Data Mining and Machine Learning (Group Coursework)

F21DL (2024/2025)

In this group coursework, you and your team will work on a data mining and machine learning project using GitHub as your collaborative platform.

The project will cover various aspects of data mining and machine learning, including basic concepts, exploration and data analysis, generative models, discriminative learning, and practical application of these techniques. Sometimes we will refer to this project as your "Machine learning portfolio"

The project is designed to give you hands-on experience with real-world data and machine learning algorithms while also teaching you how to effectively collaborate.

Note: Check with your course coordinator on the exact tools you will be using, such as GitHub for version control, documentation, code reviews, regular commits, testing, and data cleaning.

1. Coursework Objectives

1. Gain a solid grasp of the fundamental concepts in data mining and machine learning, including datasets, dealing with missing data, classification, and supervised vs. unsupervised learning.
2. Explore generative models such as naïve Bayes, probabilistic graphical models, and cluster analysis (including k-means clustering and the EM algorithm).
3. Implement discriminative learning methods such as linear models, linear regression, decision tree learning, perceptron, and advanced models like multi-layer perceptron and deep learning architectures.
4. Develop practical skills by working on real-world datasets, conducting data preprocessing, implementing machine learning models, and evaluating their performance.
5. (**Campus-dependent**) Learn how to effectively collaborate with your team members using GitHub for version control, code sharing, and project management. Your coursework documentation and evidence should be hosted on GitHub.

2. Summary of the Group Project

- Each group will select a broad topic of application, some examples are provided in Section 3.1 – Topics, below.
- Then they should look for publicly available real-world datasets (at least one and at most three), suitable for the topic of choice to perform data mining and machine learning analysis (see §3.2.1 to find datasets, at the end of this document for resources, more details are provided in canvas week1 material).
- The group will pitch the project title, identify the datasets, and formulate the research question (with any accompanying hypotheses) in Week 4 via an assessed short presentation (**Deliverable 1)**.
- The project should incorporate both generative and discriminative tasks, showcasing various algorithms discussed in the syllabus that are relevant to the chosen dataset(s).
- The project must include key components such as data preprocessing, model training, performance evaluation, and data visualization.
- Teams should provide a comprehensive documentation detailing their approach, results, discussions, and analysis.
- (**Campus-dependent**) All documentation, commentary, and code must be hosted on GitHub (e.g., README/Jupyter Notebook files/Documentation).
- The final three deliverables are i) a 6-page report in pdf format and a zipped folder of the project code (**Deliverable 2**); ii) a 15-minute mini-viva consisting of a 5-slide 7-minute presentation followed by a Q&A session (**Deliverable 3**), iii) a peer-assessment form using MS Forms (**Deliverable 4 -0marks**).
- The grading is based on the marking rubric (on CANVAS).

Group size: 4-5 people in each group.
Effort: approx. 40+ hours per student
Credit: 40% of overall mark
Coursework timeline: 9th September 2024 (Coursework Released)
Deliverables:

Week 4 – Project Pitch; short presentation to ensure project meets the coursework criteria;
Week 11 – 6-page Project Report and Code.
Week 12 – Project Presentation; 15-minute mini-viva;
Week 12 – Peer Assessment (0 marks); via MS Forms.

NOTE: Assessment will be continuous throughout course (attendance, discussions, checks).

Deadline Submission (Project Report): 15:30 (Uk time) on 22th November 2024 (Week 11)

3. Coursework Overview

The group project at hand involves choosing one topic of application and then selecting and analyzing 1–3 unique datasets while dividing the project into five distinct phases. The project's overarching goal is to explore, analyze, and apply machine learning techniques to these datasets, contributing valuable insights and solutions to relevant research questions. The project emphasizes collaboration, data exploration, and the application of various machine learning methods, including clustering, decision trees, and neural networks.

3.1. Topics

Here is a list of quite broad topics; you can use these as a guideline for choosing a domain of application and then proceed to identify suitable dataset(s).

