Yale University

From the SelectedWorks of Shuangge Ma

June 6, 2009

A Tale of Two Streets: Incorporating grouping structure in high dimensional data mining

Shuangge Ma, Yale University



A Tale of Two Streets

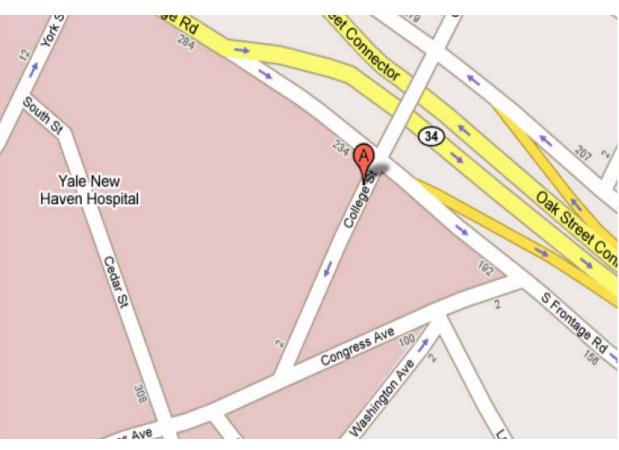
Incorporating grouping structure in high dimensional data mining

马双鸽

Assistant Professor School of Public Health, Yale University

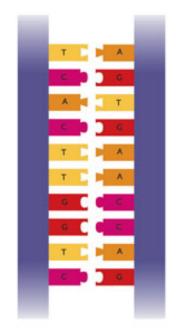
College Street (New Haven, CT) School of Public Health, Yale University





Cancer Informatics

- Genetic mutations and defects are responsible for cancer occurrence and progression
- Our research: Predict occurrence of cancer using genetic/genomic measurements obtained through genome profiling





Cancer Informatics

 Occurrence of cancer: Yes/No

Human Genome:

- 23 chromosomes
- 20,000+ genes
- 250,000+ SNPs (single nucleotide polymorphisms)
- 6,000,000,000 basepairs

Cancer Informatics: A high dimensional data mining problem

- Consider the 20,000+ genes
- · Our problem: Binary classification/prediction
- Data characteristics:
 - High dimensionality
 - Most genes are just noises
 - Some genes work together

Wall Street (Manhattan, NY)



CITY London M Herald

FRIDAY 25% OCTOBER 1929

WALL STREET CRASH!

Black Thursday in America

Stocks Plunge and Eleven Commit Suicide

Principaliting bandle New York New York Market person the execute the thirt series the perfect of a per-

Married with territorise's participation of a property erform attemption of obtain and punis. Restore his but the market with and real form their liganous formelt allump petition.

to the situation and bed with bird, the market balls (bear situate) with the of balls of the second marks for the content layout (balls over trade) for any principal the value of well companies which during the process of the successing

problems of National State Special Sections and the Laters Eastly was with hadroneously support of the matter section of the the said or 10 the hadroneously the section of the the said or 10 the hadroneously the section of the the said or 10 the hadroneously the said or 10 the hadroneously the said of the said or 10 the hadroneously the said of the sai According to America

If Margan A Co and mitter the Book State therefore term place recovered significant terms of the Book State therefore the Parties of the Book State St

the high suly enemy. Wallington and bound STATE OF THE PERSON. See hers belong offi-Aprilia maning of him in claim the proposals or free landing modern cost mens most pool manifestic free flam of model for plansional

parties and rectand part, to dark actions obtained parties and demands of their The final techniques of the second se



What Went Wrong?

No. College Street Season Additional the Studies.

att of dala legal when it became about m. Se. one competitive.

the attenuency chapt has, in part throat by bestern in

desire make a me

Who's bud obscription. Subspines - Southern

cons, related of mile. City of New York, Fig. 1950s, we appealant the reside the realing operational generation. Some Finished deal that day, the Subar recording is both at those as in . Next restore and dealers.

he do doughed by depoint of he sale. Hearth Reports gill

Renmin L

Mortgage [A "simple" question] Are you going to default?

