CISC 372: Project description

Learning objectives

By the end of the project, students should be able to:

- apply a data science software tool to solve a real-life problem, and
- analyze the results obtained from the data mining software tool.

Your team should identify a data analytics problem in the areas of arts, social sciences, humanities, engineering, business, security, or in our everyday life, and then utilize a data mining software tool to solve the problem, showing your novelty and contributions.

Please refer deadlines to our online calendar

Procedure

1. **Find a partner in the class.** You have the option to work alone, but the same evaluation standard will be applied on the presentation and the final report. Under certain circumstance three people is allowed.
2. **Understand the objectives of the basic data science functionalities**, namely classification analysis, frequent patterns mining, association rules mining, cluster analysis, outlier analysis, etc. We will cover the details of these functions in the rest of the course. Yet, at this stage, you at least need to understand their objectives in order to formulate a research problem for your project.
3. **Identify a data mining problem and collect relevant data.** You can collect the required data by yourself and get the data from any other sources. The followings are some data repositories:
    - UCI Machine Learning Repository
    - UCI KDD Archive
    - National Geophysical Data
    - US Government Data
    - UK Government Data
    - The World Bank Data
    - Kaggle

    (Hint: If you do not have any idea how to get started, you may start by browsing the UCI/Kaggle datasets.)
4. **Implement your solution** using your chosen DM software tool.

Deliverables

1. **Submit your project proposal**. The proposal should include:
    - Your names (match the names in onQ)
    - E-mail addresses
    - Title of the project
    - Background and motivation

- An informal description of your problem. (The problem you attempt to solve, question you attempt to answer, functions you attempt to implement, etc.)
- General direction of your solution. (for example, are you going to use classification? Are you going to use clustering? You can do multiple!!)
- Describe the potential datasets to be used in experiments. How will you measure the performance of the proposed method? Are you going to measure the performance in terms of accuracy, efficiency, scalability? What's the reason?
- Length: 1-2 pages. Maximum 2.
- Submit a softcopy of your proposal via OnQ.
- **The project proposal will be graded as Pass or Fail.**

2. **Give a presentation and submit slides**.
   - Submit a softcopy of your slides on OnQ before the presentation class.
   - Every team has 10 minutes (3 minutes for switching computer, 13 minutes for presentation, and 2 minutes for Q&A). You need to explain what you have done in the project **as a story**, including the problem description, challenges, method, experimental results, etc.
   - The presentation order will be randomized.
3. **Submit your final report**.
   - Submit the link to your GitHub repository through OnQ (only latest commit before deadline will be evaluated) that contains your code, links to the dataset (or you can upload datasets on the *Release* tab), the slides, and the final report in PDF or Word.
   - [Report] The report is limited to 12 pages. The report may contain the following sections:
     - Abstract – 4 sentence summary (not more, not less, **exactly 4**)
     - Introduction
       - Try to **motivate** the research problem using a real-world scenario.
       - Informal description of the problem. You may consider using an **example** to describe the problem.
       - Summarize the challenges and **contributions/impact** of the paper.
     - Problem Statement
       - Clearly define your problem.
     - Proposed Method or Solution
       - An overview of your solution.
       - Detailed steps of your solution.
     - Experimental Results (if any)
     - Summary or Conclusions
   - I understand that **the problem and solution may evolve (change) as the project develops.** Therefore, it is fine if the details of the final paper are different from the details described in the project proposal.

Evaluation criteria

- **Originality, scientific contributions, and impact**

    - Are you solving a new problem? What are the strengths of your proposed solution?
    - If some materials in your paper and source code come from other papers, books, and any other sources, you must cite them properly.
- **Correctness** of applied methodology
    - Are you using the right DM method for the intended objective?
- Experimental results, findings, conclusions
    - Show and **explain** the experimental results. Provide an analytical explanation of the results, regardless it is good or bad. Show your effort. Is the actual result same as what you have expected? If the result is bad, say with low accuracy, do you have any suggestions to improve it? It is fine to show bad results. **You will not get a lower mark because of showing bad results.**
- Clarity of the final report
- Challenges of the proposed problem: The instructor will compare the level of difficulty of your project with other teams.
- Note: **You cannot reuse any projects** from other courses and submit it to this course. This violates the Code of Student Conduct and Disciplinary Procedures.

Project ideas

Here are some potential project ideas to start thinking about. You are encouraged to propose your own topic, rather than using the topics below.

Text document classification
Text document clustering
E-mail classification (spam detection, topic classification)
Data mining on fashion data
Predict disease classes using genetic microarray data
Web log mining (predict the next webpage that will be visited by the visitor)
Social networks mining
Opinion mining on tweets (e.g., classify whether a tweet is positive or negative on a product)
Income prediction