Mathematical Introduction to Data Analysis

Final Project

Instructor: Yuan Yao *Due: Tuesday January 20, 2015*

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. Brave hearts for explorations will be encouraged!
2. The first two projects continue from the first project.
3. The datasets marked by $\star$ are a bit challenging due to its non-handy size, be careful.
4. 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.
5. In the report, show your results with your analysis of the results. Remember: scientific problem and analysis 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.
6. Submit your report by email or in paper version no later than the deadline, to Teaching Assistants (TA) (datascience_hw@126.com).
7. For those who need a server, you may connect to the Linux account `einstein@162.105.68.237` which is public to the students in this class (if you can't remember the password, please ask TA or me by email). Remember to make your own directory before starting creation of your own files. For example
   - `ssh einstein@162.105.205.92`
   - `INPUT your password`
   - `mkdir [your own directory]`

2 Project 1 datasets

The first two datasets can be either converted into character-cooccurance network, or directly processed as data matrix (time series of character activities). You may design problems with PCA and its extensions such as RPCA and SPCA, manifold learning such as diffusion map, and spectral clustering etc. Random projections can be also tested here. If you already have done something in project 1 and would like a follow-up to see improvement or deeper analysis, you are welcome!

2.1 The Characters in A Dream of Red Mansion

A 376-by-475 matrix of character-event can be found at the course website, in .XLS, .CSV, and .MAT formats. For example the Matlab format is found at

`http://www.math.pku.edu.cn/teachers/yaoy/data/hongloumeng/hongloumeng376.mat`

with a readme file:

`http://www.math.pku.edu.cn/teachers/yaoy/data/hongloumeng/readme.m`

Thanks to WAN, Mengting, an update of data matrix consisting 374 characters (two of 376 are repeated) which is readable by R read.table() can be found at

`http://www.math.pku.edu.cn/teachers/yaoy/data/hongloumeng/HongLouMeng374.txt`

She also kindly shares her BS thesis for your reference

`http://www.math.pku.edu.cn/teachers/yaoy/reference/WANMengTing2013_HLM.pdf`

2.2 A Journal to the West

On course website, you may also find the link to this dataset with a 302-by-408 matrix, whose matlab format is saved at

`http://www.math.pku.edu.cn/teachers/yaoy/Fall2011/xiyouji/xiyouji.mat`

For your reference, here is a project presentation by Mr. LI, Liying (at PKU) which gives an analysis based on PCA

`http://www.math.pku.edu.cn/teachers/yaoy/reference/LiyingLI_Xiyouji2012_slides.pdf`

2.3 Finance Data

The following data contains 1258-by-452 matrix with closed prices of 452 stocks in SNP'500 for workdays in 4 years.

`http://www.math.pku.edu.cn/teachers/yaoy/data/snp452-data.mat`

2.4 Hand-written Digits

The website

`http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/zip.digits/`

contains images of 10 handwritten digits ('0',...,'9');

You may try various methods to explore this dataset, from manifold learning to the state-of-

the-art deep neural networks.

2.5 ⋆ Air Quality Weibo Data

The dataset is provided by Prof. Xiaojin Zhu from University of Wisconsin at Madison. You can login my server:

```
ssh einstein@162.105.205.92
```

using the password I provided on class.

On the read-only folder `/data/AQweibo/`, the `AQICityData/` directory contains the Weibo posts, the AQI for 108 cities with (AQI) information during the study period from 2013-11-18 to 2013-12-18 (both inclusive); Information for the spatiotempral bin (city,date) is in the directory `city_date/`. See `README.txt` for more information.

Any easy project is to predict the AQI as time series; a more challenging task is to incorporate other features discovered from weibo texts.

2.6 ⋆ SNPs Data

This dataset contains a data matrix $X \in \mathcal{R}^{p \times n}$ of about $n = 650,000$ columns of SNPs (Single Nucleid Polymorphisms) and $p = 1064$ rows of peoples around the world. Each element is of three choices, 0 (for 'AA'), 1 (for 'AC'), 2 (for 'CC'), and some missing values marked by 9.

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

Moreover, the following file contains the region where each people comes from, as well as two variables `ind1` and`ind2` such that $X($`ind1`$,$ `ind2`$)$ removes all missing values.

```
http://www.math.pku.edu.cn/teachers/yaoy/data/HGDP_region.mat
```

Some results by PCA can be found in the following paper, Supplementary Information.

```
http://www.sciencemag.org/content/319/5866/1100.abstract
```

Attention: this last dataset is relatively big with about 2GB size.

You can login my server:

```
ssh einstein@162.105.205.92
```

using the password I provided on class. On the read only folder `/data/snp/`, you will find all the data in both .txt and .mat (`data.mat, HGDP_region.mat, readme.m`).

The dataset is a bit challenging in their high dimensionality $n = 650K$ which might not be handy to deal with your laptop. Random projections will be helpful here.

3 Heart PCI Operation Effect Prediction

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

`http://www.math.pku.edu.cn/teachers/yaoy/data/heartData_20141230.xlsx`

contains 2581 patients with 73 measurements (inputs) as well as a response variable indicating if the heart operation fails (null-reflux status). The data are collected based on 2 hospital groups, Anzhen Hospital and Chaoyang-301 Hospitals, indicated by the last column. This is a classification problem, with a challenge from the large amount of missing values. Sheet 3 and 4 in the file contains some explanation of the data and variables.

