Introduction to data mining

Ryan Tibshirani Data Mining: 36-462/36-662

January 15 2013

Logistics

- Course website (syllabus, lectures slides, homeworks, etc.): http://www.stat.cmu.edu/~ryantibs/datamining
- We will use blackboard site for email list, grades
- 6 homeworks, top 5 will count
- 2 in-class exams, final project
- Final project in groups of 2-3, will be fun!

Prerequisites:

- Only formal one is 36-401
- Assuming you know basic probability and statistics, linear algebra, R programming (see syllabus for topics list)

Textbooks:

- Course textbook Introduction to Statistical Learning by James, Witten, Hastie, and Tibshirani. Get it online at http://www-bcf.usc.edu/~gareth/ISL, use the login name: StatLearn, password: book. Not yet published, do not distribute!
- Optional, more advanced textbook: Elements of Statistical Learning by Hastie, Tibshirani, and Friedman. Also available online at http://www-stat.stanford.edu/ElemStatLearn

Course staff:

- Instructor: Ryan Tibshirani (you can call me "Ryan" or "Professor Tibshirani", please not "Professor")
- ► TAs: Li Liu, Cong Lu, Jack Rae, Michael Vespe

Why are you here?

- Because you love the subject, because it's required, because you eventually want to make \$\$\$...
- No matter the reason, everyone can get something out of the course
- Work hard and have fun!

What is data mining?

Data mining is the science of discovering structure and making predictions in (large) data sets

Unsupervised learning: discovering structure

E.g., given measurements $X_1, \ldots X_n$, learn some underlying group structure based on similarity

Supervised learning: making predictions

l.e., given measurements $(X_1,Y_1),\ldots (X_n,Y_n),$ learn a model to predict Y_i from X_i

Google



Search



Ads

800	Email - Inbox (2) - ryantibaljigmail.com	
A > 1 + Messol	mail google.com i mail /hab-wardinbox	
to () III Apple No	Anni EnnyleMays VouTake Wilcipedia Henry (2,125) + Pepular +	
You Gmail Calendar	Decuments Photos Sites Search More - nyumber@gmail.com	- ¢
Caroli	Easth Mall Basets the Web Streamphatics	
Citran A	Great's patting a new look appr. Learn more Diamina	
Mal	Constal University - years Constal only - Earn Your Tennes Online Error an Armedited University - Learn More, We have	
Contacts	Collected Test Description (B) Collected States (March 1971)	
14545		
Compose mail	C Paliye Shaper Intramural Basketball league and Equitible to Wed Jan 18th 11:00pm Vicio 1	9.68 pm
Johns (B)	C (wild, we (1) New Journal of the Royal Statistical Society - Thanks, Ryan, I look forward to m t	5.33 pm
Started 12	C Casma, we (2) 111-112 Possible example for 402 - Thankal Looks cool. On Skin, Jan 15, 2012 at 1	1.65 am
Bert Mal	🔝 🗇 Rebenne Nagert 🕴 Classification Basiety Meeting June 14-10, 2012 - Hole all, Weisserve in 2012) Hope 🥒	Jan 14
Orafta	1 🖸 🗇 me, rickaj (16) Fedal - The goal is to find a good regularization parameter for each-group of predicto 🖋	Jan 13
Arbeirerh	C DiscoverMagazine This Week's Best, Stanilating Black Holes, Super Strang Silk, and Viruses Without	Jan 13
CMJ	C ne, Carl (8) Home dr on the hydraw? - His, very attorney If have to shock further. On Jan 12, :	Jan 12
Owl	I Steet, me, Lans, obert /d	Jan 11
Donations	🖂 🗋 Large and 75	Jan 11
ENAR	C black Seattish Annual Departs - downwritelawd Chrillioticial wasawrianail actediral insurantement of	Jan 18
Orante	Contraction Reside and a second state of the s	in the second se
Heleg	C C And Ballion And And And And And And And And And An	the state
in the second se		
Papers to read	CO II Charle Instantes Elbow - Grounds for priorite P	
Natoree	The state free free state and the state and the state state and the state state and the state state and the state state state and the state state state and the state st	a 2411
\$12-602	1 🗆 1 Rob, Viece, me, Zhanwa (11) 🔹 estimating partial consistion/highession with large p - Dear all, 1 connected the code a S 🖉	32911
Tax	Larg, me (3) Pollowage - I would keep taking the daily folic acid, at least for next few weeks. Original	01611
Travel	C R Anirel, me (K) Re: 20 fused lasse code - H Pyon, Could you please send me the code Council in Decord	-
COMP comments		



Gmail

Chrome

Facebook

	Fine	d Your Friends on Facebook	Q- Google	E ^R
6-0 🛄 App	ole Yahoo! Google Maps YouTube Wikipedia N	ews (857) * Popular *	, angle	
facebook	🙏 💭 🛞 Search	٩	🧱 Jessica Issler	Home 👻
People You N	lay Know	See All		
a to a second	Edward Clapp Ema Hs and 3 other mutual friends	+1 Add Friend		
1	Steve Carlson San Jose State University Melody Kennedy and 20 other mutual friends	+1. Add Friend		
-	Alexander Koller Hanover, Germany Peter Meyer and 13 other mutual friends	+1 Add Friend		
	Anke Heinen Sylvia Kraft and 15 other mutual friends	+1 Add Friend		
3	Donald Feasel Melody Kennedy and 17 other mutual friends	+1. Add Friend	1 ● Chat (13)	

