



Principles of Website Functionality & Advertisement

CS57300 - Data Mining
Spring 2016

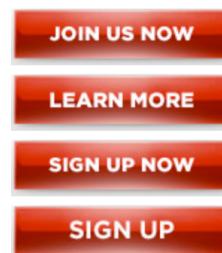
Instructor: Bruno Ribeiro

Goals

- ▶ Introduce a variety of Data Mining applications
- ▶ Explain some of the principles
- ▶ Give a roadmap for the rest of the course

A/B Testing for Political Campaigns

- ▶ Obama's 2008 campaign online effort first of its kind
- ▶ Try different strategies to get more donations



Ack: Dan Siroker
optimizely.com

Effectiveness of Distinct Strategies



Probability difference is significant

no. users donated (conversions)

no. users on each bucket

(2)	Download: XML CSV TSV Print	Chance to Beat Orig.	Observed Improvement	Conv./Visitors
0.2% - [] +	—	—	—	5851 / 77858
0.2% - [] +	100%	18.6%	6927 / 77729	
0.2% - [] +	73.5%	1.37%	5915 / 77644	
0.2% - [] +	13.7%	-2.38%	5660 / 77151	
0.2% - [] +	—	—	4425 / 51794	
0.2% - [] +	100%	13.1%	4996 / 51696	
Change Image	8.87% ± 0.2%	92.2%	3.85%	4595 / 51790
Barack's Video	7.76% ± 0.2%	0.04%	-9.14%	3992 / 51427
Sam's Video	6.29% ± 0.2%	0.00%	-26.4%	3261 / 51864
Springfield Video	5.95% ± 0.2%	0.00%	-30.3%	3084 / 51811

Ack: Dan Siroker
optimizely.com

Political Donations

- ▶ This election cycle the candidates will be doing similar experiments



v.s.



- ▶ Politician statements are similarly tested
 - No surprises on the impact on likely voters

Showing News

- ▶ Yahoo! News
- ▶ Google News

Top Stories

New York rebounds from blizzard, DC stuck in snowy gridlock

Reuters - 3 hours ago

NEW YORK/WASHINGTON Following a day of hunkering down, New Yorkers and Washingtonians surged back into the streets on Sunday after a massive blizzard brought much of the U.S.

East Coast digs out after blizzard, but dangers remain Chicago Tribune

Winter Storm In Eastern US Buries Cities, Floods Coast, Kills At Least 29 NPR

Reuters

See realtime coverage

Wall Street Journal CNN Washington Post

- ▶ What is relevant for a user?
 - How do we find out?

