VU DMT 2012 – Assignment 1

Zoltán Szlávik (Zoli)

April 3, 2012

1 Introduction

This document introduces you to the first assignment of the Data Mining Techniques 2012 course at the VU. My aim with this document is to tell you why I find it important to do this assignment, what exactly you are expected to do, what will result in a high grade for you, and what happens next.

2 Motivation

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.

3 Tasks and Expectations

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

- If you are uncertain about this choice, I recommend you to pick RapidMiner.
- Weka is not allowed for this assignment, as several of you are already familiar with it, and it wouldn't be fair to others if this assignment was a piece of cake to you.
- If you do not choose RapidMiner, you should – in one or two sentences – justify your choice. Please do not say "because I already know it", but write down why you think your chosen software is suited for DM tasks.

- 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.

3.1 Task A – Exploration

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

- Download the ODI III dataset from BB. ODI stands for Own Dataset Initiative, and it's III because I've been collecting such datasets since two years ago.
- Load the dataset, which will be in CSV format, into your software (if it requires loading). 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.

3.2 Task B – Basic classification/regression

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

- Take the ODI III 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.

3.3 Report

I would like you (as an individual, just to stress this) to prepare a report with the following in mind:

- The report should be submitted **via BB by 12/04/2012 23:59**. This is a strict deadline, please try to respect that.

- The paper should not exceed **4 pages**, with a reasonable font size and formatting in place. With the 4 pages limit, my aim is to challenge you to report only what is necessary. If you want to make my life easier, please do not use large margins (so I won't need to zoom in to read your report on my old iPad).

- Make sure I 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.

4 Grading

What happens next is that I'll collect your reports, read them, and mark them. Marking will be based on the tasks (i.e. exploration and basic modelling) as reflected by quality of the report (so content, style, etc. all matter).

You can get **maximum 10 marks** for this assignment. You will need at least 5.5 to pass. Also, 10 is only given to students whose reports are of exceptional quality, and they also should report something I did not specifically ask for (in other words, I value proactivity and creativity).

5 What's next on DMT

We'll move on with lectures, and I'll soon introduce the rest of the assignments.