People you may know

Netflix

		Net	flix Prize: View Lead	derboard			
🕨 🕂 🚖 http:/	//www.netflixprize.c	om//leaderboard?showtest=t&limit	=20		C Q+ netflix prize	2	
Apple	Yahoo! Google Ma	os YouTube Wikipedia News (2,	,326) * Popular *				
	NETFLI)	i da la companya da l					
	Netflix Prize						
	Home Rules	Leaderboard Update				_	
	Display top	20 🕴 leaders.					
		-					
	Rank	Team Name	Best Test Score	% Improvement	Best Submit Time		
	Rank Grand Pr	Team Name ize - RMSE = 0.8567 - Winning Te	Best Test Score	% Improvement	Best Submit Time		
	Rank Grand Pi 1	Team Name ize - RMSE = 0.8567 - Winning Te ellKor's Pragmatic Chaos	Best Test Score am: BellKor's Pragr 0.8567	% Improvement matic Chaos 10.06	Best Submit Time		
	Rank Grand P 1 E 2]	Team Name ize - RMSE = 0.8567 - Winning Te lellKor's Pragmatic Chaos he Ensemble	Best Test Score eam: BellKor's Pragn 0.8567 0.8567	% Improvement matic Chaos 10.06 10.06	Best Submit Time 2009-07-26 18:18:28 2009-07-26 18:38:22		
	Rank Grand P 1 E 2] 3 G	Team Name ixe - RMSE = 0.8567 - Winning Te leliKors Pragmatic Chaos he Ensemble irand Prize Team	Best Test Score eam: BellKor's Pragn 0.8567 0.8567 0.8582	% Improvement natic Chaos 10.06 10.06 9.90	Best Submit Time		
	Rank Grand P 1 5 2 7 3 6 4 6	Team Name ize - RMSE = 0.8567 - Winning Te tellKor's Pragmatic Chaos The Ensemble 'and Prize Team 'pera Solutions and Vandelay United	Best Test Score eam: BellKor's Pragn 0.8567 0.8567 0.8562 0.8588	% Improvement matic Chaos 10.06 10.06 9.90 9.84	Best Submit Time		
	Rank Grand P 1 5 2 1 3 0 4 0 5 5	Team Name ize - RMSE = 0.0567 - Winning Te teliKor's Pragmatic Chaos he Ensemble izand Prizz Team ipera Solutions and Vandelay United iandelay Industries I	Best Test Score 0.8567 0.8567 0.8582 0.8588 0.8591	% Improvement matic Chaos 10.06 10.06 9.90 9.84 9.81 9.81	Best Submit Time 2009-07-26 18:18:28 2009-07-26 18:38:22 2009-07-10 21:24:40 2009-07-10 01:12:31 2009-07-10 00:32:20		
	Rank Grand P 1 E 2 3 3 G 4 G 5 3 6 E	Team Name	Best Test Score am: BellKor's Pragn 0.8567 0.8567 0.8582 0.8588 0.8591 0.8594 0.8594	% Improvement matic Chaos 10.06 9.90 9.84 9.81 9.77 0.70	Best Submit Time		
	Rank Grand P 2] 3 G 4 G 5 S 6 F 7 F	Team Name ite - RMSE = 0.8567 - Winning To lelikors Pranmate Chaos hand Prizz Team para Solutions and Vandelay United 'andelay Industries.j tramate/Theory elikor in BigChaos	Best Test Score 0.8567 0.8567 0.8567 0.8582 0.8581 0.8591 0.8594 0.8601 0.850	☆ Improvement matic Chaos 10.06 9.90 9.84 9.81 9.77 9.70 0.50	Best Submit Time		
	Rank Grand P 2] 3 (2 5) 6] 7] 8] 8] 9]	Team Name	Best Test Score 0.8567 0.8567 0.8567 0.8582 0.8588 0.8594 0.8594 0.8601 0.8512 0.8612 0.8612	☆ Improvement matic Chaos 10.06 9.90 9.84 9.81 9.77 9.70 9.59 0.49	Best Submit Time 2009-07-26 18:18-28 2009-07-26 18:38-22 2009-07-10 21:24:40 2009-07-10 11:231 2009-07-20 10:32:20 2009-06-24 12:06:56 2009-07-24 17:16:43 2009-07-24 17:16:43		
	Rank Grand P 1 5 2 2 3 2 4 2 5 3 6 5 7 5 8 5 9 5	Team Name	Best Test Score am: BellKor's Pragr 0.8567 0.8567 0.8588 0.8591 0.8584 0.8591 0.8594 0.8601 0.8612 0.8622 0.8622 0.8622	★ Improvement matic Chaos 10.06 10.06 9.90 9.84 9.81 9.77 9.70 9.59 9.48 0.47	Best Submit Time 2009-07-26 18:18:28 2009-07-26 18:18:28 2009-07-26 18:82:2 2009-07-10 01:12:31 2009-07-10 01:12:31 2009-07-20 01:22:0 2009-06-24 12:06:56 2009-07-21 13:11:51 2009-07-12 13:11:51		
	Grand P 1 5 4 5 6 6 7 6 8 6 9 6 10 6	Team Name Eta = NBS = 0.055 - Vinning Te Mission Statution (Statution) In Statution In	Best Test Score am: BellKor's Pregr 0.8567 0.8582 0.8582 0.8584 0.8594 0.8601 0.8612 0.8622 0.8623 0.8623	⅔ Improvement matic Chaos 10.06 9.90 9.84 9.81 9.77 9.70 9.59 9.48 9.47	Best Submit Time 2009-07-26 18:18:28 2009-07-26 18:38:22 2009-07-10 21:24:40 2009-07-10 01:12:31 2009-07-10 10:32:20 2009-05-41 20:656 2009-05-13 08:14:09 2009-07-24 17:18:43 2009-07-24 13:11:51 2009-07-24 13:11:51		