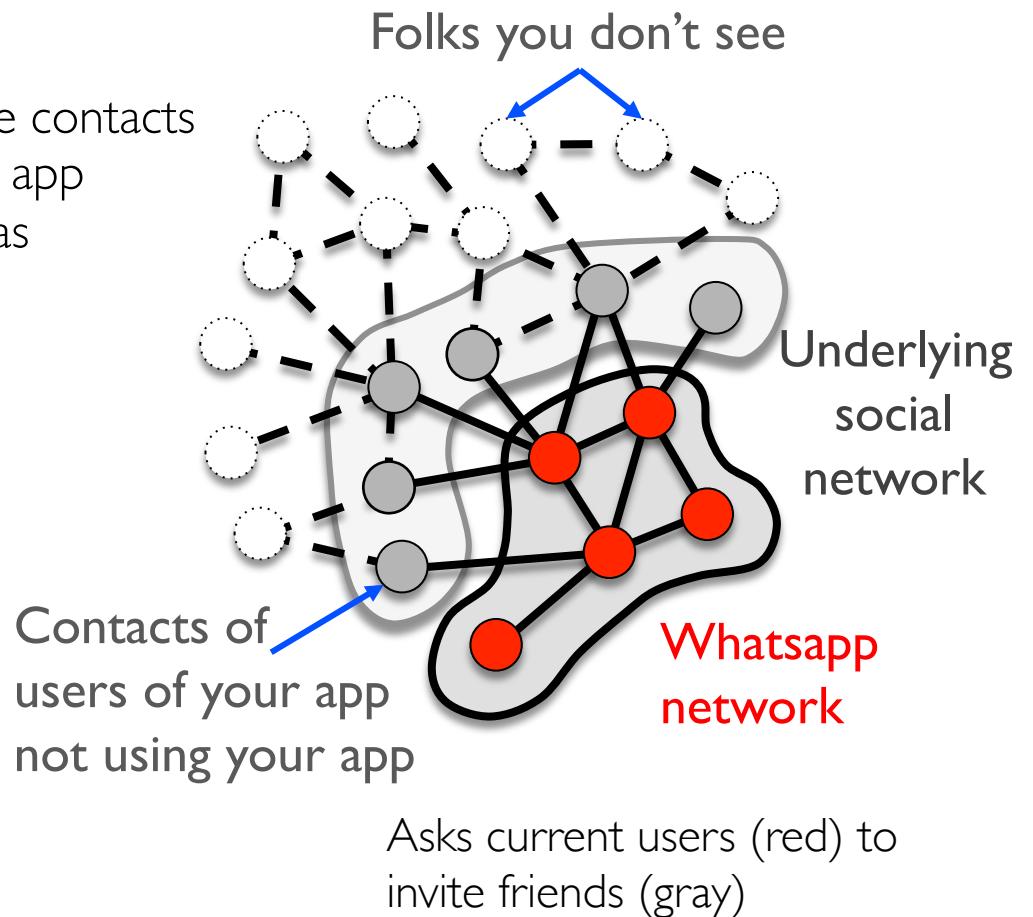
Hypothesis Testing (Sequential Analysis)

- ▶ Hypotheses are tentative statements of the expected relationships between two or more variables
- ▶ Formulate null and alternative hypothesis
 - H_0 : Angry Trump donations = Calm Trump donations
 - H_1 : Angry Trump donations \neq Calm Trump donations
- ▶ Gather a sample statistic (e.g., μ = estimate of Angry Trump donations)
- ▶ Determine the sampling distribution for the statistic under the null hypothesis
- ▶ Use the sampling distribution to calculate the probability of obtaining the observed value of μ , given H_0
 - If the probability is low, reject H_0 in favor of H_1

Problems in Network Analysis

Why Social Networks Grow

- ▶ Facebook / Twitter growth
 - Suggests new users (link prediction)
- ▶ Whatsapp growth:
 - Populate user list with existing phone contacts
 - Suggests user can send messages via app
 - Message arrives to non-subscribers as invitation to subscribe
- ▶ Recruit users via social networks
 - Online & Offline social networks
- ▶ Network effect:
 - Value of service monotonically increases with no. users
 - Positive feedback loop
 - How useful is having an email address if nobody has one?

