Data Mining

Homework 4

Due: 9/1/2023, 23:59

Instructions

You must hand in the homework electronically and before the due date and time.

This homework has to be done by each **person individually**.

Handing in: You must hand in the homework by the due date and time by an email to Gianluca (`decarlo@diag.uniroma1.it`) that will contain as attachment **(not links to some file-uploading server!)** a .zip file with your answers. The filename of the attachment should be `DM_Homework_4__StudentID_StudentName_StudentLastname.zip`;
for example:
`DM_Homework_4__1235711_Robert_Anthony_De_Niro.zip`.
The email subject should be
`[Data Mining] Homework_4StudentID StudentName StudentLastname`;
For example:
`[Data Mining] Homework_41235711 Robert Anthony De Niro`.
After you submit, you will receive an acknowledgement email that your project has been received and at what date and time. If you have not received an acknowledgement email within 2 days after you submit, then contact Gianluca.

The solutions for the theoretical exercises must contain your answers either typed up or hand written clearly and scanned.

For information about collaboration, and about being late check the web page.

Problem 1. Graph-based fraud-detection system
In this homework we will design a fraud-detection system. We will use this dataset: `https://www.kaggle.com/datasets/ealaxi/paysim1`

1. **Data preparation and EDA**
    - Load the financial transactions dataset into a Pandas DataFrame.
    - Perform data cleaning, handle missing values, and encode categorical variables.
    - Conduct exploratory data analysis (EDA) to understand key characteristics of the dataset.
    - Visualize statistics such as transaction amounts, distribution of transaction types, and any other relevant features.
    - Explore the imbalance in the target variable (*isFraud*) and consider strategies for handling it.
2. **Building a graph-based fraud-detection system**
    - Construct a graph representation of the financial transactions data. Nodes represent customers (both originators and recipients), and edges represent transactions.
    - Design a Graph Neural Network (GNN) for fraud detection. The goal is to predict whether a transaction is fraudulent based on the graph structure and transaction features.

- Implement the GNN using PyTorch or DGL. Experiment with different GNN architectures, layers, and hyperparameters.
- Task: Binary classification of transactions (fraudulent or not).
- Divide the dataset into training and testing sets. Evaluate the model performance on the test set using metrics like accuracy, precision, recall, and F1-score.
- Investigate the impact of different features on fraud detection performance.

3. **Model explainability**

- Apply an explainability methodology to interpret and explain the predictions made by the GNN model.
- Choose an explainability technique for GNNs.
- Explain a subset of predictions from the test set and discuss the key features contributing to the model's decisions.
- Provide visualizations and insights into how the model makes predictions, especially in cases of both true positives and false positives.
- Discuss the trade-offs and limitations of the explainability methodology chosen.

Deliverables

- Code for data preparation, graph construction, GNN implementation, and model explainability.
- A report (4–6 pages) describing the data processing steps, graph construction, GNN architecture, model performance metrics, and the model explainability methodology.
- Visualizations of key statistics, the graph structure, performance metrics, and explanations for selected predictions.
- Compare the results of the various parameters that you have tried with traditional machine learning methods for fraud detection.
- Discuss the challenges of detecting fraud in financial transactions, the advantages of using GNNs, and the insights gained from the explainability analysis.