\$1M prize!

eHarmony



Falling in love with statistics

FICO



An algorithm that could cause a lot of grief

FlightCaster

00		Flight	Caster			
- □	│	edia News (2,327) * Popu	ılar v	C Q ⁺ eharmony.com	O	
Iback	Flight Caster	iction	Home Abou	it Sample Blog FAQ Contact Flight No O By Route		
Feed	Control of the second s		Airline Number Date	Today Gun Jan 15th)		
	Get the apps!	What people are saying		Travel Delays De-Mystified		
	Available on the App Store	THE WALL STREET JOUR frequent fliers car flights earlier and o	NAL. n know to rebook ccasional fliers can	Why shouldn't I rely entirely on airlines or other alert systems?		

Apparently it's even used by airlines themselves

IBM's Watson



A combination of many things, including data mining

Handwritten postal codes



(From ESL p. 404)

We could have robot mailmen someday

Subtypes of breast cancer

Fig. 1. Performance of a "wound response" gene expression signature in predicting breast cancer progression



Chang, Howard Y. et al. (2005) Proc. Natl. Acad. Sci. USA 102, 3738-3743



Subtypes of breastcancer based on wound response

Predicting Alzheimer's disease



(From Raji et al. (2009), "Age, Alzheimer's disease, and brain structure")

Can we predict Alzheimer's disease years in advance?

Banff 2010 challenge



Find the Higgs boson particle and win a Nobel prize! Will it be found by a statistician?

What to expect

Expect both applied and theoretical perspectives \dots not just a course where we open up R and go from there, we also rigorously investigate the topics we learn

Why? Because success in data mining comes from a synergy between practice and theory

- You can't always open up R, download a package, and get a reasonable answer
- Real data is messy and always presents new complications
- Understanding why and how things work is a necessary precursor to figuring out what to do

Reoccuring themes

Exact approach versus approximation: often when we can't do something exactly, we'll settle for an approximation. Can perform well, and scales well computationally to work for large problems

Bias-variance tradeoff: nearly every modeling decision is a tradeoff between bias and variance. Higher model complexity means lower bias and higher variance

Interpretability versus predictive performance: there is also usually a tradeoff between a model that is interpretable and one that predicts well under general circumstances

There's not a universal recipe book

Unfortunately, there's no universal recipe book for when and in what situations you should apply certain data mining methods

Statistics doesn't work like that. Sometimes there's a clear approach; sometimes there is a good amount of uncertainty in what route should be taken. That's what makes it so hard, and so fun

This is true even at the expert level (and there are even larger philosophical disagreements spanning whole classes of problems)

The best you can do is try to understand the problem, understand the proposed methods and what assumptions they are making, and find some way to evaluate their performances

Next time: information retrieval

	but	cool	dude	party	michelangelo	raphael	rude	• • •
1	19	0	0	0	4	24	0	
2	8	1	0	0	7	45	1	
3	7	0	4	3	77	23	0	
4	2	0	0	0	4	11	0	
5	17	0	0	0	9	6	0	
6	36	0	0	0	17	101	0	
7	10	0	0	0	159	2	0	
8	2	0	0	0	0	0	0	
су	1	1	1	1	1	1	1	
	1 2 3 4 5 6 7 8 y	but 1 19 2 8 3 7 4 2 5 17 6 36 7 10 8 2 cy 1	but cool 1 19 0 2 8 1 3 7 0 4 2 0 5 17 0 6 36 0 7 10 0 8 2 0 cy 1 1	but cool dude 1 19 0 0 2 8 1 0 3 7 0 4 4 2 0 0 5 17 0 0 6 36 0 0 7 10 0 0 8 2 0 0 cy 1 1 1	but cool dude party119002810370442005170063600710008200cy111	but cool dude party michelangelo11900042810073704377420004517000963600017710000159820000ry11111	but cool dude party michelangelo raphael1190004242810074537043772342000411517000966360001771017100015928200000ry111111	but cool dude party michelangelo raphael rude119000424028100745137043772304200041105170009606360001592071000000082000000Ty1111111