VRIJE UNIVERSITEIT AMSTERDAM

Assignment 1
Data Mining Techniques
Data Mining Practice and Theory

Deadline: 21/04/2024, 23:59

INTRODUCTION

This document introduces you to the first assignment of the Data Mining Techniques course at the VU. This is a group task (3 members), please make sure all team members contribute to the work as expected. The assignment follows the structure of the CRISP-DM process model and the first five lectures of the course. You will study and apply the different subjects and algorithms that have been treated during these lectures to get hands on experience and gain a deeper understanding.

TOOLS/PROGRAMMING LANGUAGE: You have complete freedom to use any tool or programming language and can also combine them (e.g. data preparation in one tool and classification using the resulting dataset in another). There is no need to implement your own algorithms unless it is explicitly mentioned in the assignment, you can just use packages that are readily available. If you are in doubt on your choice, we would recommend using Python as that is the most frequently used language and the algorithms treated during the lecture are widely available in packages (e.g. scikit learn).

DATASET: The group following Data Mining Techniques is quite diverse, an advanced and basic version of this first assignment is available. Many components of this assignment revolve around analyzing a particular dataset. The basic version is based on a simpler dataset, while the advanced version focuses on a more challenging, complex dataset. The latter should

only be followed by those that already have some experience with machine learning and want to challenge themselves. A bonus point will be give to those who do the advanced version. In case there are specific instructions for the advanced dataset in the assignments, these will be clearly indicated.

- *Basic dataset* The basic dataset is the dataset we collected during the first lecture containing information on all of you. In case this dataset does not satisfy your interest, you are also allowed to use an alternative dataset from for instance Kaggle, but please discuss this with your TA first. Note that there should be a classification as well as a regression problem in there and the dataset should not be "pre-cleaned".
- *Advanced dataset* The domain from which the dataset originates is the domain of mental health. More and more smartphone applications are becoming available to support people suffering a depression. These applications record all kinds of sensory data about the behavior of the user and in addition frequently ask the user for a rating of the mood. A snapshot of the resulting dataset is shown in Table 1. The dataset contains ID's, reflecting the user the measurement originated from. Furthermore, it contains time-stamped pairs of variables and values. The variables and their interpretation are shown in Table 2. The goal of the dataset will be to predict the mood of the next day for the subjects in the dataset (being the average of the mood values measured during that day).

Table 1: Snapshot of the advanced data

ID	**Timestamp**	**Variable**	**Value**
AS14.01	26-02-2014 15:00.00	mood	6
AS14.01	26-02-2014 15:21.00	activity	0.031
AS14.01	26-02-2014 15:55.00	screen	103.1
AS14.01	27-02-2014 16:00.00	mood	6
AS14.01	27-02-2014 12:00.00	appCat.builtin	0.052

REPORT: We would like you as a group of 3 to prepare a report with the following in mind:

- The report should be submitted via Canvas by 21/04/2024, 23:59. This is a strict deadline, please try to respect that, otherwise points will be deducted (1 full point per day)
- Please format the document according to the LNCS guidelines. Templates are available on Canvas for both LateX and Microsoft Word, do not deviate from these templates (so no adjusting the margins, etc.). Note that you do not need to include an abstract in your report, but do start with a brief introduction. The paper **should not exceed 14 pages including all figures and tables, but excluding references** (references do not count for the number of pages to encourage you to cite all relevant work). With the page limit, our aim is to challenge you to report only what is necessary.

Table 2: Variables in the advanced dataset

Variable	**Explanation**
mood	The mood scored by the user on a scale of 1-10
circumplex.arousal	The arousal scored by the user, on a scale between -2 to 2
circumplex.valence	The valence scored by the user, on a scale between -2 to 2
activity	Activity score of the user (number between 0 and 1)
screen	Duration of screen activity (time)
call	Call made (indicated by a 1)
sms	SMS sent (indicated by a 1)
appCat.builtin	Duration of usage of builtin apps (time)
appCat.communication	Duration of usage of communication apps (time)
appCat.entertainment	Duration of usage of entertainment apps (time)
appCat.finance	Duration of usage of finance apps (time)
appCat.game	Duration of usage of game apps (time)
appCat.office	Duration of usage of office apps (time)
appCat.other	Duration of usage of other apps (time)
appCat.social	Duration of usage of social apps (time)
appCat.travel	Duration of usage of travel apps (time)
appCat.unknown	Duration of usage of unknown apps (time)
appCat.utilities	Duration of usage of utilities apps (time)
appCat.weather	Duration of usage of weather apps (time)

- Make sure we can identify your report, i.e., your group number, names and student numbers should be in the document's header.
- Structure the report following the assignments, be selective in what you report as space is limited, but do provide good descriptions of the steps you have taken (to make it reproducible) and the rationale behind your choices.

GRADING: You can earn a total of 100 points. 10 bonus points are given to those selecting the advanced dataset. Your grade for the assignment is the number of points divided by 10 (with 10 being the maximum grade). At the end of this assignment you can find the grading scheme used.

