VU Data Mining 2016 – Assignment 1 (“advanced” variant)

Mark Hoogendoorn, Ward van Breda

Start of assignment: March 30, 2016

Deadline: April 17, 2016 23:59

1 Introduction

This document introduces you to the first assignment of the Data Mining Techniques 2016 course at the VU. This is a group task (maximum 3 members), please make sure all team members contribute to the work as expected. The assignment essentially consists of creating predictive models for a challenging dataset, not from the perspective of the size of the dataset, but more the nature of the data. You can earn a total of 100 points.

2 Dataset and problem description

As a first step, let us look at the dataset we are faced with. 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 behaviour of the user and in addition frequently ask the user for a rating of the mood. A snapshot of the resulting dataset is shown below.

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

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 1.

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)

Table 1. Variables measured

Using this dataset, we would like to build a predictive model that is able to **predict** the **average mood** of the user **on the next day based on the data** we obtained from the user **on the days before**. This is illustrated graphically in Figure 1 below.

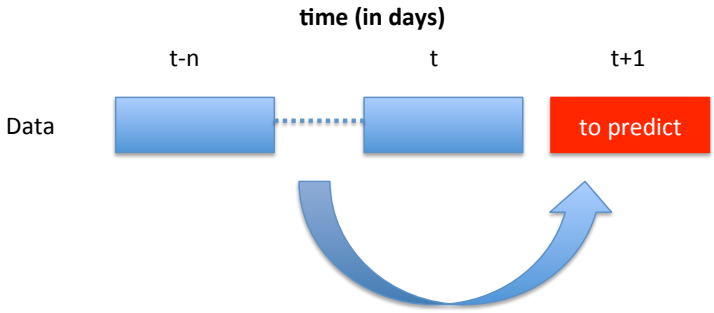
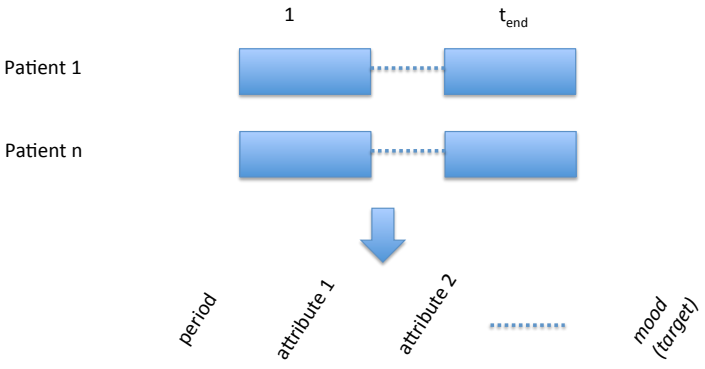


Fig. 1. Predictive model

In order to create such a predictive model, we need to perform some transformations and we need to decide on what features we want to use for these predictions.

Task 1: Pre-process the dataset (40 points)

Essentially there are two approaches you can consider to create a predictive model using this dataset: (1) use an machine learning approach that can deal with temporal data (e.g. ARIMA, 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. We are going to focus on such a transformation in this part of the assignment. What we are trying to do is illustrated in Figure 2.



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

Fig. 2. Temporal abstraction

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 field of temporal data mining.
2. Define the target by averaging the mood over the entire day.
3. Create an instance-based dataset as described in Figure 2.

Make an appointment with Mark to discuss your ideas before finalizing them. In the end, your should describe and argue your choices clearly and back them up with scientific literature. Next, perform a preliminary analysis (e.g. look at correlations) on the usefulness of the attributes you have defined.

Task 2: Learn using the dataset (40 points)

In the next step, we are going to use our dataset to create a predictive model. You can make your own choice whether you want to create individual models per patient or a single model for all patients. You will need to study three variants of predictive models:

1. A variant where you use the pre-processed dataset you identified in Task 1 in combination with a machine learning technique you consider appropriate.
2. A variant where you apply a learning algorithm that is able to cope with this temporal data (e.g. ARIMA, recurrent neural networks, etc.).
3. Implement a benchmark: predict the mood on the next day by just saying it is the same as the previous day.

Define a proper performance metric and create a solid evaluation setup. Describe and argue your choices again and show the results you have obtained. Create graphs to illustrate the performance in an insightful way.

Task 3: Evaluate and reflect on your results (20 points)

Finally, analyse the results in detail both using a more statistical view and by means of your interpretation. Argue what the pros and cons of the different approaches are.

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 BB by 17/04/2016 23:59. This is a strict deadline, please try to respect that, otherwise points will be deducted.
- Please format the document according to the lncs guidelines. The lncs format is used for scientific papers published by the Springer, where lncs stands for Lecture Notes in Computer Science, see http://www.springer.com/computer/lncs?SGWID=0-164-6-793341-0, Note that you don't need to include an abstract in your report. The paper should not exceed 8 pages. With the 8 pages limit, my aim is to challenge you to report only what is necessary.
- Make sure we can identify your report, i.e., at least a subset of the (name, student number, vu-netID) triplet should be in the document's header.
- Make an attempt to make the report look professional. Have a short introduction of your document, use appropriate language, etc. Let's say, if you gave your report to the manager of your DM project at a company, they would need to be able to understand it and conclude that it's a good project start.

Grading

Marking will be based on the tasks as reflected by quality of the report (so content, style, etc. all matter). You can get maximum 100 marks for this assignment. You will need at least 55 to pass. Also, 100 points are only given to students whose reports are of exceptional quality, and they also should report something we did not specifically ask for (in other words, we value proactivity and creativity).