VU Data Mining 2017 – Assignment 1 (“basic” variant)

Start of assignment: April 3, 2017
Deadline: April 23, 2017 23:59

1 Introduction

This document introduces you to the first assignment of the Data Mining Techniques 2017 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 consists of three parts: (1) the exploration of a first dataset, namely the dataset of all of you; (2) participating in a data mining competition for a relatively simple dataset, and (3) gaining some insights into research and some more theoretical parts of data mining. You can do this assignment with groups of three students. You can earn a total of 100 points.

Task 1: Explore a small dataset (40 points)

When you do data mining (DM), you will need to check for data quality, manipulate data, as well as to build and evaluate models. You can write your own applications for this purpose, and in many commercial projects this is exactly what you are expected to do, but to start with DM and to carry out typical tasks I find it better to use a software that is specifically written for this purpose. In this first assignment, the aim is that you get familiar with your chosen DM software. You should learn to use it on a basic level, and carry out elementary DM tasks, so you will be prepared to take on more challenging ones coming to you in the form of subsequent assignments.

Tasks and Expectations

You (as a group) should start by choosing a certain DM software. Here's a list of corresponding rules:

- If you are uncertain about this choice, I recommend you to pick RapidMiner or WEKA. If you have a favourite programming language, you can also look for packages available in that language.
- This is not a choice for the whole course. If you find out later that another software is better suited for a task at hand, you can switch to that in subsequent assignments.

Task 1.A – Exploration (20 points)

Once you have chosen (and perhaps installed) the software, here is your first task:

- Download the ODI dataset from BB. ODI stands for Own Dataset Initiative.
- Load the dataset, which will be in CSV format as well as Excel, into your software. CSV stands for comma separated values, though it is more common these days to separate values with semi-colon (;). Nevertheless, the format is still called CSV even if the separator is a tab character or whatever else defined by the dataset creator.
- 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, and so on. 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 (see details in a later section).

To sum up, you will need to explore the dataset, and report findings that are essential to get an idea about the data, and also, findings that make it possible to learn something interesting about the dataset.

Task 1.B – Basic classification/regression (20 points)

Our main goal with this task is that you learn to run simple experiments. Here's the task list:

- Take the ODI dataset and load it. Alternatively, you can download a dataset of your own choice from the web, and load that. If you opt for a downloaded dataset, write down why that interests you, and why it is suited for classification/regression.
- Design and run at least one classification/regression experiment on the data, with cross validation. You will probably need to go through a couple of tutorials to accomplish this task. Don't worry if you don't know what cross validation is, we will cover that later. I just require you to use that to avoid that people do this task with one line of code.
- Note the setup you use, the results you get, and try to understand what happened, what models have been built, what numbers have been outputted by the algorithm you used.
- Try at least two algorithms, and try to interpret the differences in outcome of the experiments. This doesn't need to be a deep analysis, remember that this assignment is only to get you started. We will learn more about performance measures and comparison later.

Once you are done with this task, write up your findings.

Task 2: Compete in a Kaggle Competition to Predict Titanic Survival (30 points)

The previous task consisted of a rather artificial dataset to explore and allowed you to get acquainted with the tools you selected. The next assignment is more competitive as you are going to participate in a so-called Kaggle competition. The main goal is to pre-process the data as good as you can, select the right techniques and obtain a good score. The Kaggle competition we are going to focus on is the Titanic competition, which can be found on www.kaggle.com/c/titanic-gettingStared/. The main idea behind the competition is to come up with a model to predict whether someone survived the Titanic disaster or not based on a training set of people for whom you know whether they survived or not.

Task 2.A – Preparation (10 points)

Once you have registered on Kaggle (use VU-DM-group as your team name where group is your group number in the course), do the following:

- Download the Titanic training set and explore the data (what attributes are there, what are their types, what are their distributions, do you see any obvious correlations). Describe you most important findings.
- Based on your findings during the exploration phase, prepare the data for a learning algorithm: what attributes should you select, should you transform certain attributes to make them usable in the learning algorithm, etc. Describe which selections/transformations you have made and provide a rationale for you choices.

Task 2.B – Classification and evaluation (20 points)

Now that you are familiar with the dataset you can apply different classifiers.

- Create a setup that enables you to evaluate a classifier (i.e. create a training and test set based upon the dataset you have downloaded from Kaggle).
- Apply and evaluate at least two classification algorithms based on your evaluation setup.
- Take your best classifier and apply it to the test set posted on Kaggle, upload your results. Report on your score on the leadership board and explain whether this is accordance with your expectations.

Task 3: Research and theory (30 points)

The following assignment go into more detail on state-of-the-art DM approaches as well as the more theoretical aspects of DM.

Task 3.A – Research: State of the art solutions (10 points)

Find a data mining related competition that's already finished. Your task is to describe the approach of the winner.

The following sites might serve as starting points:

- http://www.kaggle.com/ - DM competitions
- http://www.sigkdd.org/kddcup/ - KDD Cup
- http://trec.nist.gov/tracks.html - Mostly text mining/retrieval, not specifically competition, though there are always “best submissions".
- 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.B – Theory: MSE verse MAE (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).
- Run an experiment on any dataset obtained from the web, measure MSE and MAE of different regression methods, and discuss the differences you find. (Make sure to include the link where you got the data from, add a sentence about why you chose that dataset, and another describing its size, attributes, etc.)

Task 3.C – Theory: Analyze a less obvious dataset (10 points)

Consider the dataset SmsCollection.csv, this collection contains 5574 text messages from the UK labeled either as spam or ham. Ham signifying it is a bonafide text message, spam meaning it is an unwanted message.

- The data in this collection is only regular texts, which modelling techniques would be suitable to use with this type of data?
- Which data transformations would you use on this dataset, how would these transformation improve the quality of the data? Apply a selection of these data transformations on the data and then…
- Build a model on the data, using label as the class attribute. Describe its quality and whether it could be improved further in any way.

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 23/04/2017 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).