VU DMT 2012 – Assignment 2

Z. Szlávik (Zoli)

VU University Amsterdam
z.szlavik@vu.nl

Abstract. This document describes Assignment 2 of the Data Mining Techniques course of 2012. There are tasks to be carried out individually, and tasks to be done in groups of two. Submitting one document per group is required that should report on both members' individual tasks as well as on group tasks.

1 Motivation

I have several aims with this assignment: I'd like to test your data-related common sense, which is important in a business environment, I believe. Also, I'd like you to do a bit of research on data mining oriented algorithms so you won't try to start solving a difficult problem without checking what others have done so far about it. Furthermore, I'd like you to go into the details a bit, so you will be able to decide between options you can have. Finally, I'd like you to be able to present a DM application, find the key points to mention, be concise, and be creative during the process. Don't worry if you don't find these tasks 'technical' or 'practical' enough, you'll get to do enough of those things in Assignment 3. However, I can assure you that the skills you can gain by completing this assignment will be useful for later.

Let's see the tasks then.

2 Tasks and Expectations

In this section, I'm going to describe your tasks, and what I expect you to deliver in your reports. In other words, given my input tasks, this is section gives you the content and format you need to produce as output.

2.1 Task A (individual)

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. KDD Cup 2011 was last year's Assignment 3. (KDD Cup 2011 has a separate website)
- 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 your understand a technique that beats other techniques under certain conditions (specified by the task and data at hand).

Here's what I'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:

- Do not choose the same competition/year as your teammate.
- 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.

Maximum marks you can get for Task A of this assignment: 8.

2.2 Task B (individual)

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 an 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.)

A rule: don't use the same dataset as your partner.
Maximum marks you can get for Task B of this assignment: 8.

2.3 Task C (group)

Take the dataset '`OneOutlierName.csv`'. The dataset doesn't have attribute names, consider that when loading the data. This is a sort of explorative-creative task, and I need you to answer the following questions:

- What is this dataset about?
- What are (possibly) the attributes?
- What is the name of the most obvious outlier of the dataset? (yes, you have to find a name for it)
- Where did I possibly get the data from, and how did I create this dataset?

Apart from basic DM tool usage, answering the first three questions requires, let's say, high school level knowledge of one particular subject (that I'm not going to disclose because then the task would be too easy, but it's not maths).

Maximum marks you can get for Task C of this assignment: 8.

2.4 Task D (group)

Consider the dataset '`ZAVdataset.csv`'. There are three columns in there, named z, a and V, where the latter is the class attribute (label). Your task is to

- Build a model on the data, using V as the class attribute. Describe its quality and whether it could be improved further in any way. Once you have a sufficiently good model, you should be able to...
- Describe the dataset with one single word (i.e., give it a much better name than it has now). Note that this step requires some creativity (or maybe intuition is a better word but I'm not sure. You'll need to see/identify something there, and keep in mind that I have a particular kind of humour).

- Cluster the dataset in whatever way you choose (maybe experiment with several algorithms and parameter values). How many clusters did you find? Do they make sense? What might those clusters mean?
- How did I generate the dataset? (roughly; adding specific findings might be possible though not expected)

Maximum marks you can get for Task D of this assignment: 8.

2.5 Task E (group)

This is the presentation task I mentioned earlier in this document. It also happens to be directly connected to one of the ODI III questions. Here you are:

- Find an example of a successful data mining application. You can do this by just searching the web, already knowing one that uses data mining and is successful because of it, you read about it in a newspaper, saw something in the street, heard from a friend, etc..
- Once you make your choice, you will have to present this application to the world (but primarily to me and your classmates as target audience). You might want to address issues such as why did you choose it, why it is successful, how do they do things, what techniques they might be using, etc.. And here's the twist: the presentation should be in the format of a **maximum two minutes long video**! Be creative!

The video can to be submitted to me in any common video format, and I would appreciate if it was below 200MB in size. The resolution should be 'reasonable'. There are a couple of options for submission: 1) include a link in your report that I can use to download the file, 2) bring it on a USB stick, 3) upload it to youtube and give me the link (though I prefer files), 4) whatever else as long as it gets to me.

Important: I will need you to state specifically if you give me your consent to use your video 1) in class (as I want to show you selected presentations) and 2) for educational purposes in general (for example, to show your video to students of next year). This is an important ethical issue to sort out. Your grade will not be affected, in any sense, by whatever your answers to these two questions will be. There is a Consent form attached to this document, please make sure I get it back, even if you don't sign it anywhere, only add you names. I accept hard copies or scanned PDFs (separate, or added to the report as last page).

Here's a couple of ideas about what the video might be like:

- Create a powerpoint presentation, add your voice and record the presentation. But I beg you, make this your very final resort to turn to.
- Demonstrate the application by using some screen and voice capture software (HyperCam, Camtasia, etc.).
- Make an animated movie (with a good script). Example: `http://www.youtube.com/watch?v=ax0tDcFkPic`
- Sit down behind a desk and describe the application (Evening news).
- Figure out why an application is successful by asking people on the street/campus.
- Make a nice clip with music. Example: `http://www.youtube.com/watch?v=4B2xOvKFFz4`
- Make a short documentary. Example: `http://dataminingtools.net/blog/2011/02/23/data-mining-hip-hop/`
- Record a short, but good quality, talk... while on the bike/train/bus/tram/metro/boat/hill/tree/TV tower. (you are responsible for your own safety!)
- Whatever else that can make sense for the assignment.
- Combination of the above.

It's all about content, structure and presentation (and these attributes are certainly not independent).

Maximum marks you can get for Task E of this assignment: 8.

2.6 Report

I would like you to create **one report per group**. Include parts for the group work, then have a section with the individual work of person A in the group, and then a section with person B's work (*please use student numbers/names in the individual sections' titles so I can identify you*).

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. I have uploaded a template to blackboard, for LaTeX, see the appropriate files in the zip file (I modified the template slightly though). If you want to use Word, you can find a template at http://www.springer.com/computer/lncs?SGWID=0-164-6-793341-0, if you do that, please be aware that I use slightly different margins, so you might want to adjust those to make the doc look like this document (which is in the required format). Note that you don't need to include an abstract in your report.

The report should be **maximum 12 pages** long. This might be a stretch for some of you (who usually produce long reports, or who have a lot of figures they want to show) but I value to-the-point and concise reports (and managers value those, too). If you have too many figures, keep the most informative ones, and perhaps mention in a couple of sentences what is (less) interesting in the figure you did not include. In case you are wondering, it is 12 pages *with* possible reference lists and appendices, but I don't think you will need these too much, particularly the appendices.

Submission should be done **using BB**.

The submission deadline is **01/05/2012 23:59**. This time I'll set the deadline in BB to exactly that date.

3 Grading and next steps

I'll mark your reports as soon as I can, and let you know when I'm done. As you could already calculate, you can get maximum 40 marks for Assignment 2, 8 marks for each task. (You can achieve 16 marks individually, and 24 marks for the group tasks.)

Good luck!