Statistical Learning **March 23, 2013**

Project I

Instructor: Jinzhu Jia and Yuan Yao *Due: April 11, 2013*

1 Requirement

1. Pick up ONE (or more if you like) favorite problem below to attack. If you would like to work on a different problem outside the candidates we proposed, please email course instructors about your proposal.
2. Team work: we encourage you to form small team, up to THREE persons per group, to work on the same problem. Each team just submit ONE report, with a clear remark on each person's contribution.
3. In the report, show your results with your careful analysis of the results. Remember: scientific analysis and reasoning are more important than merely the performance results. Source codes may be submitted through email as a zip file, or as an appendix if it is not large.
4. Submit your report by email or paper version no later than the deadline, to Teaching Assistants (TA), Xinyu Chen (xycbaker@gmail.com) and Bowei Yan (yanbowei@gmail.com).

2 Problem I (Regression): Keyword Pricing

The following data, collected by Prof. Hansheng Wang in Guanghua Business School at PKU,

`http://www.math.pku.edu.cn/teachers/yaoy/math2010_spring/Keyword/SE.csv`

contains two columns: the first column is a list of keywords; the second column is the profit value (positive for earning and negative for loss). Figure 1 gives some example.

'乌鲁木齐-阿克苏-机票'	14.1200
'乌鲁木齐阿克苏飞机票价'	9.0600
'乌鲁木齐到阿克苏-机票'	-1.1800
'乌鲁木齐到阿克苏打折机票'	-0.4800
'乌鲁木齐到阿克苏机票'	31.9400

Figure 1: Keywords and profit value

The purpose is to predict the profit value based on features extracted from the keywords, which might be linguistic, geographic, and any new features in your creation. Since the profit values are of real numbers, this problem is regarded as a regression problem by default.

3 Problem II (Classification): click-prediction

The way to access data is announced in class. Here we give one example for Linux users:

1. `sftp einstein@162.105.68.237`
2. `INPUT your password`
3. `cd /data/ipinyou/`
4. `get all_txt_data.zip` # containing all .txt data files, of size 170MB
5. `quit`

Unzip the file `all_txt_data.zip` to your local directory, you will find the following files:

- bid.20130301.txt: Bidding log file, 1.2M rows, 470MB
- imp.20130301.txt: Impression log, 0.8M rows, 360MB
- clk.20130301.txt: Click log file, 796 rows, 330KB
- conv.20130301.txt: Conversion log file, 1 rows, 809B.
- Region&citys.txt: Region and City code
- region_en.txt: region name in English
- region_zh.txt: region name in simplified Chinese

which are just some sample data from iPinyou. Further updates of training data will be released soon. A short introduction on the data format can be found at

`http://www.math.pku.edu.cn/teachers/yaoy/Spring2013/dsp_bidding_data_format.pdf`

As a warm-up, the first stage in the DSP-bidding competition is the so called *click-prediction*: find a model to predict the click of impressions. To be specific, the file `imp.20130301.txt` (0.8M rows, 360MB) contains 0.8M impressions after winning the bidding, among which only 796 impressions are clicked in the file `clk.20130301.txt` (796 rows, 330KB). The task is to design your features and predict the class of clicks.

For those R users, the following commands can be used to read the data

```
imp <- read.table("/data/ipinyou/imp.20130301.txt", sep='\t', comment.char='')
```

This is because that R `read.table`by default uses '#' as a comment character, that is, it has `comment.char = '#'` parameter by default. But the user-agent field in data file may have '#' character. Hence to read correctly, it is necessary to turn off `comment.char=''`.

One challenge lies in the imbalance of the problem, with 1000 zeroes for a single one. So this is a rare event classification problem. The following paper, Logistic Regression in Rare Events Data, by King and Zeng 2001, might be helpful for you.

`http://gking.harvard.edu/files/abs/0s-abs.shtml`

4 Problem III (Classification): Heart Operation Effect Prediction

The following data, provided by Dr. Jinwen Wang at Anzhen Hospital,

`http://www.math.pku.edu.cn/teachers/yaoy/data/HeartData_20130201.zip`

contains 2581 patients with 73 measurements (inputs) as well as a response variable indicating if after the heart operation there is a null-reflux state. This is a classification problem, with a challenge from the large amount of missing values.

The following two reports by LU, Yu and WANG, Qing, are probably inspiring to you.

`http://www.math.pku.edu.cn/teachers/yaoy/reference/LuYu_201303_BigHeart.pdf`

`http://www.math.pku.edu.cn/teachers/yaoy/reference/WangQing_201303_BigHeart.pdf`