Important Note: Some topics (indicated with an asterisk *) are considered more advanced, hence you should consult with the teaching team first during the labs before committing:

Entertainment

1. Genre prediction of a song based on metadata/visual (album cover)/audio features
2. Popularity prediction of a song based on metadata/visual (album cover)/audio features
3. Genre prediction of a game based on textual description*
4. Movie/Book/Game Review classification*

Sports

5. Player/Team Rating and Match forecasting

Natural Sciences

6. Quality prediction of a particular food item or beverage
7. Forecasting of natural phenomena such as weather systems, wildfires, or earthquake aftershocks

Medical

8. Disease diagnosis (e.g., cancer) based on metadata/medical imaging data (e.g., mammograms, radiographs)

Misc.

9. Customer (e.g., train/flight/grocery store) satisfaction prediction

3.2. Datasets

Once you have selected a topic, the next step is to select **up to three distinct datasets** (these should be unique to your group project). These datasets will serve as the foundation for the entire project, guiding the research direction and analysis.

It is imperative you ensure that your dataset(s) are relevant to the topic and can complete the course requirements (see below). You must take into consideration dataset complexities and size when planning your project. We suggest you choose at least one image data set (for later part of the course when we talk about CNNs).

Note related to Deliverable 1: Week 4 – Project Pitch. You only need to decide on the datasets to use and formulate the research question and hypotheses. You don't need to make any modelling decisions just yet.

3.2.1. Useful links to find datasets

The following list has some pointers to places where you might get some inspiration for data mining challenges together with associated data such as evaluation criteria and comparative performance data

- https://www.kaggle.com - source of lots of different data mining competitions.
- http://www.drivendata.org - source of lots of different data mining competitions with an emphasis on saving the world.
- http://www.kdnuggets.com/competitions/past-competitions.html - list of past data mining competitions; data and evaluation criteria is likely to be available for many of these.
- http://webscope.sandbox.yahoo.com - publicly available research datasets from Yahoo!
- http://www.kdd.org/kdd-cup - KDD Cup is an annual data mining competition run by ACM SIG KDD; datasets, evaluation criteria, and info previous winners are available (note that the most recent competitions are actually hosted on kaggle.com).
- https://huggingface.co/datasets — consolidation of publicly available image, text, and audio datasets
- http://multimediaeval.org/datasets/ - a range of data and evaluation criteria for different types of data mining problems involving multimedia and multi-modal data.

4. Coursework Requirements

Your group project must use your chosen datasets and complete the following requirements (R1–R5). R1 will be assessed both in deliverables D1 and D2-3, whereas the rest in D2 and D3 only (see Section 5 – Deliverables):

R1. Project Topic, Direction, and Questions
In this phase, the group will define a set of clear questions, and objectives based on the selected topic/datasets. Identify the specific problems or hypotheses to be addressed using the datasets (D1-D3).

R2. Data Analysis and Exploration
Comprehensive data preprocessing and cleaning to ensure data quality. Exploratory data analysis (EDA) to gain insights into the datasets. Visualization techniques to present data patterns and trends (D2-D3). Feature selection and evaluation should be included.

R3. Clustering
Implement an appropriate clustering algorithm to show aspects of the data with similar characteristics. Evaluate the performance of clustering algorithms. Interpret the results and discuss their implications (D2-D3).

R4. Baseline Training and Evaluation Experiments

Apply *at least three (3) machine learning algorithms* to build predictive models, including Decision Trees and two more such as Naïve Bayes, Linear Models, Perceptron and k-Nearest Neighbours. Note your task can be either classification and/or regression, depending on the topic and dataset(s) you choose. Evaluate and compare the models in terms of appropriate metrics for the respective tasks, e.g., accuracy, precision, recall, F-1 or error measures (like RMSE). In the case of Decision Trees discuss their practical applications in the context of the selected datasets/topic (D2-D3).

R5. Neural Networks

Implement neural network models, such as Multi-Layer Perceptrons (MLPs) and Convolutional Neural Networks (CNN), for tasks such as classification or regression. Train and fine-tune the neural network models. Assess the performance of neural network models and compare them with other techniques (D2-D3).

