VU DMT 2012 – Assignment 3

Z. Szlávik (Zoli)

VU University Amsterdam
z.szlavik@vu.nl

Abstract. This document describes Assignment 3 of the Data Mining Techniques course of 2012. Please make sure you read it thoroughly and carefully. This is a group task, and please make sure all team members contribute to the work as expected. There will be two things to be submitted: 1) A report, 2) A prediction vector. For details, please see below.

1 Motivation

By now should have a fair idea about techniques we can use, and should also have some practical experience with mining datasets. In this third assignment, you will gain more experience, explore various techniques (and whether they work in this situation), and hopefully learn a lot.

The topic of this assignment is football (or soccer, in US English). I opted for it for the reason that right after this course finishes, the European Championships will begin, and the whole country will dress up in orange from time to time; things we can't avoid, and actually may look forward to.

The motivation for the assignment is not that we want to build models that make us money at the bookie's. Actually, it comes from the fact that I had the idea not too long ago that the FIFA team ranking method doesn't really have any good predictive power, i.e. we could come up with models that predict match outcomes better than what we could expect from that ranking (i.e. if I'm ranked higher, it's more likely that my team will beat a low ranked one).

So, the task, in short, is to build a model, using historical data (or some additional data you can add), to predict outcomes of matches. We will focus on the period until the beginning of the European Championships.

2 Dataset

The dataset, which can be downloaded from BB, is **created from data available at the FIFA website**. A student of mine, Jan Lasek, was kind enough and extracted the data from there, cleaned it where it was necessary, and transformed it into a nice table format. He also provided us with a **short description** of the dataset, the corresponding file is also available via BB. (By the way, he's also working on this topic, although his focus is not simply on prediction, but prediction via effective ranking schemes.)

3 Detailed task description

To make things easier, I'll use a **DM process model** to describe your task:

3.1 Business understanding

Your task is to **predict outcomes of international football matches**. I assume you all know what football is about (*"22 people envying the ball from one another"*), so you don't need you to describe that, unless there's something you want to emphasize, e.g. some big events or changes in rules that might affect your models.

Another part of the business understanding is **how prediction is done by other people**. Team rankings can be used for prediction, player statistics, news, etc. can also be used. Can you find some relevant work that also attempts to predict outcomes of football matches? Please spend a couple of paragraphs on this topic (i.e. related work) in your report.

3.2 Data understanding

Essentially, this is a subtask that requires you do **exploratory data analysis (EDA)**. Explore the dataset, count, summarise, plot things, and report findings that are useful for your task. For instance, if you want to use information about where a game is played, you might want to explore games continent by continent. This was just an example, you should be able to find more very easily.

Remember that EDA is not necessary done once and then you move on. It might very well be possible that you do some EDA, build some models, then some idea comes up, do some more EDA, modify your model according to what it shows, and so on.

3.3 Data preparation

You'll certainly need to work on the dataset, to create, modify or add new features, but certainly to separate the training set from the test set.

Training and testing You can use the dataset for **training** up to the date 17/05/2012, i.e. until which match results are recorded in the dataset. When estimating the performance of your prediction methods (**validation**), please consider that the dataset uses time (dates), and so you should make sure that you don't train on the future and test on the past. For example, you can use a match in March 2012 to predict and check an outcome in April 2012, but the reverse order is methodologically incorrect.

The **test data** I will need you to submit predictions for is between 18/5/2012 and 5/6/2012. This covers 114 matches. Scheduled matches after this period are not part of the assignment, though you can use them for your own amusement.

In short:

1. Data about **up to 17/05/2012 is for training and validation**,
2. between **18/5/2012 and 5/6/2012 is for testing**,
3. after 5/6/2012 is not part of the assignment.

New attributes The task certainly requires you to get a bit dirty with the data (you may also need to write some programmes, scripts, or macros), and here's a couple of things you might want to do:

- **Create new attributes from existing ones:** for instance, you might want to calculate goal differences, use month information, or create attributes which are very brief summaries of previous matches (e.g. number of matches won out of the last 5 appearances, etc.). Depending on your perception of what will work, you can get very creative here, or don't even bother creating new attributes this way.
- **Creating new attributes using external information:** you now have cities, what if you replace them by their corresponding countries or continents? If it's some qualifier, what if you add which round the match was from? You can add player statistics, bookmaker predictions, weather information, city altitude, and so on. Depending on time and ideas, many things are possible here. Note that you should be able to create these new attributes (considered to have predictive power) for the instances in the test set as well!

3.4 Modelling and Evaluation

Naturally, once you prepare the dataset, you should be able to **build models**. Use whatever technique you want, though the choice might be influenced by how I would like to measure your predictions at the end (see Section 5.1).

Regarding evaluation, as already mentioned, please make sure you don't train on the future, and test on the past. This implies that **traditional cross-validation will not be applicable**. One way you might want to estimate the quality of your predictions is as follows: Take a sliding window approach, which I'd like to illustrate with an example described in the next paragraph.

Let's say you want to use 4 years of data to predict 1 month worth of matches (numbers here inspired by FIFA ranking, but you can choose others if you want to). Then here's what you can do: train on data from Jan 2006 to Dec 2010, test (validate) on Jan 2011, and record evaluation score (accuracy, RMSE, your choice, though it might be influenced by my choice of performance metric, Section 5.1). Then you shift, and train on data from Feb 2006 to Jan 2011, test on Feb 2011, record the score. And so on, until, say, you arrive at the most recent data you have results for. Then you take the average score. Once you experiment with several modelling methods, you should be able to choose the one that has the highest average validation score, make predictions on the (real) test set with the winning model, and submit your predictions to me.

