Statistical Learning

Final Project

Instructor: Yuan Yao *Due: Sunday December 27, 2015*

1 Requirement

1. Pick up ONE (or more if you like) favorite problem *below* or *from the data in textbook* 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 last few projects continue from the first project.

3. Team work: we encourage you to form small team, up to FIVE persons per group, to work on the same problem. Each team just submit ONE *poster* report, with a clear remark on each person's contribution. A sample poster file with PKU logo can be found at
`http://math.stanford.edu/~yuany/course/reference/poster_v5.pdf`
whose source LATEX codes can be downloaded at
`http://math.stanford.edu/~yuany/course/reference/pkuposter.zip`

4. In the report, show your results with your 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.

5. Submit your report by email or in paper version no later than the deadline, to Teaching Assistants (TA) (statlearning_hw@126.com). We plan a poster session during 3-6pm on Monday December 28 for peer reviews.

2 Heart PCI Operation Effect Prediction

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

`http://math.stanford.edu/~yuany/course/data/heartData_20140401.xlsx`

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. 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.

   The following two reports by LV, Yuan and LI, Xiao, respectively, might be interesting to you:

   `http://math.stanford.edu/~yuany/course/reference/MSThesis.LvYuan.pdf`

   `http://arxiv.org/abs/1511.04656`

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

   `http://math.stanford.edu/~yuany/course/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 first warm-up project to show your improvements.

3 Identification of Raphael's paintings from the forgeries

The following data, provided by Prof. Yang WANG from HKUST,

`https://drive.google.com/folderview?id=0B-yDtwSjhaSCZ2FqN3AxQ3NJNTA&usp=sharing`

contains a 28 digital paintings of Raphael or forgeries. Note that there are both jpeg and tiff files, so be careful with the bit depth in digitization. The following file

`https://docs.google.com/document/d/1tMaaSIrYwNFZZ2cEJdx1DfFscIfERd5Dp2U7K1ekjTI/edit`

contains the labels of such paintings, which are

1 Maybe Raphael - Disputed

2 Raphael

3 Raphael

4 Raphael

5 Raphael

6 Raphael

7 Maybe Raphael - Disputed

8 Raphael

9 Raphael

10 Maybe Raphael - Disputed

11 Not Raphael

12 Not Raphael

13 Not Raphael

14 Not Raphael

15 Not Raphael

16 Not Raphael

17 Not Raphael

18 Not Raphael

19 Not Raphael

20 My Drawing (Raphael?)

21 Raphael

22 Raphael

23 Maybe Raphael - Disputed

24 Raphael

25 Maybe Raphael - Disputed

26 Maybe Raphael - Disputed

27 Raphael

28 Raphael

Can you exploit the known Raphael vs. Not Raphael data to predict the identity of those 6 disputed paintings (maybe Raphael)? The following paper by Haixia Liu, Raymond Chan, and me studies Van Gogh's paintings which might be a reference for you:

`http://dx.doi.org/10.1016/j.acha.2015.11.005`

4 Ising Models for Biological Sequences

The problem is to estimate an Ising model for multiple aligned sequences of proteins in the same family. The data is provided by Dr. John Barton from MIT, in the following zip file,

`http://math.stanford.edu/~yuany/course/data/protein2014.zip`

where you will find

- pro-binary.dat: A set of 10579 binarized sequences, one sequence per row, taken from the real sequence database
- pro-model-binary.dat: A sample of 10000 binary sequences sampled from the model, in the same format as above
- pro-couplings.dat: The inferred model parameters

In the third model file, the first N=99 rows of the couplings file are the fields for sites 1 through 99, and the remaining N*(N-1)/2 entries are the couplings between sites, i.e. the entries are

h_1
h_2
...
h_{99}
$J_{1,2}$
$J_{1,3}$
...
$J_{1,99}$
$J_{2,3}$
...
$J_{98,99}$

People use different conventions for the energy function, so just to be clear the convention I am using is that the energy of a configuration $x = x_1, , x_N$, $x_i \in \{0, 1\}$ is

$$E(x) = -\sum_{i=1}^{N} h_i x_i - \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} J_{i,j} x_i x_j,$$

and the probability distribution over configurations x is $p(x) = \exp(-E(x))/Z$ with Z the partition (normalization) function.

This project is to learn an Ising model from multiple aligned sequences. This may contains the following 2 challenges

1. Learn an Ising model from simulated data, e.g. the second data file above with model in the third file. You may use 2 ways to evaluate your estimator: 1) the l_2 distance between the parameters you learned and the true parameters, or; 2) use your models to generate new sequences and test if the marginal distribution and correlation matrix meets the data.

