Data Science Term-Project Rules

Prof M.-S. Chen & Prof M.-L Lo

Outline

- Schedule
- Rules
- Grading
- Introduction of Datasets

Schedule

Date	Todo
10/31 (Fri.) 09:10	final project rules announce
11/07 (Fri.) 23:59	team makeup due
11/18 (Tue.) 23:59	proposal due
12/01 (Mon.) 23:59	Presentation slides due
12/02 (Tue.) 09:10	project presentation 1
12/09 (Tue.) 09:10	project presentation 2
12/21 (Sun.) 23:59	project report due

Rule - Team

- 2-3人一組, 請於2025/11/7 Fri. 23:59前在以下網址填寫分組資訊

  [https://docs.google.com/spreadsheets/d/1oJysa4XbHqMV6_WP7gbUI75npKhoBBA8wCPVFE3CQDY/edit?usp=sharing]

  (期限後仍未填寫, 助教會幫忙分組)

Rule - Proposal

- 以下四個資料集擇一
  - Learning social circles in networks [link]
  - Web traffic forecasting [link]
  - Santander Bank Product Recommendation [link]
  - News Category Dataset [link]
- 根據所選資料集，題目自訂
- 請於2025/11/18 Tue. 23:59前在NTU COOL上傳Proposal (組長繳交即可)
- Proposal 格式: [template]
- 檔名 : group<group_id>_proposal.pdf (e.g. group1_proposal.pdf)

Rule - Use Generative AI

- 允許使用, 但不強迫使用
- 以輔助為主, 其僅能參與其中一小部分, 而非用來進行預測的核心技術
- 使用方式:呼叫線上API或本地部署皆可
- 使用情景舉例
  - 將資料分析結果包裝成prototype
    Ex. 預測完銀行推薦產品後, 依據客戶資料客製化產生產品推薦報表(「因為你有每年穩定收入..., 我們建議理財產品A, 可以幫助達成..., 每年預估回報...」)
  - 模擬使用情景
    Ex. 對新聞資料進行分群後, 使用LLM產生虛擬觀眾資料, 驗證新聞資料分群結果

Rule - Presentation slides submission

- 此檔案為Oral Presentation時使用, 報告時長為八分鐘, 請自行斟酌頁數
- 請於 2025/12/01 (Mon.) 23:59前 繳交
  - 上傳COOL 作業區: “Final Project Presentation”
  - 組長繳交即可
- Presentation slides
  - 檔名 :group<group_id>_slides.pdf (e.g. group1_slides.pdf)
  - 包含 : 題目, 組別, 組員

Rule - Oral Presentation

- Dates: 12/2 (Wed.) & 12/9 (Wed.)
- 每組 10 min.（8 min. 報告 ＋ 2 min. Q&A）
  - 請提早 15 min. 到教室測試檔案（ short video + slides ）
  - 使用自己筆電的組別也請提前試投影設備

Rule - Report submission

- 請於 2025/12/21 (Sun.) 23:59前 繳交
  - 上傳到 COOL 作業區: "Report"
  - 組長繳交即可
- Report (書面報告, 不是投影片!!!)
  - 檔名 :group<group_id>_report.pdf (e.g. group1_report.pdf)
  - 包含 : 題目, 組別, 組員 (學號及姓名)
  - 10-12頁, 中英文不限, 論文格式 1 column or 2 columns 皆可

Grading

- Project = Proposal (5%) + Oral presentation (10%) + Final Report (18%)
- The term project accounts for 33% of the overall your class grade.
- We will evaluate your work based on
  - Problem definition clarity: How clearly you define the research objectives
  - Methodology correctness: Whether your approach is appropriate and correctly executed
  - Analysis and results: Quality of experiments, data analysis, and depth of findings
  - Originality and contribution: Novelty and meaningful contribution of your work
  - Presentation quality: Organization, storytelling, visual design
  - Report completeness: Structure, grammar, references, and overall readability
  - Team collaboration: Work division and contribution from each member

Dataset 1 : Learning social circles in networks [link]

- Facebook user profile & friends
- Files
  - egonets.zip - A list of the user's friends (UserId: Friends)
  - features.zip - Facebook profiles (UserId feature1 feature2 feature3 …)
  - featureList.txt - Discription of features (birthday, education, work, ...)
  - Training.zip - Connected social cirles (circleID: friend1 friend2 friend3 …)
- Reference
  - Leskovec, J., & Mcauley, J. J. (2012). Learning to discover social circles in ego networks. In Advances in neural information processing systems (pp. 539-547) (http://i.stanford.edu/~julian/pdfs/nips2012.pdf)

Dataset 2 : Web traffic forecasting [link]

- Include 145k time series in training datasets.
  - Each of these time series represent a number of daily views of a different Wikipedia article
- Files
  - train_*.csv - contains traffic data. This a csv file where each row corresponds to a particular article and each column correspond to a particular date.
  - key_*.csv - gives the mapping between the page names and the shortened Id column used for prediction
  - sample_submission_*.csv - a submission file showing the correct format

Dataset 3 : Santander Bank Product Recommendation [link]

- 2015-01-28～2016-06-28, 1.5 years of customers behavior, to predit what new products customers will purchase.
- Files
  - train_ver2.csv.zip - products columns #25-#48
  - Test_ver2.csv.zip
    - products : column #25-#48
- task examples:
  - product recommendation systems
  - Statistical analysis over potential consumer groups.
  - Analysing cusomer's behavior.

Dataset 4 : News Category Dataset [link]

- Contains around 210k news headlines from 2012 to 2022 from HuffPost
- Files
    - News_Category_Dataset_v3.json: contains all the data with features such as category, headline, date etc.
- Task Examples
    - News category classification
    - News correlation analysis
- Reference
    - Rishabh Misra (2022), News Category Dataset (https://arxiv.org/abs/2209.11429)

Thanks for your attention：）