Data Mining

CS57300
Purdue University

January 17, 2018

Goals

- Introduce a variety of Data Mining applications
- Explain some of the principles behind today's Web

The Anatomy of the Today's Web

User interaction

These days, even walking around is a form of interaction

Data

Sells

Mining

insurance, advertisers, etc.

Predictions/ Descriptions

Descriptive vs. predictive modeling

- Descriptive (generative) models **summarize** the data
  - Provide insights into the domain
  - Focus on modeling joint distribution *P(X) or P(X,Y)*
  - 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

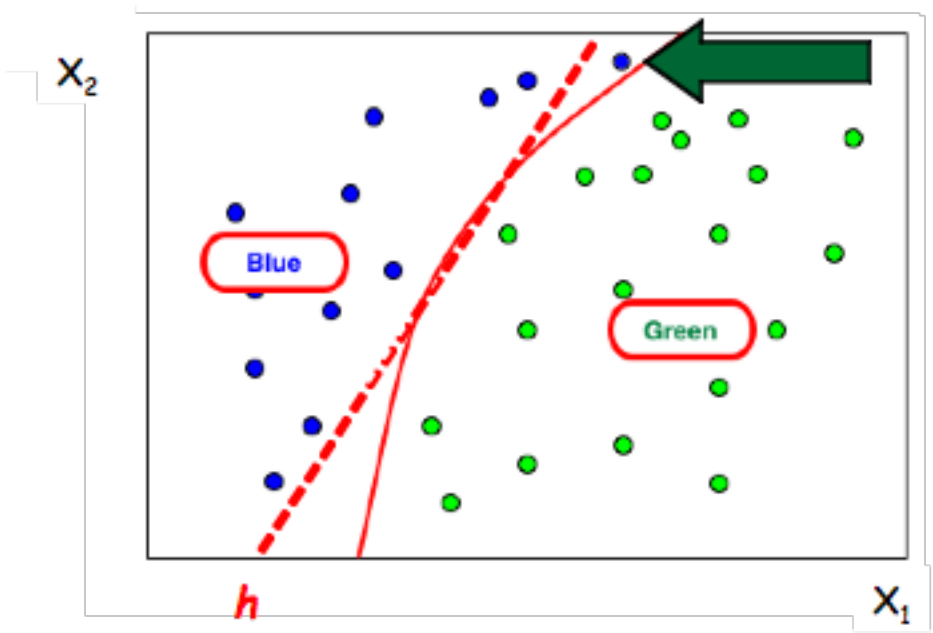
isSpam	$word_1$	$word_2$	$word_3$	...	$word_n$
+	1	0	1	...	1
-	0	0	0	...	1

Predictive modeling

- Data representation:
  - Training set: Paired attribute vectors and class labels $(\mathbf{x}_i, y_i)$ or n×p tabular data with class label *y* and p-1 attributes ***x***
- Task: estimate a predictive function $f(\boldsymbol{x}; \boldsymbol{W}) = y$
  - Assume that there is a parametric function $y = f(x; \boldsymbol{W})$ that **maps** data instances **X** to class labels Y using function parameters ***W***
  - 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: $\mathbf{x}=\{x_1,x_2,...,x_n\}$ is a set of attributes, function f assigns a label y to input **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 $p(y|\boldsymbol{x})$
  - *Allows for more refined reasoning about sets of instances*

Discriminative classification

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

Probabilistic classification

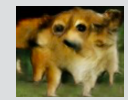
- Model the underlying probability distributions
  - Posterior class probabilities: $p(y|\boldsymbol{x})$
  - Class-conditional and class prior: $p(\boldsymbol{x}|y)$ and $p(y)$
- Maps from inputs **x** to class label y indirectly through posterior class distribution $p(y|\boldsymbol{x})$
- Examples:
  - Naive Bayes classifier, logistic regression, linear regression, most neural networks classification/regression tasks

Examples: Predictive/Descriptive(Generative) Models

- **Classification/Regression task**
  - Given an example (x_i, y_i)
  - Wants to learn relationship between X and Y (often a probability distribution p(Y | X)
  - To use it, we need x_i and the output is the predicted class: $\hat{y}_i = \arg\max_y p(y|x_i)$