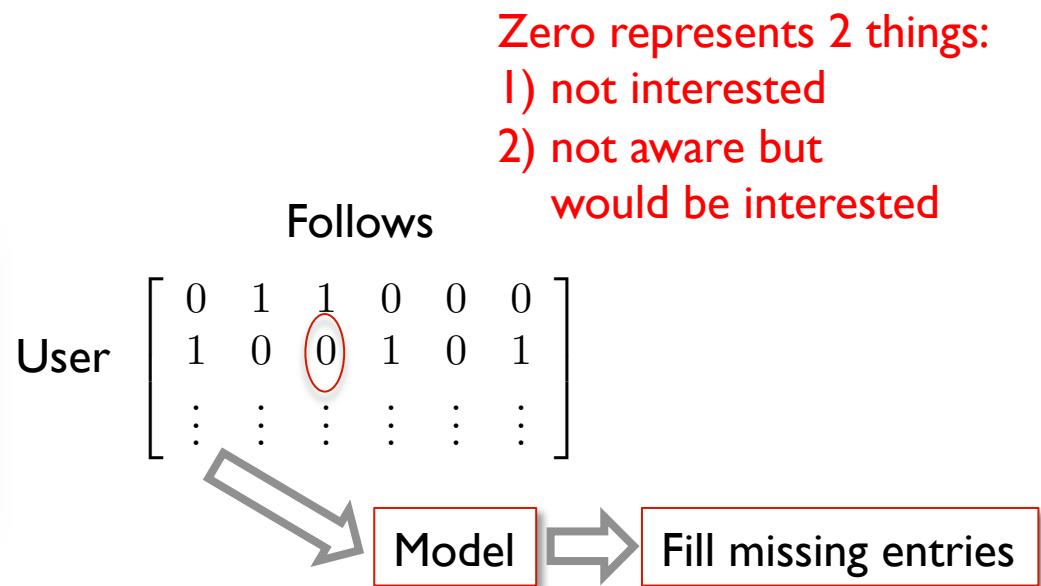


Facebook / Twitter

- ▶ Link prediction

The screenshot shows a list of three Twitter accounts:

- Google Students** (@googlestudents) - Google news and updates especially for students. Tweets from Sarah in Student Development.
- Twitter Engineering** (@TwitterEng) - The official account for Twitter Engineering. Followed by Johan Ugander and Bruno Gonçalves.
- Turing Institute** (@turinginst) - The twitter feed for the Alan Turing Institute. The UK's national institute for data science. Followed by Andrea Baroni, Charles Sutton and Brian Keegan.



Low Rank Reconstruction

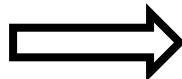
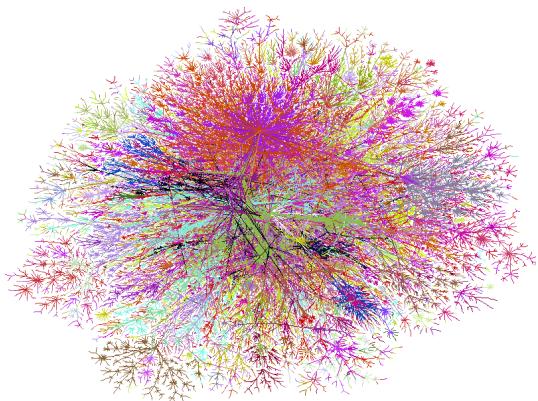
- ▶ Represent the adjacency matrix A with a lower rank matrix A_k .

$$\boxed{A} \approx \boxed{U_{k \times n}} \boxed{V_{n \times k}} = \boxed{A_k}$$

- ▶ If $A_k(u,v)$ has large value for a missing $A(u,v)=0$, then recommend link (u,v)

Graph Applications

▶ Node Importance



Web Graph

purdue airport

All Maps News Images Shopping More Search tools

About 999,000 results (0.74 seconds)

general information - Purdue University
www.purdue.edu/pat/mainnav/airport/air_geninfo.html ▾ Purdue University ▾
The Purdue University Airport encompasses 500 acres divided into airside and landside facilities. The airside includes two runways, a system of parallel ...

airline reservations - Purdue University
www.purdue.edu/pat/.../airport/airline_reserve.html ▾ Purdue University ▾
We do not currently have commercial airline service at the Purdue University Airport (LAF). The closest airport with airline service is in Indianapolis, 65 miles ...

Transportation - Visitor Information - Purdue University
<https://www.purdue.edu/visit/.../transportation.html> ▾ Purdue University ▾
Purdue is accessible from the Indianapolis and Chicago O'hare international airports. The Purdue University Airport is less than one mile from campus and is ...
You've visited this page 3 times. Last visit: 12/29/15

Parking, Airport and - Purdue University
www.purdue.edu/pat/ ▾ Purdue University ▾
FAQs, statistics, history, map, flight schedules, weather, and information on the fixed base operators and ground transportation.

