Statistical Learning

Final Project

Instructor: Yuan Yao *Due: Saturday June 21, 2013*

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 first two projects continue from the first project.

3. 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 *poster* report, with a clear remark on each person's contribution. A sample poster file with PKU logo can be found at
`http://www.math.pku.edu.cn/teachers/yaoy/reference/poster_v5.pdf`
whose source LATEX codes can be downloaded at
`http://www.math.pku.edu.cn/teachers/yaoy/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 on Saturday June 21 (evening) for peer reviews.

2 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_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.

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 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 first warm-up project to show your improvements.

3 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`

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

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

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@162.105.205.92`
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.

4 Keyword Pricing (Regression)

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.

A reference can be found in Mr. Jiaqi Zhu's bachelor thesis work:

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

5 Beer Popularity and Rating

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

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

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

The following data, provided by Ms. 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 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://www.math.pku.edu.cn/teachers/yaoy/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.

(Hint) You may consider to try Xue-Zou-Cai's composite penalized conditional likelihood method. But I would recommend a recent ICML 2013 paper about Minimum Probability Flow (MPF) Learning method, downloadable at

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

MPF is superfast! Try it, and you won't be disappointed!

8 *Social Network Data: 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`

Among various choices of analysis, with this data matrix X, you may form a weighted graph $W = X * X'$, pursue PCA of X, and sparse SVD of X etc. As an example, here is a project presentation by LI, Liying which gives an analysis of A Journal to the West (by Chen-En Wu) based on PCA, for the class Mathematical Introduction to Data Science in Fall 2012 where you may find more interesting approaches.

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

On course website, you may also find the link to this dataset.

9 *Neural Network and Deep Learning

The following project on deep learning for reconstructing a 2-D Gaussian Mixture Model, is proposed by Dr. Lei Jia from Baidu and posted on page 25-30 in my lecture slides

`http://www.math.pku.edu.cn/teachers/yaoy/Spring2014/Lecture13.pdf`

For those who are interested in Restricted Boltzman Machine and MNIST experiments, Hinton's matlab codes are good demonstration

`http://www.cs.toronto.edu/~hinton/MatlabForSciencePaper.html`