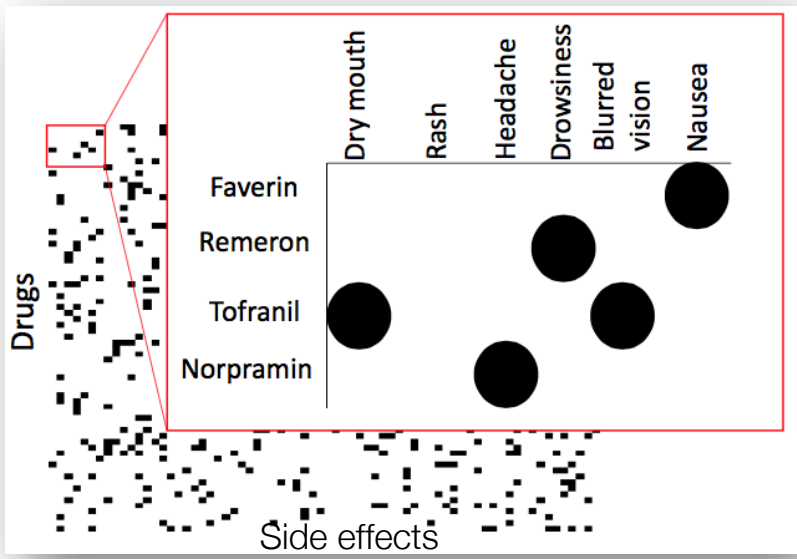
e.g.: $x_i =$  , $y_i = \text{dog}$

- **Generation (descriptive) task ("closer to real intelligence")**
  - Given an example (x_i, y_i)
  - Wants to learn joint probability $p(y_i, x_i)$
  - To use it, we sample another example from $p(y, x)$, the output is an entirely new example (it understands "behavior" and can simulate it)

e.g.: $x =$  , $y = \text{dog}$

Example Classification Task: 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

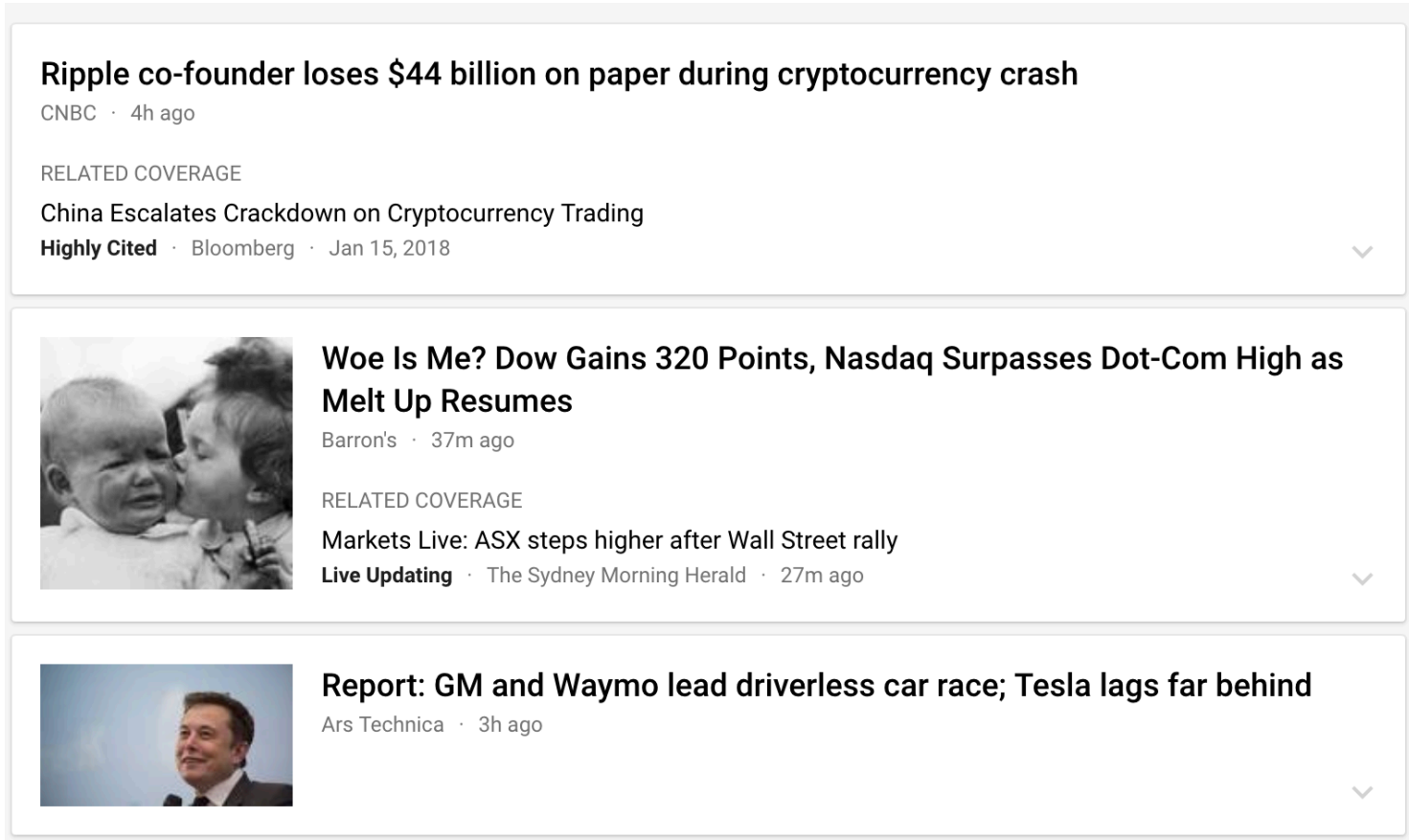
DrugBank dataset: http://www.drugbank.ca/downloads