Throughout this group coursework, the project team will work collaboratively to navigate through the various requirement phases, from selecting appropriate datasets to applying advanced machine learning techniques. The project's outcomes will contribute to the understanding of the selected datasets, provide solutions to implementation problems , and demonstrate proficiency in data analysis and machine learning. This structured approach will enable you to master essential data science and machine learning skills.

Each group is responsible for project planning, task allocation among team members, and ensuring efficient communication. Regular interactions with instructors and peer code reviews are encouraged to ensure the success of the group projects. Additionally, ethical considerations regarding data usage and privacy must be strictly adhered to when working with real-world datasets.

Finally, it is highly recommended that your dataset(s) satisfies all the above course requirements. If you are unsure whether your choice will satisfy the above requirements, you should ask as soon as possible.

5. Deliverables

There are three deliverables for this coursework:

D1. **Week 4 – Project Pitch**

A 10-minute presentation (we recommend presenting using a short slide-deck, but it's not compulsory). The group will pitch the project title, identify the datasets, formulate the research question and outline any relevant hypotheses. You should also provide a tentative project plan and timeline (deliverables/stages/milestones). You can describe initial thoughts on which approaches they are going to use (or are thinking of taking). There is no requirement to discuss about models yet. This will be arranged via a booked session, where groups will be allocated a time.

D2. **Week 11 – Project Report and Code**

a. A 6-page report in PDF format with the following sections: Introduction, Related work, Dataset Description and Analysis, Experimental Setup, Results, Discussion, and Conclusion. Your report should cover all course requirements described in Section 4 – R1-R5, explaining your approach in each part, summarizing results in graphs/tables when possible and explaining insightful conclusions drawn.

b. A '.zip' of your entire repository. Your final project should demonstrate a clear understanding of the covered concepts and techniques in data mining and machine learning.

D3. **Week 12 – Project Presentation**

A 15-minute mini-viva comprising a short slide-deck of up to 5 slides (7 minutes of presentation), followed by a Q&A session (8 minutes) conducted by members of staff/lab helpers. This will be your team's chance to present your work and describe approaches you have taken. This will be arranged via a booked session, where groups will be allocated a time. You might be asked to demo your code.

D4. **Week 12 – Peer Assessment**
This is a MS Form for providing a peer assessment review of your group. This peer review form is done privately and **only** the course lecturers will have access to your answers.

Important notes:

- ***Regular attendance in labs is required by all team members*** (actively demonstrate and present work during scheduled sessions)
- ***Any data or resources used for the project must be accessible*** (e.g., downloadable/run/test any submission on another machine)
- ***Consult rubric on canvas for mark breakdown for D1-D3.***

7. Plagiarism

In any coursework, it is imperative that students uphold the principles of academic integrity and ethical scholarship. Plagiarism, the act of presenting someone else's ideas, work, or words as our own without proper acknowledgment, is strictly prohibited. We are committed to producing original and authentic content, and any external sources, whether they be ideas, data, text, images, or any other material, must be appropriately referenced and acknowledged using the prescribed citation style. This not only ensures the credibility of our work but also demonstrates our respect for the intellectual contributions of others. Together, let us maintain the highest standards of honesty and integrity throughout our collaborative efforts.

https://www.hw.ac.uk/uk/students/studies/examinations/plagiarism.htm

9. Resources/Useful Links

The following list has some pointers to places where you might get some inspiration for data mining challenges together with associated data such as evaluation criteria and comparative performance data:

9.1. Getting started with Python

We recommend using conda to create a virtual environment (with Python 3.11) for the project.

- Installation: https://docs.conda.io/projects/conda/en/latest/user-guide/install/index.html
- Managing your environment: https://conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html#activating-an-environment

9.2. Framework recommendations for the course

We recommend the frameworks used in the labs, however if you want to use others like PyTorch and Hugging Face you should feel free to do so. Remember, documentation is your best friend.