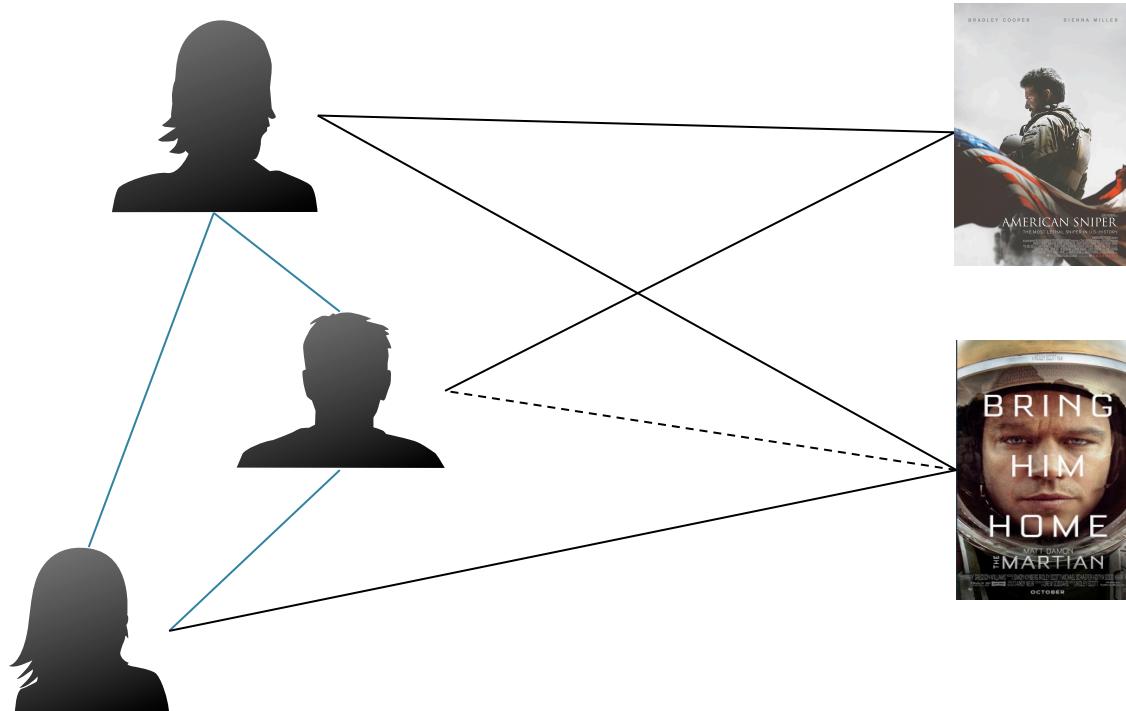
Purdue University Airport - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Purdue_University_Airport ▾ Wikipedia ▾
History[edit]. Purdue University Airport was the first university-owned airport in the United States. In 1930, inventor-industrialist David Ross (one of two people for ...

▶ Spammer Detection

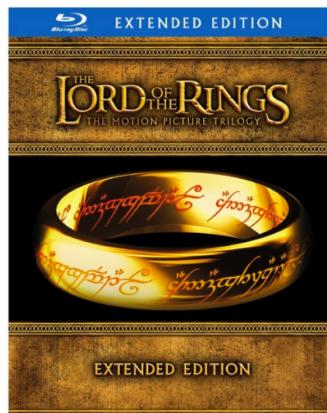
- Twitter spammers tend to form strongly connected components

Collaborative filtering (Graph-based Approaches)

- ▶ (Social) Collaborative Filtering
 - - Based on homophily: items or users
 - - Rely on user serendipity
 - + Bias recommendations towards truly relevant items



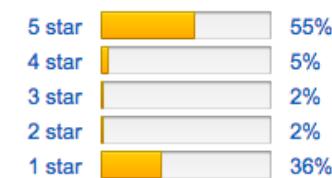
Collaborative filtering (example)



Customer Reviews

★★★★★ 9,154

3.4 out of 5 stars ▾



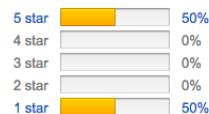
[See all 9,154 customer reviews ▾](#)

- When searching for action movies, should we show LORT Blue Ray edition?

Amazon (Fraud)

▶ Detecting Fraud

Review Fraud?