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

- **Descriptive vs Predictive**
  - Tylenol and Advil are good for pain (descriptive)
  - Drug X will reduce fever Y by at least 5% (predictive)

Click-through Prediction Task:
Google News predicts probability you will click to read the news

- Google News
  - Ranked list of news from highest probability of clicking to lowest
  - “Filter bubble”



Example Regression Task:

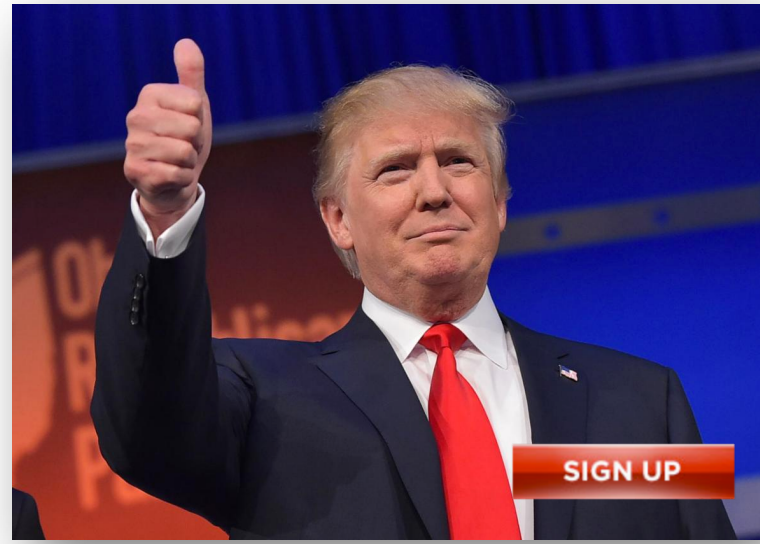
- Obama's 2008 campaign online effort first of its kind
- Try different strategies to get more donations
- Given website layout, predict wording that gives largest average campaign contribution

Ack: Dan Siroker
optimizely.com

Regression Task (cont)

- In 2016 candidates did similar experiments

v.s.



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

All Predictive Models Must be Tested

- **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 ≠ 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

Two Types of Hypothesis (Model) Testing

- **Offline Testing**
  - Test if predictions are accurate over held out data (data not used to train the prediction model)
  - Why can't we use the original (training) data, from which we constructed our prediction model?
    - Most prediction methods are not robust to overfitting
  - Examples of techniques:
    - Boostrapping
    - Cross-validation

- **Online Testing**
  - We "test" predictions in the live system
  - It is not quite a test, because we don't really care which hypothesis (model) is more accurate
  - Examples:
    - Reinforcement learning
    - Multi-armed Bandits

The New York Times Daily Dilemma

- Select 500 users to see headline chosen by model A
  - Titanic Sinks
- Select 500 users to see headline chosen by model B
  - Ship Sinks Killing Thousands


"All the News That's Fit to Print."

The New York Times.

THE WEATHER.

NEW YORK, TUESDAY, APRIL 16, 1912.—TWENTY-FOUR PAGES. ONE CENT

TITANIC SINKS FOUR HOURS AFTER HITTING ICEBERG; 866 RESCUED BY CARPATHIA, PROBABLY 1250 PERISH; ISMAY SAFE, MRS. ASTOR MAYBE, NOTED NAMES MISSING

Col. Astor and Bride, Isidor Straus and Wife, and Maj. Butt Aboard.

"RULE OF SEA" FOLLOWED

Women and Children Put Over in Lifeboats and Are Supposed to be Safe on Carpathia.

PICKED UP AFTER 8 HOURS

Vincent Astor Calls at White Star Office for News of His Father and Leaves Weeping.

FRANKLIN HOPEFUL ALL DAY

Manager of the Line Insisted Titanic Was Unsinkable Even After She Had Gone Down.

HEAD OF THE LINE ABOARD

J. Bruce Ismay Making First Trip on Gigantic Ship That Was to Surpass All Others.

Biggest Liner Plunges to the Bottom at 2:20 A. M.