	4411				
Culturu Residential Loan Application		Section Chair States	pr Tourism III	the Carry	Spine Stational
to again to a support to to expect the agreement of the fraction agreement and require the first of the confirmation agreement and the confirmation are the confirmation and the confirmation and the confirmation are the confirmation and the confirmation and the confirmation are the confirmation and the confirmation and the confirmation are the confirmation are the confirmation and the confirmation are the			West Come		World Communication
Assistance in Regard to the content to the principal of the forest contents about contents that the content of the forest content of		headled and head	Total	Table (See	tente.
and it is easily purple and it is because of the purple and and a purple and a purp					territo (ci mod)
		Section Control	Total Comm		Ball has
Region 216 (Throndon St Throndon State Sta		hate/fortpartness	name Park	The state of the s	tanto Mand
Sangton				The state of the s	
the state of the s		Sering book* 1 1		to Magazini	
April pol (alle (as 4) and E)		Name -		to form (III)	
springing Pringing (and design forward)		Total officers		tertractan	
North Chair Chairm Chairman Ch		20.00		boson to the	
Spirit Frank residence in		See 1	-		
Total Signature Security of State State Class State State Class State State Class State State Class State St		Total Charles			
Nagati Mariji Ma		-1	The Server III a Parks		lane.
The Superfice Secretarian Superfictions Superfictions State Superfictions State Superfictions State Superfictions Superfictions State Superfiction Superfiction Superfiction Superfiction Superfiction Superfiction Superfic		_			
1 1 54				-	
No of the date and Ready Section (No. of Section Sec				and the fallowing state or support	all the last of th
terror fina (para) filation (bags sale Neskat (basis) page)				Negati	Charlest Charles
A STATE OF THE STA			and the County		-
Manufacture Manufacture Manufacture (Manufacture Manufacture Manuf		C Colonial Parished P	10000		
Market State Control of		Total and the Control of the Control	Securities of Super	The second	
The State of Particular Control of Particula					
Titlean of Timestricides () Septiment printed by the second of Timestricides () Septiment of the second of timestricine () Septiment of timestricine ()		ten I	-		
Translation (s. ed. 17) 17 to 17 t		No. of the Co. of the Co.	Non-selviden d'Imper	Hypothesis	
(mile para displacion para april 6 (delay) Anni Alexandria, des (M) — (Mile — Mile —		Service Co. M Seller	Security of Super	Heaten	•
Name and Associated State of S					
Topication. Topic					
		ALCOHOL: SALES	************	_	Seattle to the seattl
to top to top top top top top top top to		2		- 10 h	
Production on Control of the Control		The State of the S	to had also peeced by making peaks		
		I have departed the parties		:: ::	
Management Service Ser					
		physical desir			
New office (March 1997) Programme Transfer (March 1997) Progra	· beautiful	Contract and place			
The state of the s		- Bayanta materia			
Manufacture of the Control of the Co	- Martin Science Street	Of the last of the			
The state of the s		N. A. Princery and Desire of Arts			
Tel September 1 Se					
Name And State State And S	-	which the complement of the self-dependent of The S. Take and Court of September 1 and September 1 and September 1 of Stating a September 1 and September 1 and September 1 of September 1 and September 1 and September 1 and September 1 of September 1 and September 1 and September 1 and September 1 of September 1 and September 1 and September 1 and September 1 of September 1 and September 1 and September 1 and September 1 of September 1 and September 1			
Market Annual Parket Annual Pa	died desprie to desprié la	at the section process to be because the party			
See	Control of the last of the las		7 market 1 m	Sent Surger of La	
The test based 3 MARKAGE (15 on one one of 1	described by the con-	agent besty activistic garder any other of the hour, is not to prove to the hour, to any legislate for the proper benefit any come, but the		and a second of the second of	
Technic hours 1	Security Space	Ton School	- System	**	
Territoria 2 Section 2 Sec	Salating should be appear		and the second below to the second se	and the latest two says.	
100 algorithms (100 algorithms				Type to all the base to	
M Non-Samples Sales Samples Sales Sa	AND THE PERSON		THE R. LANSING MICH.		
	Name of Street of Street of Street Original St	The Sharthetone See	Section 1 to 1	Clin bein	
The second of th	The fact was	-	Section 1984		
And the second s	Service and a se	Secretar's New Jose of Sal	No. of Street,		
The state of the s	Total Control	Santa Again	-		
- Santa 1 Santa Sa		Secret Section (at secre			
Manufal M					
T MODIFICATION OF THE PROPERTY	tentered to	Term .		*******	