[See both customer reviews >](#)

Most Helpful Customer Reviews

★★★★★ Realistic Looking Security Camera
By Kim S on September 5, 2015

Verified Purchase

This is the second on of these from two different sellers replaced. It is not convenient to keep replacing batteries one that had a light, but this is no good, just like the old ones flashes and I only have to replace the battery once a length of time.

[Comment](#) | Was this review helpful to you?

0 of 1 people found the following review helpful

★★★★★ It is a very good product. I'm glad to buy it
By coffee on April 1, 2015

It is a very good product. I'm glad to be able to buy a

[Comment](#) | Was this review helpful to you?

coffee reviewed a product.
Nov 2, 2015

Women's Fashion Trendy Button Down Faux Suede High ...
★★★★★ It is a very good product. I'm glad to be able to buy ...
It is a very good product. I'm glad to be able to buy a practical product. I want to recommend this to others.

coffee reviewed a product.
Sep 22, 2015

Arshiner Laugh & Learn Say Please Magical Tea Set Magi...
★★★★★ It is a very good product. I'm glad to be able to buy ...
It is a very good product. I'm glad to be able to buy a practical product. I want to recommend this to others.

[Checkout → Pay](#) [My Account](#)

Buy Amazon Reviews

coffee reviewed a product.
Apr 7, 2015

Rockport Men's Lead The Pack Wingtip Oxford,Black Water...
★★★★★ It is a very good product. I'm glad to be able to buy ...
It is a very good product. I'm glad to be able to buy a practical product. I want to recommend this to others.

coffee reviewed a product.
Sep 22, 2015

MAYTAMA 100% Very Rare Maragogipe Honey Processed...
★★★★★ It is a very good product. I'm glad to be able to buy ...
It is a very good product. I'm glad to be able to buy a practical product. I want to recommend this to others.

[Home](#) [Buy Reviews](#) [Contact Us](#) [FAQ](#)

Buy Amazon Reviews



Never has it been easier to get multiple 4 and 5 star reviews on your Amazon product page. We provide real reviews from aged accounts with real buying activity. Most products in the Amazon marketplace will never even be seen. The more positive reviews you have the better your chances are.



Amazon is a highly competitive marketplace. Standing out from the crowd is the only way you are going to sell products. When you buy Amazon reviews you can rest assured that your product listing will rank higher in searches, get more attention and most importantly make more sales.

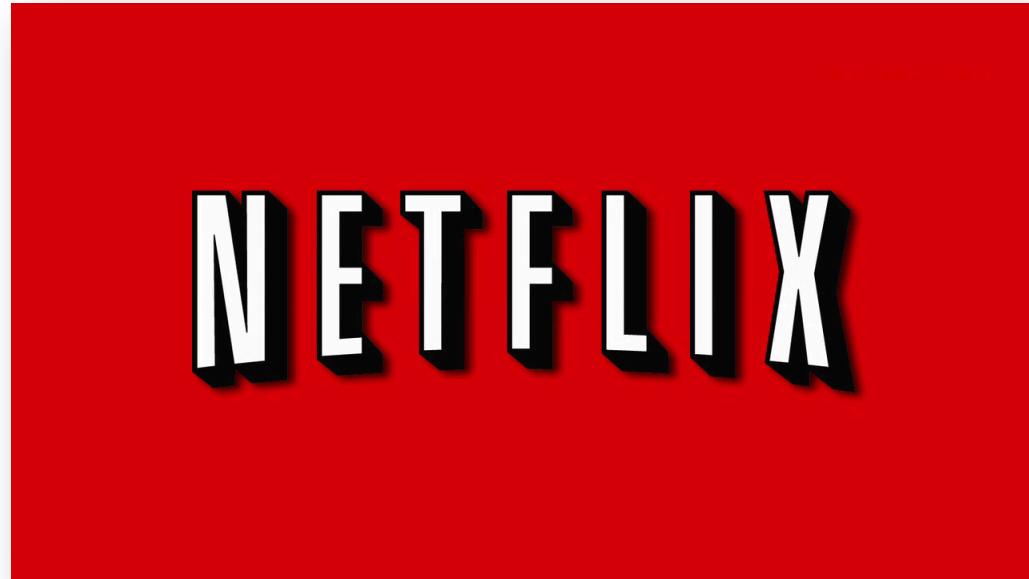


We Review Any Product

We utilize the talents of a professional and diverse writing team. You can buy Amazon reviews for any type of product. We write reviews for music, eBooks, supplements, cosmetics etc.. We won't just copy reviews from elsewhere and rewrite them. Your reviews will be 100% unique.

