Cyber Data Analytics Assignment 1

INTRODUCTION

One of the big tasks in cyber data analytics is detecting malicious or fraudulent data records. This is a machine learning task that is extra complicated due to several properties of the data and its domain:

- *The large majority (typically over 99%) of the data is benign*
- *Malicious users actively try to hide*
- *The amount of data to learn from is enormous*
- *Often this data is unlabeled*

In this assignment, our focus will be on the first property: how to deal with large class imbalances in machine learning. Dealing with the second property does influence the data and taking it into account will be beneficial for your performance. Developing methods for dealing with this is part of the bonus task. The final two properties are the topic of follow-up assignments.

LEARNING OUTCOMES

After completing this assignment, you will be able to:

- *Correctly apply machine learning methods to real data*
- *Modify machine learning algorithms to deal with class imbalance*
- *Analyze the outcomes of machine learning for fraud detection*

INSTRUCTIONS

Visualization task – 1 A4

Load the fraud data into your favorite analysis platform (R, Matlab, Python, Weka, KNIME, ...) and make a visualization showing an interesting relationship in the data when comparing the fraudulent from the benign credit card transactions. You may use any visualization method such as a Heat map, a Scatter plot, a Bar chart, a set of Box plots, etc. as long as they show all data points in one figure. What feature(s) and relation to show is entirely up to you. Describe the plot briefly.

Imbalance task – 1 A4

Process the data such that you can apply SMOTE to it. SMOTE is included in most analysis platforms, if not you can write the scripts for it yourself. Analyze the performance of at least three classifiers on the SMOTEd and UNSMOTEd data using ROC analysis. Provide the obtained ROC curves and explain which method performs best. Is using SMOTE a good idea? Why (not)?

Classification task – 2 A4

Build two classifiers for the fraud detection data as well as you can manage:

1. A black-box algorithm, ignoring its inner workings: it is the performance that counts.
2. A white-box algorithm, making sure that we can explain why a transaction is labeled as being fraudulent.

Explain the applied data pre-processing steps, learning algorithms, and post-processing steps or ensemble methods. Compare the performance of the two algorithms, focusing on performance criteria that are relevant in practice, use 10-fold cross-validation. Write clear code/scripts for this task, for peer-review your fellow students will be asked to run and understand your code!

Bonus task – 1 A4

Try to add context to your classifier by linking/grouping the transactions based on IP, or card number, or transaction date, or country code, etc. For example, you can first aggregate the data before learning a classifier, or post-process the classifier decisions into improved estimates. Do whatever you can to improve your classifiers performance by not seeing every table row as an individual record, they are linked to others in many different ways.

RESOURCES

Slides from Lectures 1 and 2

"Learning from imbalanced data" paper, by He and Garcia.
"An introduction to ROC analysis" paper, by Fawcett.
"SMOTE: Synthetic Minority Oversampling Technique" paper, by Chawla et. al .
"MetaCost: A General Method for Making Classifiers Cost-Sensitive" paper, by Domingos.

All are made available through Brightspace

Your favorite analysis platform (R, Matlab, Python, Weka, KNIME, ...)

Real-world fraud detection data provided by Adyen, available through Brightspace

PRODUCTS

A small report (max 4 pages, 5 including bonus), and the code used to obtain the results. Both will be assessed using the below criteria.

ASSESSMENT CRITERIA

The assignment will be assessed by peer review. The login details will be provided in the week of the deadline.

Knockout criteria (will not be evaluated if unsatisfied):

Your code needs to execute successfully on computers/laptops of your fellow students (who will assess your work). You may assume the availability of 4GB RAM and a Linux operating system, possibly a virtual machine. Please test your code before submitting. In addition, the flow from data to prediction has to be highlighted, e.g., using inline comments.

Your report needs to satisfy the page limit requirements for the different parts. When working in a data analysis notebook, you have to copy and paste the text and results into a printable document satisfying the requirements.

Submissions submitted after the deadline will not be graded.

The report/code will be assessed using these criteria:

Criteria	*Description*	*Evaluation*
Visualization	*Shows an interesting relationship in the data. The relationship is relevant for fraud detection.*	*0-5 points*
ML Workflow	*The data preprocessing is sound. The flow from data to prediction is correctly implemented. At least three different classifiers have been tested on the data.*	*0-5 points*
Modification	*The machine learning algorithms have been modified at the correct point in the data-prediction flow. The modification and obtained ROC curves are sound.*	*0-5 points*
Performance	*At least 100 fraudulent cases are found in the test data, with at most 1000 false positives. Worse performance means fewer points.*	*0-5 points*
Analysis	*The analysis is correct and the conclusions are reasonable. The conclusions are relevant for fraud detection in practice.*	*0-5 points*
Bonus	*Creative solution, correctly implemented.*	*0-5 points*
Report and code	*The data-prediction flow is clearly described, including preprocessing and post-processing steps.*	*0-5 points*

Your total score will be determined by summing up the points assigned to the individual criteria, and averaging to account for the number of peer reviews. In total 35 points can be obtained in each course assignment, the total number of obtained points will be divided by 90 to determine the final grade. In case one of the reviews is significantly worse than (at least 10 points difference) the others, this review score is not taken into account.

You will receive a penalty of 5 points for each peer review not performed. Significantly different reviews will be subject to investigation. If deemed badly done by the teacher or TA, you will also receive 5 penalty points.

SUPERVISION AND HELP

We use Mattermost for this assignment. Under channel Lab1, you may ask questions to the teacher, TAs, and fellow students. It is wise to ask for help when encountering start-up problems related to loading the data or getting a machine learning platform to execute. Experience teaches that students typically answer within an hour, TAs within a day, and the teacher the next working day. When asking a question to a TA or teacher, your questions may be forwarded to the channel to get answers from fellow students. Important questions and issues may lead to discussions in class.

There is no separate lab session hosted at the university, it is your own responsibility to start and finish on time.

SUBMISSION AND FEEDBACK

Submit your work in Brightspace, under assignments. Within a day after the deadline, you will receive several (typically two) reports to grade for peer review as well as access to the online peer review form. You have 5 days to complete these reviews. You will then receive the anonymous review forms for your groups report and code.

There is the possibility to question the amount of points given to your work, up to one week after receiving the completed forms. You should do so via a private message to the teacher and TA in Mattermost.