Statistical Learning

Final Project

Instructor: Jinzhu Jia and Yuan Yao *Due: June 30, 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. The first three 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 report, with a clear remark on each person's contribution.
4. 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.
5. Submit your report by email or in 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.

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.

If you have worked on this problem before, make a comparative study on how did you improve over previous work.

'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

3 Problem II: CTR (Click-Through-Rate) Prediction

Original data can be downloaded from iPinYou Global Bidding Algorithm Competition at

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

where Stage II started on June 2nd and will last through September 30th. To register the contest and submit your algorithm, you need to fill in your algorithm in a Java interface found at

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

Part of data is stored in the server for students in this class. Here we give one example to access the data 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`

The problem of *click-through-rate prediction*: find a model to predict the probability of 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`

For those who already worked on this problem, you might consider extensions to online logistic regression (e.g. `http://hunch.net/~jl/projects/interactive/sparse_online/paper_sparseonline.pdf`).

We invited the winning team of Stage I, `ml_rush`, to give a talk in our class on June 6th, which might interest you if you wanna take part in the contest.

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`

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`

5 Problem IV (Graphical Model): Protein Folding Prediction by Sequences

The problem is to predict the *contact map* of proteins by multiple aligned sequences in the same family. Three examples are given in the data

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

where you will find PF00013 (PCBP1_HUMAN/281-343, PDB 1WVN), PF00018 (YES_HUMAN/97-144, PDB 2HDA), and PF00254 (O45418_CAEEL/24-118, PDB 1R9H). Data format information can be found at

`http://www.math.pku.edu.cn/teachers/yaoy/data/protein/readme.txt`

For example, file `http://www.math.pku.edu.cn/teachers/yaoy/data/protein/sequence/PF00018_match.aln` contains 3610 sequences of length 48 for the same family PF00018, where the first sequence is

`----------------ENEIVQVFSIVDESWWSGKLRRNGAEGIFPK`

Here

- `-` denotes the gap,
- other alphabets denotes the Amino Acid code, from 20 characters.

Therefore in total the sequence is coded by 21 characters. Correspondingly file `http://www.math.pku.edu.cn/teachers/yaoy/data/protein/sequence/PF00018_2HDA.pdb` contains the 3D coordinates of alpha-carbons for a particular amino acid sequence in the family, YES_HUMAN/97-144, read as

`VALYDYEARTTEDLSFKKGERFQIINNTEGDWWEARSIATGKNGYIPS`

where the first line in the file is

97 V 0.967 18.470 4.342

Here

- '97': start position 97 in the sequence
- 'V': first character in the sequence
- $[x, y, z]$: 3D coordinates in unit $\mathring{A}$.

Figure 2 gives the 3D representation of its structure.

Given the 3D coordinates of the amino acids in the sequence, one can computer pairwise distance between amino acids, $[d_{ij}]^{l\times l}$ where l is the sequence length. A *contact map* is defined to be a graph $G_\theta = (V, E)$ consisting l vertices for amino acids such that and edge $(i, j) \in E$ if $d_{ij} \leq \theta$, where the threshold $\theta = 8\mathring{A}$ here.

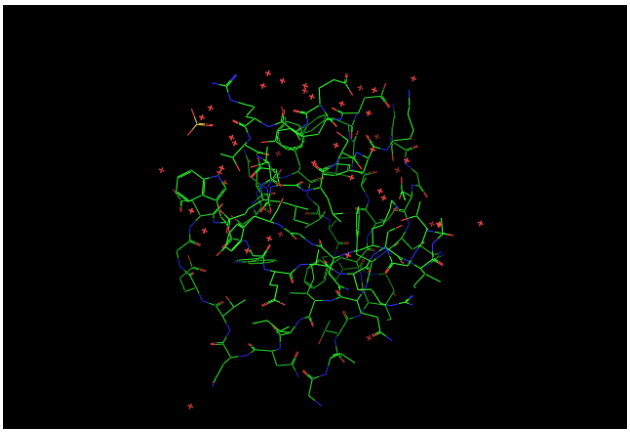

Figure 2: default

Non-local contact map $G_{\theta,\tau}$ considers the restricted contact map with only edges (i, j) with i and j are τ-separated way in sequence distance. Here we choose $\tau = 5$, *i.e.* $|i - j| > 5$.

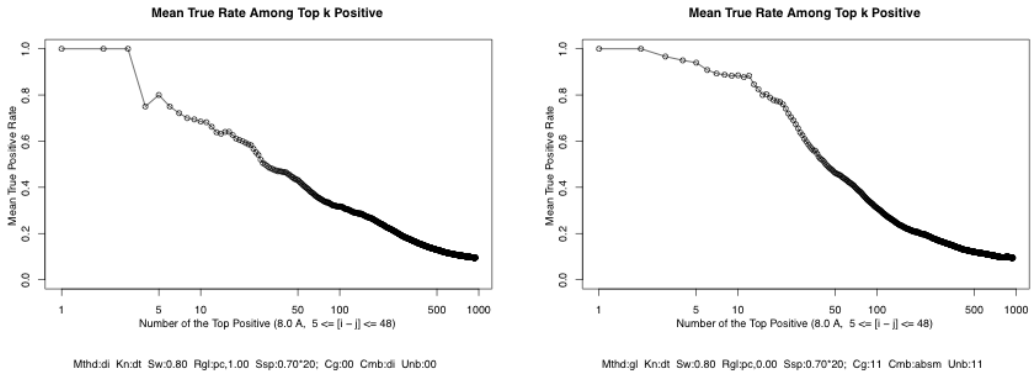

Figure 3: True Positive Rates on non-local contact predictions by Directed Information vs. Graphical Lasso on Yes_Human, courtesy by Chendi Huang, indicating that graphical lasso performs better.

This project is to learn a graphical model from multiple aligned sequences, to predict the non-local contact map $G_{\theta=8\mathring{A},\tau=5}$. Performance is evaluated in terms of the fraction of correct predicted non-local contacts (true-positive-rates) among the top k pairs with highest scores, *e.g.* $k = l/5, l/3, l/2, l$, etc. Figure 3, courtesy by Chendi Huang, gives you a reference on comparing the Directed Information by Morcos and the graphical lasso. For your reference, Chendi's report can be found at

`http://www.math.pku.edu.cn/teachers/yaoy/reference/Huang_protein_report_2013-04-28.pdf`

6 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.

7 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 24-29 in my lecture slides

`http://www.math.pku.edu.cn/teachers/yaoy/Spring2013/Lecture11.pptx`

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`