Netflix

- ▶ Joining multiple sources of information:
Co-factorization



Browse Movies

Don't Forget ... Spend a few seconds here to personalize your Netflix experience.

Rate your Recent Returns

Clash of the Titans (No Opinion) The X-Files: Fight the Future (No Opinion)

Click the stars to tell us how much you enjoyed these movies.

Rate More Movies

Movies for You

You Have Recommendations!

Friends

YOUR FRIENDS' AND FAVES' ACTIVITY

Date	Action	Item
03/29	Queued	The Parent Trap
03/27	Returned	24: Season 1, Disc 1 (6-Disc Series)
03/27	Shipped	Australia
03/25	Returned	Chasing
03/25	Shipped	Rachel Getting Married
03/24	Shipped	The Love Guru
03/24	Returned	24: Season 1, Disc 2 (6-Disc Series)
03/24	Shipped	24: Season 1, Disc 3 (6-Disc Series)

Profile

Ratings

Reviews & Lists

Friends & Faves

Notebook

Community Home

User

B

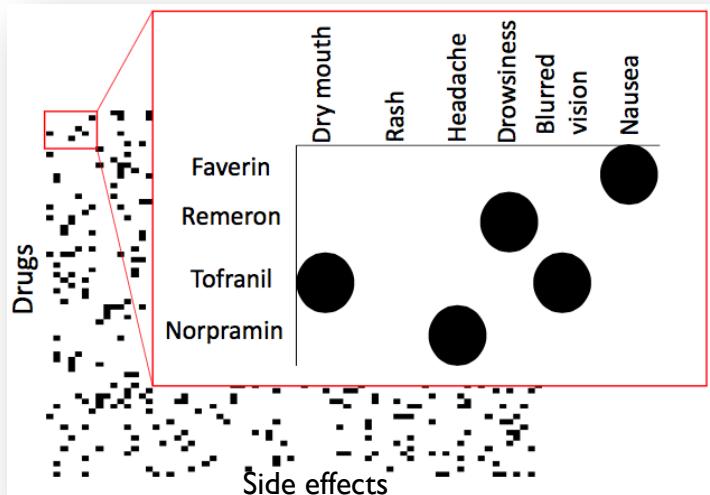
User

A

Other forms of data

Retail & Healthcare

- Classification (drug safe or not safe, user buys or does not buy)



Bryan Hooi, Hyun Ah Song, Evangelos Papalexakis,
Rakesh Agrawal, Christos Faloutsos, PAKDD'16

DrugBank dataset: <http://www.drugbank.ca/downloads>

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

Ack: Neville

- Descriptive vs Predictive
 - Tylenol and Advil are good for pain (descriptive)
 - Drug will reduce fever (predictive)

Modeling approaches

Descriptive vs. predictive modeling

- ▶ Descriptive models **summarize** the data
 - Provide insights into the domain
 - Focus on modeling joint distribution $P(X)$
 - May be used for classification, but prediction is not the primary goal
- ▶ Predictive models **predict** the value of one variable of interest given known values of other variables
 - Focus on modeling the conditional distribution $P(Y | X)$ or on modeling the decision boundary for Y

Example: SPAM

- ▶ I was reading a little more about Tsalling entropy and trying to figure out whether it would be appropriate for relational learning problems. One possibility is to use it for exponential random graph models, which have features like the number of triangles in the graph. Since these grow with graph size, it seems to be an "extensive" property that the Tsalling entropy is trying to model...
- ▶ Don't Be Silly To Pay Hundred\$ Or Thousand\$ You Can Have The Exactly Same Licensed Software At 5%-10% Of The Retail Price : All popular softwares for PC & MAC : Language available: English, Deutsch, French, Italian, Spanish : Buy & start downloading right after you paid : You will be given your dedicated PERSONAL LICENSE right after you paid....

