Introduction to Data Science (Shanghai Jiaoda)

Course Project

Instructor: Yuan Yao *Due: July 28?, 2013*

1 Requirement

1. Pick up ONE (or more if you like) favorite problem below to attack.]
2. 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. Give a good reference on those reports you benefit from as well as credits to your peers who collaborate on the same project.
3. Submit your report to TAs and Instructor by email no later than the deadline.
4. TAs: Haixia Liu (hxliu@math.cuhk.edu.hk) and Yaoyu Zhang (agatespace@qq.com)

2 Problem I (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, 42 of them are before the operation) 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. We have the following two tasks:

- Use all the measurements to predict the null-flux status;
- Use only the 42 measurements before the operation to predict the null-flux status.

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`

I may send you more reference on this problem.

3 Problem II (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 $\mathring{A}$.

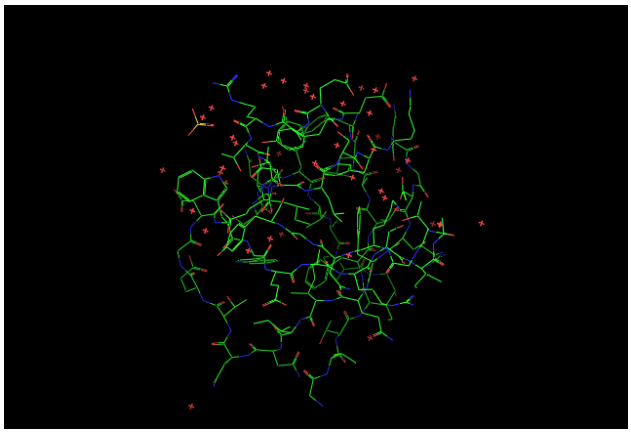

Figure 1: default

Figure 1 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.

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.* $|i - j| > 5$.

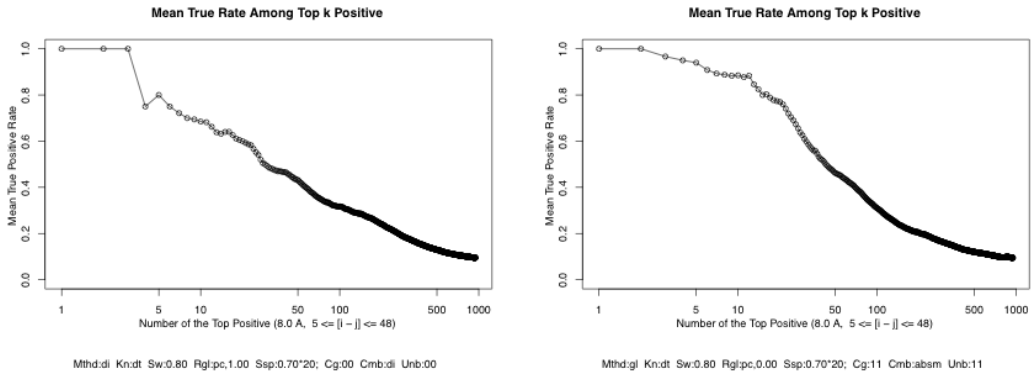

Figure 2: True Positive Rates on non-local contact predictions by Directed Information vs. Graphical Lasso on Yes_Human, courtesy by Chendi Huang, indicating that graphical lasso performs better.

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 2, 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`

With other references can be found on my course web: Lecture 13 at `http://www.math.pku.edu.cn/teachers/yaoy/Spring2013/`

- Marcos et al., Direct-coupling analysis of residue coevolution captures native contacts across many protein families, PNAS, 2011, 108(49): E1293-E1301.
- Jones et al. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments, Bioinformatics. 2012, 28(2):184-90.
- Ravikumar, Wainwright and Lafferty (2010). High-dimensional Ising model selection using l1-regularized logistic regression. Ann. Stat. 2010, 38(3): 1287-1319.

4 Problem III: Social Network Data: Dream of Red Mansion and A Journey to the West

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 as well as the dataset.

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