实验二 / Experiment 2

请于 2021 年 5 月 11 日 23:59 之前提交至课程邮箱 ustcweb2019@163.com
并于 2021 年 5 月 12 日（周三）课上进行现场汇报

总体实验要求：
请组成 4-6 人小组，围绕指定数据集进行社交决策分析实验，记录实验过程并撰写实验报告。

数据背景：
数据集来自著名在线活动组织网站 Meetup，其机制为：

- 整个 Meetup 社区由若干社团组成，每个社团有若干名用户，用户可以随时加入或退出。
- 活动以社团为主体，由若干名社团成员发起组织（部分活动组织者缺失）。仅有社团成员会收到邀请。
- 社团成员可以选择是否参加活动（Yes/No/Maybe），但不是所有人都会回应。
- 部分活动会注明限制人数（Headcount），但不一定会起到约束作用。

实验数据：
数据中所有元素编号经过哈希处理，包括社团（G 开头）、活动（E 开头）、成员（M 开头）、主题（T 开头）四种类型，具体包含以下三部分内容：

（1） Group-Topic：社团信息
该文件用于描述社团主题，每个社团对应 2 行，各个元素之间用**空格**分隔，具体格式如下：
社团 ID 社团发起人 ID
社团主题 ID1 社团主题 ID2 社团主题 ID3…
例如：
G174 M41736
T109 T1991 T2151 T7929 T9753 T15052 T17407 T18240 T19163

（2） Member-Topic：成员信息
该文件用于描述成员画像，每个成员对应 2 行，各个元素之间用**空格**分隔，具体格式如下：
成员 ID
成员主题 ID1 成员主题 ID2 成员主题 ID3…
例如：
M19524
T17866 T17030 T8652 T4392 T15117 T491 T1122 T130 T131 T133 T235

（3） Group-Event：活动信息
该文件用于描述活动响应情况，每个社团对应一个文件，其中列出了该社团组织的所有活动，并包含了**活动发生的时间**。
每个活动对应 5 行，各个元素之间用**空格**分隔，具体格式如下：
活动 ID 活动限额 时间（时间戳格式）
活动组织者 ID1 活动组织者 ID1…
选择 Yes 的用户 ID1 选择 Yes 的用户 ID2 选择 Yes 的用户 ID3…
选择 No 的用户 ID1 选择 No 的用户 ID2 选择 No 的用户 ID3…

选择 Maybe 的用户 ID1 选择 Maybe 的用户 ID2 选择 Maybe 的用户 ID3…
例如：
E47 11 1239370200000
M24444
M24444 M33638 M38989 M42317 M62021 M66864 M68349 M68660 M69527
M36257 M4372 M10604 M29227 M35326 M36958 M38481 M40515 M41215 M42249
M60073

实验内容：
要求对于指定数据，预测列表中用户对于活动邀请的回应（Yes/No/Maybe），并**自行**根据真实数据验证算法效果。
要求进行五折实验，即每次采用 80%的数据作为训练数据，20%的数据作为测试数据，重复五次直至覆盖全体数据，并汇报五次实验各自的效果及平均值、方差。数据的切分方式由自己决定。

具体实验内容包括：
（1） 不考虑社交因素的预测（基本推荐）
由于数据本身没有提供社交信息，请先将每位用户视作相互独立，进行个性化的决策预测。
该部分要求对以下内容进行分析（并不限于以下内容）：
- 社团在个人决策中的作用
- 成员标签对于决策的影响
- 个人偏好随时间推移的演化现象
- 冷启动用户的有效处理（即测试数据中的成员未在训练样本中出现）
- 不同验证指标的相互比较

可选内容包括：
- 数据切分方式对于效果的影响
- 参数敏感性讨论（如果涉及）
- 活动限额的影响与匹配问题

（2） 考虑社交约束的预测（社交推荐）
为实现社交约束下的预测，首先需要自行构造社会网络。常见的构造方式如根据成员之间的标签相似性、成员共同参加的社团或者成员共同参加的活动等进行构造（可以采用相似性或共现次数进行加权）。
该部分要求对以下内容进行分析（并不限于以下内容）：
- 不同网络构造方式对于结果的影响
- 网络结构演化对于结果的影响
- 网络是否加权对于结果的影响
- 社交约束与社交级联设计下的效果对比

可选内容包括：
- 不同社交约束设计方式对于效果的影响
- 网络稀疏性问题的有效处理
- 参数敏感性讨论（如果涉及）

提交说明：

以 PDF 或 DOC 格式提交，实验报告提交文件及邮件标题命名格式统一为“社会计算第一次实验报告_学号_姓名”。

- 例如：“社会计算第二次实验报告_SA20011999_法外狂徒张三”
- 标题仅写明小组内一位成员学号及姓名即可，其他成员请在文中注明学号及姓名。
- 因未署名造成统计遗漏责任自行承担。
- 实验报告请务必独立完成，如果发现抄袭按零分处理。
- 请注明所采用的算法，并列举必要的参考文献。
- 请采用必要的图表以更清晰地展示实验结果。
- 提交报告的同时请提交**源代码**以供检查。
- **除非特殊情况并事先征得许可，否则迟交报告将不再被接收，并取消答辩资格**。