Data representation

- ▶ Class label: isSpam {+, -}
- ▶ Attributes?
 - Convert email text into a set of attributes

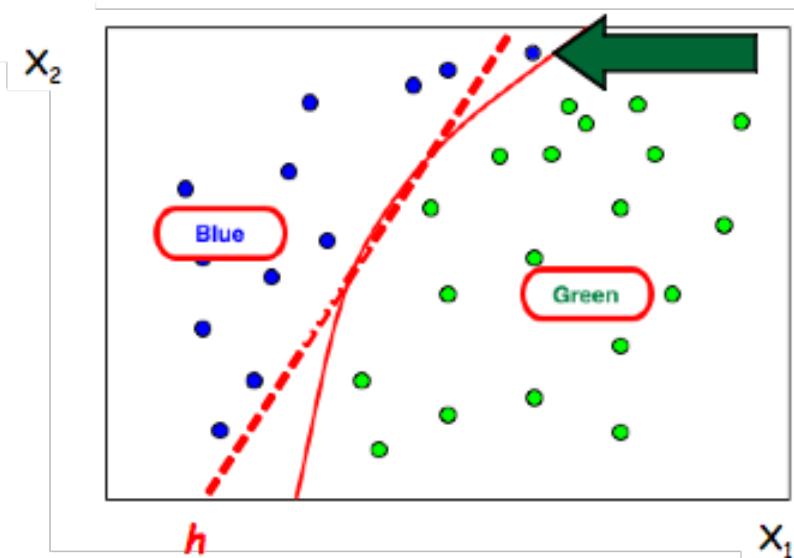
isSpam	word ₁	word ₂	word ₃	...	word _n
+	1	0	1	...	1
-	0	0	0	...	1

Predictive modeling

- ▶ Data representation:
 - Training set: Paired attribute vectors and class labels $\langle y(i), x(i) \rangle$ or $n \times p$ tabular data with class label (y) and $p-1$ attributes (x)
- ▶ Task: estimate a predictive function $f(x; \theta) = y$
 - Assume that there is a function $y=f(x)$ that **maps** data instances (x) to class labels (y)
 - Construct a model that approximates the mapping
 - Classification: if y is categorical
 - Regression: if y is real-valued

Classification

- ▶ In its simplest form, a classification model defines a decision boundary (h) and labels for each side of the boundary
- ▶ Input: $\mathbf{x}=\{x_1, x_2, \dots, x_n\}$ is a set of attributes, function f assigns a label y to input \mathbf{x} , where y is a discrete variable with a finite number of values



Classification output

- ▶ Different classification tasks can require different kinds of output
 - Each requires progressively more accurate models (e.g., a poor probability estimator can still produce an accurate ranking)
- ▶ Class labels — Each instance is assigned a single label
 - Model only need to decide on *crisp class boundaries*
- ▶ Ranking — Instances are ranked according to their likelihood of belonging to a particular class
 - Model implicitly explores many potential class boundaries
- ▶ Probabilities — Instances are assigned class probabilities $p(y|x)$
 - Allows for more refined reasoning about sets of instances

Discriminative classification

- ▶ Model the decision boundary directly
- ▶ Direct mapping from inputs \mathbf{x} to class label y
- ▶ No attempt to model probability distributions
- ▶ May seek a discriminant function $f(\mathbf{x}; \boldsymbol{\theta})$ that maximizes measure of separation between classes
- ▶ Examples:
 - Perceptrons, nearest neighbor classifiers, support vector machines, decision trees

Probabilistic classification

- ▶ Model the underlying probability distributions
 - Posterior class probabilities: $p(y|x)$
 - Class-conditional and class prior: $p(x|y)$ and $p(y)$
- ▶ Maps from inputs x to class label y indirectly through posterior class distribution $p(y|x)$
- ▶ Examples:
 - Naive Bayes classifier, logistic regression, linear regression, probability estimation trees