TASK 1: DATA PREPARATION (30 POINTS)

As discussed during Lecture 2, the first phase of a Data Mining project typically includes getting familiar with the domain and pre-processing the dataset in a suitable manner. In this part of the assignment, we will go through those steps.

TASK 1A: EXPLORATORY DATA ANALYSIS (10 POINTS)

Start with exploring the raw data that is available:

- Notice all sorts of properties of the dataset: how many records are there, how many attributes, what kinds of attributes are there, ranges of values, distribution of values, relationships between attributes, missing values, and so on. A table is often a suitable way of showing such properties of a dataset. Notice if something is interesting (to you, or in general), make sure you write it down if you find something worth mentioning.
- Make various plots of the data. Is there something interesting worth reporting? Report the figures, discuss what is in them. What meaning do those bars, lines, dots, etc. convey? Please select essential and interesting plots for discussion, as you have limited space for reporting your findings.

TASK 1B: DATA CLEANING (10 POINTS)

As the insights from Task 1A will have shown, the dataset you analyze contains quite some noise. Values are sometimes missing, and extreme or incorrect values are seen that are likely outliers you may want to remove from the dataset. We will clean the dataset in two steps:

- Apply an approach to remove extreme and incorrect values from your dataset. Describe what your approach is, why you consider that to be a good approach, and describe what the result of applying the approach is.
- Impute the missing values using two different approaches. Describe the approaches and study the impact of applying them to your data. Argue which one of the two approaches would be most suitable and select that one to form your cleaned dataset. Also base yourself on scientific literature for making your choice.

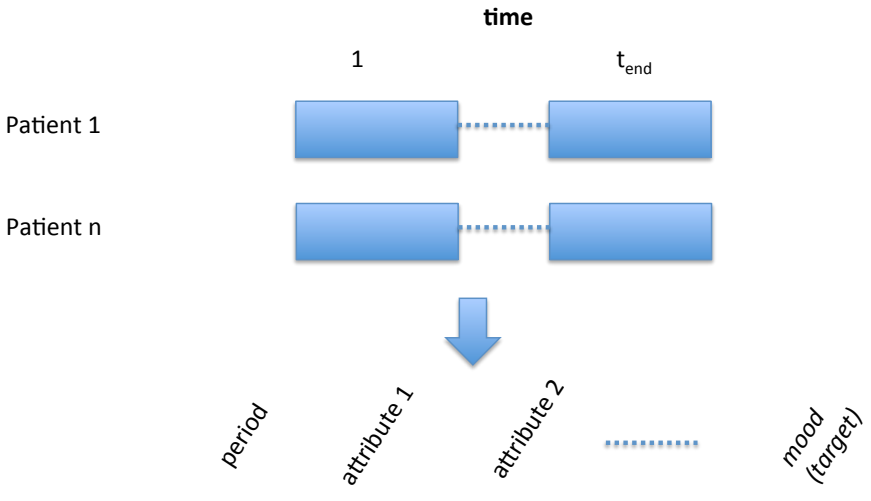
Advanced: The advanced dataset contains a number of time series, select two approaches to impute missing values that are logical for such time series and argue for one of them based on the insights you gain. Also consider what to do with prolonged periods of missing data in a time series.

TASK 1C: FEATURE ENGINEERING (10 POINTS)

While we now have a clean dataset, we can still take one step before we move to classification or regression that can in the end help to improve performance, namely feature engineering. As discussed during the lectures, feature engineering is a creative process and can involve

for example the transformation of values (e.g. take the log of values given a certain distribution of values) or combining multiple features (e.g. two features that are more valuable combined than the two separate values). Think of a creative feature engineering approach for your dataset, describe it, and apply it. Report on why you think this is a useful enrichment of your dataset.

Advanced: Essentially there are two approaches you can consider to create a predictive model using this dataset (which we will do in the next part of this assignment): (1) use a machine learning approach that can deal with temporal data (e.g. recurrent neural networks) or you can try to aggregate the history somehow to create attributes that can be used in a more common machine learning approach (e.g. SVM, decision tree). For instance, you use the average mood during the last five days as a predictor. Ample literature is present in the area of temporal data mining that describes how such a transformation can be made. For the feature engineering, you are going to focus on such a transformation in this part of the assignment. This is illustrated in Figure 1.



Patient 1	t = 1-6	value for t=1-5			value for t=6
Patient 1	t = 2-7	value for t=2-6			value for t=7
Patient 2	T = 1-6	value for t=1-5			value for t=6

Figure 1: Predictive model

In the end, we end up with a dataset with a number of training instances per patient (as you have a number of time points for which you can train), i.e. an instance that concerns the mood at t=1, t=2, etc. Of course it depends on your choice of the history you consider relevant from what time point you can start predicting (if you use a windows of 5 days of history to create attributes you cannot create training instances before the 6th day). To come

to this dataset, you need to:

1. Define attributes that aggregate the history, draw inspiration from the scientific literature.
2. Define the target by averaging the mood over the entire day.
3. Create an instance-based dataset as described in Figure 1.