RESCUERS THERE TOO LATE

Except to Pick Up the Few Hundreds Who Took to the Lifeboats.

WOMEN AND CHILDREN FIRST

Cunarder Carpathia Rushing to New York with the Survivors.

SEA SEARCH FOR OTHERS

The California Stands By on Chance of Picking Up Other Boats or Rafts.

OLYMPIC SENDS THE NEWS

Only Ship to Flash Wireless Messages to Shore After the Disaster.

- We often refer to decision of choosing A or B as choosing an **action** (or arm)
- Do people click more on headline of models A or B?
  - If action A much better than action B, we are wasting users 500 users on a bad model… can we do better?

Truth is…

- Sometimes we don't only want to quickly find whether hypothesis (model) A is better than hypothesis (model) B
- We really want to use the best-looking hypothesis (model) at any point in time
- Deciding if H_0 should be rejected is irrelevant

Real-world Problem

- Websites in perpetual state of testing

- Goal:
  Acquire just enough information about suboptimal action (headline) to ensure they are suboptimal. Looking for action (headline) i of user k that maximizes $\mathrm{E}[\mathrm{X}^{(\mathrm{i})}{}_{\mathrm{k}}]$

$$X_k^{(i)} = \begin{cases} 1 & \text{, if } k\text{-th user seeing headline } i \text{ clicks} \\ 0 & \text{, otherwise} \end{cases}$$

(A) Titanic Sinks

reward

$$X_k^{(1)} = \begin{cases} 1 & \text{, with probability } p_1 \\ 0 & \text{, otherwise} \end{cases}$$

(B) Ship Sinks Killing Thousands

$$X_k^{(2)} = \begin{cases} 1 & \text{, with probability } p_2 \\ 0 & \text{, otherwise} \end{cases}$$

Role of Dependencies in Data

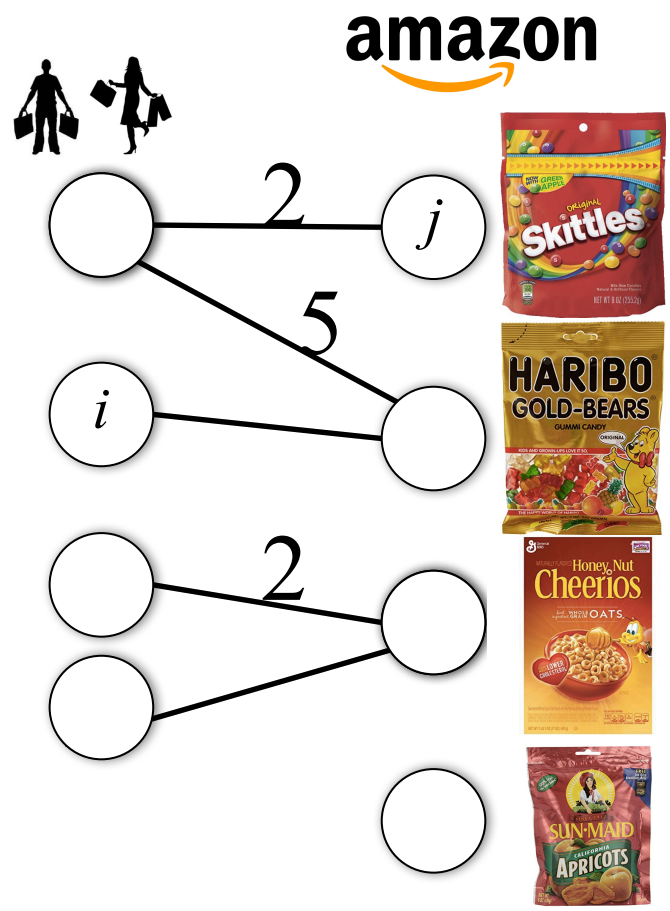
Relationship in Data

- Data is rich and sometimes dependent
  - Examples (x_i, y_i) and (x_j, y_j) may have dependencies
    - Say, i and j are friends on Facebook. X is a set of observable online behaviors and Y is whether they vote for the same party

  - Many very important prediction tasks using dependent data are related to graphs

Product Recommendation

Example of predicting links in a bipartite graph



Twitter's Who to Follow

Another example of link prediction application

Graph-based Prediction Tasks also Find Fraud

- Detecting Fraud

Review Fraud?

5 star 50%
4 star 0%
3 star 0%
2 star 0%
1 star 50%

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 sell
replaced. It is not convenient to keep replacing batte
one that had a light, but this is no good, just like the
flashes and I only have to replace the battery once a
length of time.
Comment | Was this review helpful to you? Yes

0 of 1 people found the following review helpful
It is a very good product. I'm glad t
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? Yes