2. Learn an Ising model from real data, e.g. the first data file. Only the second method above can be applied to evaluate your estimator in this setting, since you don't know the ground truth parameters.

You may explore Gaussian (corpula) graphical models for such sequence data.

(Hint) You may consider to try or compare the following algorithms:

- Xue-Zou-Cai's composite penalized conditional likelihood method,

  `http://www.personal.psu.edu/lxx6/softwares/sparseising_1.2.tar.gz`

- Minimum Probability Flow (MPF) method at ICML2013,

  `https://github.com/Sohl-Dickstein/Minimum-Probability-Flow-Learning`.

- Linearized Bregman Algorithms

  `https://cran.r-project.org/web/packages/Libra/index.html`

Note that the binary coding for Ising model in statistics is often $\{-1, 1\}$ rather than $\{0, 1\}$. A poster from early classes can be found at

`http://math.stanford.edu/~yuany/report/Poster07.Ising_Yichen_Zhang_Jiyi_Liu_Rujing_Zhao.pdf`

5 Co-appearance data in novels: Dream of Red Mansion and Journey to the West

A 374-by-475 binary matrix of character-event can be found at the course website, in .XLS, .CSV, .RData, and .MAT formats. For example the RData format is found at

`http://math.stanford.edu/~yuany/course/data/dream.RData`

with a readme file:

`http://math.stanford.edu/~yuany/course/data/dream.Rd`

`http://math.stanford.edu/~yuany/course/data/readme.m`

Thanks to WAN, Mengting, who helps clean the data and kindly shares her BS thesis for your reference

`http://math.stanford.edu/~yuany/report/WANMengTing2013_HLM.pdf`

Moreover you may find a similar matrix for the Journey to the West (by Chen-En Wu) at:

`http://math.stanford.edu/~yuany/course/data/west.RData`

6 Jiashun Jin's data on Coauthorship and Citation Networks for Statisticians

Thanks to Prof. Jiashun Jin at CMU, who provides his collection of citation and coauthor data for statisticians. The zipped data file (14M) can be found at

`http://math.stanford.edu/~yuany/course/data/jiashun/Jiashun.zip`

with an explanation file

`http://math.stanford.edu/~yuany/course/data/jiashun/ReadMe.txt`

You may feel free to explore this data using regression (e.g. logistic regression), discrete graphical models (e.g. Ising, Gaussian corpula graphical models), and/or other methods.

7 CTR (Click-Through-Rate) Prediction in Bidding Algorithm

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`

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. Remember to make your own directory before starting creation of your own files. For example

1. `ssh einstein@162.105.68.237`
2. `INPUT your password`
3. `mkdir [your own directory]`

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.

8 Keyword Pricing (Regression)

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

`http://math.stanford.edu/~yuany/course/data/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.

A sample study can be found at

`http://math.stanford.edu/~yuany/report/Poster06.Keyword.pdf`

9 Beer Popularity and Rating

The following data, provided by Mr. Richard (`sun.richard@yahoo.com`) from Shanghai,

`http://math.stanford.edu/~yuany/course/data/Beers_20140514.xlsx`

contains 877 brands (rows) of beers in Chinese market, with a few attributes about ingradients, alcoholicity, price (and unit price), reviewers count, mean scores, and as well as sources of reviewers (e.g. amazon, jd, yhd etc.). Two questions are interesting to explore such data

1. What factors are highly correlated with the popularity of beers indicated by reviewers count?
2. What factors accounts for the mean rating scores? Why are those beers lowly rated?

Note that the data does not contain lots of attributes, so think about your goal before you take a try.

10 Animal Species Sleeping Hours Regression

The following dataset contains $n = 51$ species with several features including the average sleeping hours per day.

`http://math.stanford.edu/~yuany/course/data/sleep1.csv`

Explore *what affects the sleep an animal needs*? A sample R code can be found at

`http://math.stanford.edu/~yuany/course/Fall2015/Lecture19.R`

You need to create your own design of analysis.

11 Crime Rate

Explore the following dataset about crime rates in 59 US cities during 1970-1992.

`http://math.stanford.edu/~yuany/course/data/crime.zip`

12 Radon Measurement

The following data set contains Radon measurements of 12,687 houses in U.S.

`http://math.stanford.edu/~yuany/course/data/radon.csv`

13 Switch unsafe wells

The following data set contains decision of switching unsafe wells for arsenic pollution in Bangladesh.

`http://math.stanford.edu/~yuany/course/data/wells.csv`