TASK 2: CLASSIFICATION (30 POINTS)

Now that we have formed our final dataset, we can move to some modeling. First, we will focus on a classification task.

TASK 2A: APPLICATION OF CLASSIFICATION ALGORITHMS (20 POINTS)

Identify the target (i.e. the class you want to predict) for your dataset. In case you use the dataset we collected you are free to choose whatever you like. Split up your data in a train and test set and apply two classification algorithms, at least one of them should have been discussed during the lectures. Optimize the hyperparameters of the approaches. Measure and discuss the performance using a performance metric and argue why that is a suitable metric. Describe all steps in your process clearly and fully to make sure it is reproducible.

Advanced: For the advanced assignment you go through the same steps (and shape it into a classification problem for predicting the mood of the next day), however you are required to use two different types of classification algorithms, namely one that uses the dataset you formed in Task 1C (e.g. using a random forest) and an algorithm that is inherently temporal (e.g. recurrent neural networks). Also consider a good evaluation setup given the nature of the dataset.

TASK 2B: WINNING CLASSIFICATION ALGORITHMS (10 POINTS)

Machine learning techniques that are used in Data Mining projects develop quickly these days. One nice way to track these developments is to see which algorithms win competitions on websites such as Kaggle. Your task is to describe the approach of the winner of one of those competitions that focus on a classification tasks. The following sites might serve as starting points:

- `http://www.kaggle.com/` - DM competitions
- `https://www.kdd.org/kdd-cup` - KDD Cup
- Etc. - You should be able to find other relevant competitions by searching the Web.

The main goal is that you can demonstrate that you understand a technique that beats other techniques under certain conditions (specified by the task and data at hand). Here's what we'd like you to include in the report for this task:

- A description of the competition: what competition, when was it held, what data they were using, what task(s) they were solving, what evaluation measure(s) they used.
- Who was the winner, what technique did they use?
- What was the main idea of the winning approach? (Typically this would come from a paper written by the winners.)
- What makes the winning approach stand out, or how is it different from standard, or non-winning methods?

Particular rules and points to consider:

- A suggestion: 1 page should be more than enough for this task.
- Needless to say, but for the record, please do not copy and paste from papers. Always cite (properly) the source of the paper you are using.

TASK 3: ASSOCIATION RULES (10 POINTS)

We have seen the APRIORI algorithm during the lecture that targets finding associations in datasets, predicting that an item is likely to be bought given other items that are in the shopping basket already. As mentioned during the lecture, many innovations have been made to improve the APRIORI and other methods. One category of improvements involves grouping of products into higher level product categories (e.g. a Pizza Margherita and Pizza Quattro Formaggio are both pizza's). Find an approach that aims to do this and describe it. Discuss the pros and cons of such an approach.

TASK 4: NUMERICAL PREDICTION (10 POINTS)

Similar to Task 2A, apply two machine learning algorithms to your dataset, but now focus on predicting a numerical target (i.e. a regression problem). Describe similar details as you have for the classification problem. Highlight the differences you see between the two types of prediction tasks.

TASK 5: EVALUATION (20 POINTS)

As a final part of the assignment, we will study the impact of your evaluation metrics and their characteristics.

TASK 5A: CHARACTERISTICS OF EVALUATION METRICS (10 POINTS)

Consider the following two error measures: mean squared error (MSE) and mean absolute error (MAE).

- Write down their corresponding formulae.
- Discuss: Why would someone use one and not the other?
- Describe an example situation (dataset, problem, algorithm perhaps) where using MSE or MAE would give identical results. Justify your answer (some maths may come handy, but clear explanation is also sufficient).

TASK 5B: IMPACT OF EVALUATION METRICS (10 POINTS)

Apply the MSE and MAE as evaluation metrics to the numerical prediction problem you have worked on under Task 4. Describe how the model behaves under the different characteristics and describe the implications.

Table 3: Grading scheme

Task	**Grading Component**	**Weight**
1A	Description of the dataset (statistics)	3
	Plots of some features	3
	Interpretation and rationale	4
1B	Description of approach and results to remove outliers	2
	Description of two approaches to impute missing values and comparison	3
	Interpretation and rationale	5
1C	Description of feature engineering approach	5
	Interpretation and rationale	5
2A	Description of classification approaches	4
	Description of hyperparameter optimization	2
	Description of evaluation setup	4
	Description of results	5
	Interpretation and rationale	5
2B	Description of the competition	2
	Describe winning technique	4
	Analysis/comparison	4
3	Description of algorithm	5
	Discussion pros and cons	5
4	Description of regression approaches	2
	Description of hyperparameter optimization	1
	Description of evaluation setup	2
	Description of results	2
	Interpretation and rationale	3
5A	Formulae	2
	Example	4
	Rationale	4
5B	Description of results	6
	Interpretation and rationale	4
Deductions	Extra page	-10
	Late (per day)	-10
	Wrong formatting	-10
Total		100