Mortgage (as well as many other business sectors)

· Default: Yes/No

Applicant's information:

- 1. Financial situation
- 2. Education
- 3. Family members' information

..... A long list of variables from questionnaires

Cancer/Mortgage A high dimensional data mining problem

- · Our problem: Binary classification/prediction
- · Data characteristics:
 - High dimensionality
 - Most measurements are just noises
 - Some measurements work together

Cancer/Mortgage A high dimensional data mining problem

- Our problem: Binary classification/prediction
 Construct binary classification models
- · Data characteristics:
 - High dimensionality
 - Most measurements are noises
 - Variable selection to reduce dimension & remove noises
 - Some measurements work together Account for the grouping structure!!!

Penalized Regularization in Data Mining

- Binary outcome Y = 0, 1
- Covariate $X = (X^1, X^2 \dots X^K)$
- X^i is the m_i dimensional measurements of the i^{th} group (of variables)
- For example: X^i =(year of education; highest degree; area of education...)

Penalized Regularization in Data Mining

• Assume the logistic regression model $Pr(Y = 1|X) = logit(\sum_{i=1}^{K} \beta^i X^i)$. Denote $\beta = (\beta^1 \dots \beta^K)$

• Assume *n* iid observations

• Log-likelihood
$$R_n(\beta) = \sum_i Y_i \log \left(\frac{\exp(\beta X_i)}{1 + \exp(\beta X_i)} \right) + (1 - Y_i) \log \left(\frac{1}{1 + \exp(\beta X_i)} \right)$$

Penalized Regularization in Data Mining

- Penalized regularization $\hat{\beta} = argmaxR_n \lambda_n Pen(\beta)$
- $Pen(\beta) = \sum_{i} ||\beta^{i}||_{2}^{\gamma}$ with $0 < \gamma < 1$

So what we do is:

- ·We maximize the likelihood function
- But, we add a penalty, which gives us a penalized likelihood function
- Most recently, penalized data mining is one of the hottest areas

So, what's the deal with $Pen(\beta) = \sum_i ||\beta^i||_2^{\gamma}$

- · It is called the "group bridge" penalty
- When sample size < number of covariates, it can lead to a unique estimate > Regularization
- Many of the estimates are zero→Selection
- The penalty is on the "group norm". Thus, it can account for the grouping structure

We bother with group bridge because ...

- It is estimation consistent
- It is variable selection consistent, meaning: we can separate "important variables" from "noisy ones"
- So in the future, instead of asking for a huge number of questions, we can ask only a few
- It saves time and cost

We bother with group bridge because ...

 Variable selection consistency + estimation consistency. The Oracle is with us!



We bother with group bridge because ...

- Many other data mining methods
 - cannot account for the grouping structure
 [you lose efficiency]
 - Although they may work well with a few datasets, without statistical backup, it is hard to say how well they work in general

Some non-trivial tasks

- Computing [the objective function is not concave]
 - An iterative approach has been developed
- Tuning parameter selection
 - Cross validation
- Evaluation of significance and reproducibility
 - Occurrence index

All right. Now the real question:

- · How much do we really gain in practice?
- No one likes a toy that works only "in theory"
- In extensive numerical studies, the group bridge can beat alternatives:
 - A few percent in classification error
 - A lot in variable selection

What's next

- On College street: let's cure cancer!
 - Diagnosis and prognosis gene signatures are being constructed for lymphoma, breast cancer, leukemia ...
- On Wall street: let's make money! Or at least lose less
 - Efficient penalized data mining tools are being introduced

Acknowledgements

- · Conference organizers and RenMin University
- Funding support from NIH/NSF, USA
- Main collaborators: Dr. Jian Huang, University of Iowa

Thanks for your attentions!