individual diversity）从单个用户的角度来度量推荐的多样性，主要考察系统能够找到用户喜欢的冷门项目的能力；总体多样性（aggregate diversity）主要强调针对不同的用户的推荐应当尽可能的不同，部分研究利用了长尾理论（long-tail theory）来考察推荐系统多样性对产品推荐的影响，多样性程度越高的推荐系统更有利于带动那些不太流行的商品的销售量；时序多样性（temporal diversity）考虑了更加现实的情况，由于新产品的出现、用户兴趣的动态变化或者用户情境的变化，用户偏好会发生改变，将时间情境融入到推荐系统中，构建用户—项目—时间三维推荐模型，可以有效提高个性化推荐的时序多样性。

现在，请你选择一个角度（不限于上述提及的 diversity、accuracy、scalability），做出一些对现有工作的**调研**（鼓励引用参考文献），包括：1）对问题的分析；2）已提出的方法。同时，请选择一个已有的方法或者提出一个新的方法**实现**来提高模型的性能（提出新的改良方法最多可以加分 20%），需要给出**图表分析**。

实验报告要求：

1）实验报告要求给出对数据进行预处理的步骤；（10%）

2）需要详细描述你的算法的流程（以伪代码的形式给出）（这里使用标准的方法进行处理只能获得基本分 10%，满分 20%）；

3）在划分出的测试集上，根据评价标准，给出模型的性能评估结果；（10%）

4）给出两种方法的优缺点分析；（5%）

5）调研内容（问题分析以及已有的一些方法）、对扩展问题实现以及分析；（25%）

6）务必注明**所有组员**的姓名+学号。

提交内容：

1）数据预处理的源码（10%）、两个主体算法（**C**ontent-**B**ased Methods 和 **C**ollaborative **F**iltering Methods）的源码（5% + 5%）、扩展问题的实现（10%）。

以 python 为例，分别为 preprocessing.py、cb.py、cf.py、extension.py

源码单独保存在一个文件夹 codes 中！

2）实验报告，保存为 pdf 格式（70%）。

以上内容置于文件夹“实验二_part2_1”中

②网络中节点影响力及社区发现（Influence & Communities）

背景

建立并分析共同作者网络(Co-author Network)是确定学者学术影响力的重要方式，因为共同作者之间通常有着很强的联系。

20 世纪最有名的共同作者（Co-author）是数学家 Paul Erdös，他与超过 500 名伙伴合作过 1400 多篇学术论文。也许是其到处旅行的生活方式，以及将其大部分积蓄作为学生解决数学难题的奖励的做法，Erdös 的共同作者网络蓬勃发展，以至于很多数学家通过分析 Erdös 庞大的共同作者网络（Co-author Network），来衡量自己与 Erdös 的关系密切程度（见 http://www.oakland.edu/enp/），以在一定程度上确定自己的学术影响力。有趣的是，得益于 Erdös 与 Alfred Rényi 合著的论文《论随机图谱》(On Random Graphs)在 1959 年的出版，他也是跨学科网络科学的奠基人之一。传奇数学家 Erdös 的精彩故事可以参见很多书籍以及网站，例如：http://www-history.mcs.st-and.ac.uk/Biographies/Erdos.html。

本部分的任务建立在 Erdös 的共同作者网络的基础上。

图 1. Erdös 与幼年陶哲轩

任务描述

1. 建立 Erdös 共同作者网络(Co-author Network)

对数据集 Erdos1.html（可以自己去 https://files.oakland.edu/users/grossman/enp/Erdos1.html 下载）做适当的预处理，并建立 Erdös 共同作者网络（Co-author Network）。**注意：网络中一定不要包括 Erdös 本人**。此外，建议使用相关工具，比如 Gephi 软件，或者 Python，R 语言中的相关包，来完成网络可视化，软件会包含不同的可视化算法，会形成不同的可视化布局，你要自己考虑哪些布局是合适的，图 2 是一种可视化结果实例。如果必要的话，可以适当减小网络的规模（比如适当去除一些数据），以利于网络的建立与分析。

再次强调：**网络中无 Erdös**，因为如果把其包含进去，其将与所有节点相连，不利于网络的建立与分析。

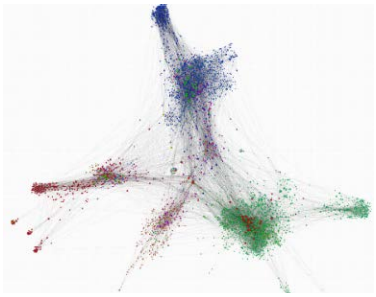

图 2. 网络可视化示例

2. 影响力分析

使用影响力度量标准（如 Degree Centrality、Closeness Centrality）与方法，确定 Erdös 共同作者网络(Co-author Network)中最有影响力的节点，思考谁的论文最有价值或者谁与 Erdös 联系最密切，并比较不同标准与方法下的结果。

3. 社区发现

使用不同的、合适的社区发现方法（如 spectral clustering）对 Erdös 共同作者网络(Co-author Network)进行分割，形成不同社区，比较不同方法的分割结果，并尝试去解释分割出的社区的含义。

实验报告要求

总要求：内容科学、无误、完整、详尽，能对问题做出很好解答，**排版科学，格式美观**。包括但不限于（括号里为评分比重参考）：

1. 报告排版要好，内容充实，要有对做法以及方法的说明，要包括网络建立、影响力分析、社区发现三部分；（>=10%）
2. 要有**必要的数据预处理**，以建立网络。此外，建议完成网络可视化（可最高奖励 15% 的分数）；（20%）
3. 相关分析要使用**多种**（>=2）方法，**方法要能很好地解决问题**，要有方法的说明以及算法伪代码，并比较不同方法的结果，并且要有对结果的分析；（40%）
4. 报告的完成要建立在代码实现（编程语言不限，**数据挖掘算法的实现不准调包**）的基础。要求提交源码；（30%）
5. 务必注明**所有组员**的姓名+学号。

提交内容

提交实验报告，源码以及其他必要的附件。

1）实验报告，保存为 pdf 格式，命名要清晰易懂。

2）源码单独保存在一个文件夹“codes”中，要包括三部分（每部分单独一个子文件夹或者文件）：1. 数据预处理以及建立 Erdös 共同作者网络的代码。2. 影响力分析源代码，包括不同方法对应的源码。3. 社区发现源代码，包括不同方法对应的源代码。文件命名以及代码要清晰易懂。

3）其他所有附件（如果有的话）单独保存在一个文件夹“supplements”中，文件命名要清晰易懂。

以上内容置于文件夹“实验二_part2_2”中。

提交方式：

1）截止时间：2019 年 1 月 18 日 23:59

2）将所有实验内容打包压缩为“学号+姓名.zip”发送至邮箱 xu_lab@sina.com

3）由于 Part II 部分是小组实验，所以每位同学在提交时，如果你是组长，打包文件里需要包含 Part I 和 Part II 两部分，其余同学仅需提交 Part I 即可

温馨提示：

1）Part II 的小组组员成绩一样，所以在分组时，你们需要分配及协调好各自的工作内容；

2）Part I 和 Part II（含①②两部分）的最后得分在实验总分里占比为 3 : 4 : 3；

3）如果实验中遇到问题，可以联系课程助教

xins@mail.ustc.edu.cn xdjcl@mail.ustc.edu.cn