报告要求：

由组长进行汇报，汇报总时长为 8 分钟，包括 6 分钟 PPT 讲解与 2 分钟提问+点评。

- 报告内容应包括两个部分各自的数据处理方式、社会网络构建方式、采用的预测/推荐模型、相关参数的设置、实验结果及其分析、组内成员分工等。
- 报告顺序按照实验报告接收的顺序为准，名单将在报告当天于课程群内公布。
- 三位助教将根据汇报内容和实验报告内容进行打分，并计入总评成绩。

如有未尽事宜，将对本说明进行进一步更新。

**********English Version**********

Please submit your report to ustcweb2019@163.com, before 23:59, May 11th, 2021.
Also, please report your experiments during the last class, May 12th, 2021.

General Requirement:

Please form a team with 4-6 members for the prediction of social-related decision-making, with recording the experimental details for the report.

Background:

The data set was extracted from the famous event-based social network service “Meetup”.

- The whole community is formed by several groups, each group contains multiple members, and all the members could join or quit the groups.
- All the events are organized by certain groups, in which some group members will act as “organizers”. Only group members will be invited.
- The group members could response to the invitations as Yes/No/Maybe, but usually, only a few of members will reply.
- Some events will have an explicit headcount limitation, but it may not limit the participators.

Experimental Data Set:

All the records in data set have been hashed, including four types, namely group, event, member,

and topic. The details are introduced as follows:

（1） Group-Topic

To describe the topics of groups. Each group corresponds to 2 lines, and factors are separated by "blank space", with the format as:
GroupID Group-organizerID
Group-topicID1 Group-topicID2 Group-topicID3…
For example:
G174 M41736
T109 T1991 T2151 T7929 T9753 T15052 T17407 T18240 T19163

（2） Member-Topic

To describe the topics of member. Each member corresponds to 2 lines, and factors are separated by "blank space", with the format as:
MemberID
MemberID1 MemberID2 MemberID3…
For example:
M19524
T17866 T17030 T8652 T4392 T15117 T491 T1122 T130 T131 T133 T235

（3） Group-Event

To describe the responses to event invitation. One group corresponds to a file, in which all the organized events are summarized and labeled with timestamp. Each event corresponds to 5 lines, and factors are separated by "blank space", with the format as:
EventID EventHeadCount EventTimestamp
OrganizerID1 OrganizerID2 OrganizerID3…
YesID1 YesID2 YesID3…
NoID1 NoID2 NoID3…
MaybeID1 MaybeID2 MaybeID3…
For example:
E47 11 1239370200000
M24444
M24444 M33638 M38989 M42317 M62021 M66864 M68349 M68660 M69527
M36257 M4372 M10604 M29227 M35326 M36958 M38481 M40515 M41215 M42249
M60073

Details for Experiments:

It is required to predict the response of individual members for given events (Yes/No/Maybe), and measure the performance by the ground truth in data set.
The 5-fold experiments are required (i.e., for each round, 80% of records will be treated as training data, and the rest 20% as test data. 5 rounds will be executed to cover all the records as test data). In the report, the 5 results, as well as the mean value and variance results should be reported.

Details includes:

（1） Prediction without Social Factor: Basic Recommendation

All the members will be treated as mutual independent. Required tasks include (but not limited to):

- The effect of group in decision-making
- The effect of topics in decision-making
- The evolution of individual preference
- Solution for cold-start problem (i.e., new members in test data)
- The comparison of different measurement indexes.

Optional tasks include:

- The effect of data division
- Parameter sensitiveness (if applicable)
- The effect of event headcount and matching problem

（2） Prediction with Social Factor: Social Recommendation

Since no social network is provided in the data set, it is required to build a network by yourself. Common-used techniques including measuring the similarity of members via topics, common groups, or common events, which could be weighted.

Required tasks include (but not limited to):

- The effect of different network building strategies
- The effect of evolving network structure
- The effect of weighted/non-weighted network
- Comparison between social-constraint and social cascade.

Optional tasks include:

- The effect of different social constraint
- Solution for network sparcity
- Parameter sensitiveness (if applicable)

Submission Instruction (if <u>ALL</u> the team members are international students <u>ONLY</u>):

- The name of submitted files, as well as the title email, should be "Social Computing Exp2"
- Name/student ID of all the members should be listed in the submissions.
- Any kind of plagiarism will be strictly prohibited.
- Please clarify the utilized techniques, and list related references.
- Necessary figures/tables are required to represent the experimental results.
- **<u>Source code</u>** should be submitted simultaneously.
- **<u>Late submissions will not be accepted, and the final reporting will be disqualified</u>**.

Details for Representation:

The team leaders are asked to represent in the final class. It will spend 6 minutes to report with PPT, and 2 minutes for Q&A.

- The report should include data pre-processing, solution for building social network, prediction/recommendation models, parameter setting, experimental results and analysis, task distribution for members, etc.
- The order of report will be announced on May 12th, 2021 based on the order of report received.