The problems are listed here:

1. The inputs (covariates) are of three kinds, measurements upon check-in, measurements before PCI operation, and measurements in PCI operations. For doctors, it is desired to find a prediction model based on measurements before the operation (including check-in). Sheet 2 in the file contains only such measurements.
2. It is an interesting problem to explore if there is any systematic difference between 2 hospital groups. Such a group effect has been addressed systematically by Andrew Gelman in his book, Multilevel models. But it is a good to start individual models for each of the two hospitals.
3. It is also an interesting problem how to predict the effect based on all measurements, with lots of missing values. Sheet 1 contains the full measurements. There are some good work by previous students, which are listed here for your reference:

   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`

   The following report by MIAO, Wang and LI, Yanfang, pioneers in missing value treatment.

   `http://www.math.pku.edu.cn/teachers/yaoy/reference/MiaoLi2013S_project01.pdf`

In the final project, it is desired to take only those measurements upon check-in to predict the probability of non-reflux (non-reflow) after PCI operations. An interpretable model adds a big value! You may compare with your early results to show your improvements. To evaluate your results, we suggest two ways

- 5-fold CV: to train any model, you may randomly split the data into 5 folds, with 3 fold for training, 1 fold for validation (optimizing parameter) and 1 fold for testing error (misclassification error). Currently the most state-of-the-art accuracy for all hospitals with pre-operation variables is about 85%, obtained by Dongming Huang, but with a large false-discovery rate compared to a small missing rate.
- AUC: compute the area-under-the-ROC, to avoid the unique choice of thresholds in classification.

4 World College Rating

- In the public server, the folder `/data/worldcollege` contains the following data

  `export_4271_ideas_20131117.csv`: 261 colleges in the world

  `export_4271_votes_20131128_n5k.csv`: about 5000 votes up to Nov 28, 2013

  `export_4271_non_votes_20131128.csv`: non-votes marked as "I can't decide" with various reasons

Explore this data with Hodge decomposition of paired comparison data. No one has ever tried it yet.

5 Social Network Data

Besides the two social networks constructed from two novels above, the following data are possible candidates for networks (from small to large). Problems like clustering, ranking, semi-supervised learning or transition path theories are all good ideas to pursue. You may explore them with different methodologies such as nonlinear Euclidean embedding and random walks on networks.

5.1 Stanford Large Network Dataset Collection

There are various datasets collected by Jure Leskovec

`http://snap.stanford.edu/data/`

where you may find undirected and directed graphs, possibly dynamic.

5.2 Mark Newman's Network Data

The following website collects Mark Newman's network data which are smaller than Jure's.

`http://www-personal.umich.edu/~mejn/netdata/`

6 Identification of Vincent van Gogh's paintings from the forgeries

The following data, provided by Dr. Haixia Liu from CUHK,

`http://www.math.pku.edu.cn/teachers/yaoy/data/vangogh-4.mat`

contains a 79-by-4 data matrix X, as 4 geometric-tight-frame features constructed from 79 paintings, in which the first 64 are attributed to Vincent van Gogh while the remaining 15 are forgeries. The IDs of those paintings are contained in string variable, `vg` and `nvg`, while the names of those paintings are listed in the file

`http://www.math.pku.edu.cn/teachers/yaoy/data/vangogh-info.pdf`

With Leave-One-Out test, our current state-of-the art classification accuracy with only these 4 features is 84%. Can you beat us?

7 Recommendation of VIP customers

The following dataset, provided by Prof. Dongdong Ge (gedong78@163.com) from Shanghai University of Finance and Economics,

`http://www.math.pku.edu.cn/teachers/yaoy/data/baixing/mydata.rda`

contains a 172,598-by-192 data matrix, as 172K phone calls made to customers from Baixing.com to upgrade their status to VIP with fees. The variable *mydata*$*success* indicates if the call is success (at rates about 3%). This is a classification problem to predict the success probability of VIP upgrade calls. The following files contain more information on variables and background problem (as I introduced in class)

`http://www.math.pku.edu.cn/teachers/yaoy/data/baixing/readme.xlsx`

`http://www.math.pku.edu.cn/teachers/yaoy/data/baixing/background.pptx`

The following file

`http://www.math.pku.edu.cn/teachers/yaoy/data/baixing/Rcode.txt`

contains some preliminary R codes with four basic models: linear models, logistic regression, decision trees, and random forests. The performance is measured by AUC, with the best achieved by random forests. Can you beat it?

8 * CTR (Click-Through-Rate) Prediction in Bidding Algorithm

This is a very challenging problem due to its big data size and rare event feature. The problem is similar to VIP recommendation above, though with only 1/1000 success rate. Original competition can be found from iPinYou Global Bidding Algorithm Competition at

`http://contest.ipinyou.com/`

where the full data (about 40GB) of 3 seasons can be downloadable at Baidu WebDrive

`http://pan.baidu.com/s/1kTkGUQN`

As part of the data, README file can be read from here:

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

More information can be found in class notes at `www.ebanshu.com`. If you have worked on this problem before, make a comparative study on how did you improve over previous work.