3.5 Deployment

Not really deployment, as you are (probably) not building a commercial product, but I'd like be sure that your results are replicable, so I might select some people (randomly, or by competition rank, I'll see it later) and ask them to demonstrate whether and how their approach works.

4 Deliverables and formats

I've covered the process above, let us see what I expect you to deliver.

4.1 Predictions

You'll need to submit a single file containing a **matrix sized 114 x 3**, where **114 is the number of matches** in the test set. The **order** in which you submit your predictions should match that in the dataset.

Three refers to the following three numbers: **1)** probability of Team 1 (column 2) winning, **2)** probability of a draw, **3)** probability of Team 2 (column 3) winning. Naturally, the three numbers in a row should **sum up to 1**.

Wins are determined based on the **whole game**, i.e. if there's extra time and/or penalties, whoever wins at the end should be predicted by you.

The submission file should be a CSV file, its name should contain the VUnetID of both team members. The contents of the file should include only the matrix, i.e. no header row is allowed. Columns should be separated by commas. Just to be sure, here's an example (Table 1).

Table 1. Submission example (tab here only for clarity)

1,	0,	0
0,	1,	0
0.8,	0.1,	0.1
0.45,	0.45,	0.1
0.2,	0.5,	0.3
...,	...,	...

I'll try to provide a submission validation tool soon (or someone can volunteer to create one).

4.2 Report

The assignment is not only about winning, but also about quality of the process and understanding of what you did. Therefore, I'd like you to write a report which should contain the following:

- **What you did** – you might want to follow the process model, and describe the steps you took. If you tried a number of things but only some worked, please mention those that did not work as well, and discuss why they might not have worked.

– **What you learned** – either inside the main part of the report, or separately in a paragraph of two, please describe what skills and knowledge you have gained from this assignment, what were the main difficulties, expected and unexpected outcomes of your experiments, etc.

Please **format the report** according to the `lncs` guidelines (only with larger margin, exactly the way you used previously when doing A2).

The report should be **maximum 10 pages** long. Again, it is 10 pages *with* possible reference lists and appendices.

5 Evaluation

Here's how you will get rewarded for your work:

5.1 Winning the competition

Let me first use the official text that tells you what happens if you produce the **best predictive performance**:

"The O'Reilly Academic Prize Scheme is run each academic year with participating Universities across the UK, Europe, Africa and the Middle East.

Every participating University is awarded £100 of print books for their library along with £100 of ebooks for their winning student. (This can be split between more than one winner).

On top of the £100 ebook prize, each winning student is entered into a prize draw for an HTC Flyer 7 inch Tablet PC.

The prize draw takes place at the end June and the winning student's lecturer will be contacted then."

And here's what will make you the winner:

I will calculate a kind of **accuracy** for your submissions, and whichever team has the highest accuracy value will be declared winner[1]. Equation 1 shows how this is calculated:

$$Acc = \frac{1}{N} \sum_i P_W(i) \cdot A_W(i) + P_D(i) \cdot A_D(i) + P_L(i) \cdot A_L(i) \tag{1}$$

where $N = 114$ is the number of matches in the test set, Ps are predictions, As are actual values, W refers to Team 1 winning (corresponding predictions will be in the first column of your submission), and D and L stand for draw and losing (of Team 1), respectively. An actual vector, naturally, will always contain one *one* and two *zeros*, depending on the actual outcome of a particular game.

5.2 Getting a high grade

You can get **50 marks** for this assignment, so half of your final grade depends on this. **80%** of this 50 (i.e. 40) marks can be achieved by **submitting a nice and thorough report**, and **20%** will come from where you end up in the **competition**. Regarding the competition-based marks, I'll see how good your submissions are, and then I'll decide if you'll get the marks based on accuracy value, or rank in the competition.

Note that, as usual, you can boost your grade by doing extra things, such as not only predicting who will win, but also who will score how many goals, etc.

[1] I also had other measures in mind but this should take care of the three-way logic, and also, possible ties that would have been problematic with standard accuracy measures

6 Submission

Report submission should be done **using BB**. Regarding prediction submission, I'll try to set up something automatic, you'll get more info about this as soon as I have it.

The submission deadline for both the report and the predictions is **03/06/2012 23:59**, but you are encouraged to submit these earlier.

7 Closing events

On the **7th June**, there will be a room booked for us in the morning, where I'll **announce the winners** of the competition, ask the top ranked groups to do a small presentation about their approach, and there will be a couple of **other students presenting their work**. These 'other students' are also working on football-related projects, and so the morning will be not only about work and results you know well, but also about similar, and interesting, projects.

On the same day, in the afternoon, there's going to be a **football tournament** in the Amsterdamse Bos, organised by the Alumni organisation of BMI. I would like to take a team there and win that competition, too. If you want to play, please **let me know**, if you want to become a top supporter, please make sure you can come. As far as I know, the tournament will be followed up by a BBQ. If I have more details, I'll let you know.

Thanks for reading this long description, and I'd like to wish you **good luck**!