Data Mining - SENG 474/CSC578D

Hung Le

University of Victoria

January 8, 2019

Teaching Staffs

- Instructor: Hung Le
  - Email: hungle@uvic.ca
  - Course Website: `https://hunglvosu.github.io/DMW19.html`
  - Office: ECS 621
  - Office Hours: 10:30 am - 12:30 pm Friday
- TAs:
  - Sajjad Azami (Email: sajjadaazami@gmail.com
  - Cole Peterson (Email: colpeterson@gmail.com)
  - Jasbir Singh (Email: jasbircheema96@gmail.com)
  - Office Hours:
    - ★ Monday 11:00-12:30 am
    - ★ Tues: 1:30-3:00 pm

Textbook

- **Mining of Massive Datasets**
  - By Jure Leskovec, Anand Rajaraman, Jeff Ullman.
  - Why? it's free (http://www.mmds.org). Educational cost is already expensive!!!
  - Most of the materials presented in this course will be drawn from there.

Logistics

	SENG474	CSC578D	Misc
4 HWs	20%	20%	5% for each, group of 2
Project	20%	25%	Exp. for 578D is higher, group of 3
Midterm	20%	15%	Wed, 13 of Feb, 2019
Final	40%	40%	Scheduled by the university

There maybe programming questions in HW. Other useful information (late homework policy, grading system, plagiarism policy):

- https://heat.csc.uvic.ca/coview/outline/2019/Spring/SENG/474
- https://heat.csc.uvic.ca/coview/outline/2019/Spring/CSC/578D

My advice: start the project and HWs ASAP.

HW discussion policy

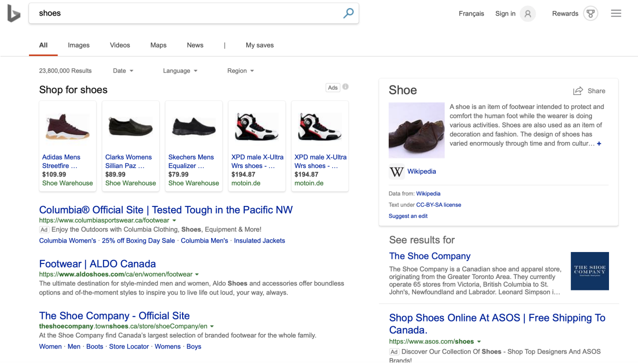
You can discuss your HW with at most one other group. However, discussions are restricted to oral only. Written similarity is regarded as plagiarism. If you group discusses with other groups, you must acknowledge the discussion in your written solution.

General expectation

- Learning useful techniques to "mine" data.
- Dealing with massive data sets.

My background

- WPR (Whole Page Relevance) - Bing



23,800,000 Results

Shop for shoes

Adidas Mens Streetfire …	Clarks Womens Sillian Paz …	Skechers Mens Equalizer …	XPD male X-Ultra Wrs shoes - …	XPD male X-Ultra Wrs shoes - …
$109.99	$89.99	$79.99	$194.87	$194.87
Shoe Warehouse	Shoe Warehouse	Shoe Warehouse	motoin.de	motoin.de

Columbia® Official Site | Tested Tough in the Pacific NW
https://www.columbiasportswear.ca/footwear
Ad Enjoy the Outdoors with Columbia Clothing, **Shoes**, Equipment & More!
Columbia Women's · 25% off Boxing Day Sale · Columbia Men's · Insulated Jackets

Footwear | ALDO Canada
https://www.aldoshoes.com/ca/en/women/footwear
The ultimate destination for style-minded men and women, Aldo **Shoes** and accessories offer boundless options and of-the-moment styles to inspire you to live life out loud, your way, always.

The Shoe Company - Official Site
theshoecompany.townshoes.ca/store/shoeCompany/en
At the Shoe Company find Canada's largest selection of branded footwear for the whole family.
Women · Men · Boots · Store Locator · Womens · Boys

Shoe

A shoe is an item of footwear intended to protect and comfort the human foot while the wearer is doing various activities. Shoes are also used as an item of decoration and fashion. The design of shoes has varied enormously through time and from cultur… +

Wikipedia

Data from: Wikipedia
Text under CC-BY-SA license
Suggest an edit

See results for

The Shoe Company
The Shoe Company is a Canadian shoe and apparel store, originating from the Greater Toronto Area. They currently operate 65 stores from Victoria, British Columbia to St. John's, Newfoundland and Labrador. Leonard Simpson i…

Shop Shoes Online At ASOS | Free Shipping To Canada.
https://www.asos.com/shoes
Ad Discover Our Collection Of **Shoes** - Shop Top Designers And ASOS Brands!

Who need to mine data

Retail [1]

- Walmart: handles more than 1 million customer transactions every hour, has more than more than 2.5 PB (2560 TB) of data.
- Windermere Real Estate: use location information from nearly 100 million drivers to help new home buyers determine their typical drive times to and from work throughout various times of the day.
- FICO Card Detection System protects accounts worldwide

[1] https://en.wikipedia.org/wiki/Big_data#Case_studies

Who need to mine data

Science [2]

- Large Hadron Collider experiments: 150 million sensors delivering data 40 million times per second and nearly 600 million collisions per second.
- The Square Kilometre Array is a radio telescope built of thousands of antennas. These antennas are expected to gather 14 EB and store 1 PB per day.
- The NASA Center for Climate Simulation (NCCS) stores 32 PB of climate observations.
- So many more in the footnote link.

[2] https://en.wikipedia.org/wiki/Big_data#Case_studies

Who need to mine data

Technology [3]

- Bay.com: two data warehouses at 7.5 PB and 40 PB as well as a 40PB Hadoop cluster.
- Amazon.com: the world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB (as of 2005).
- Facebook: 50 billion photos from its user base. As of June 2017, Facebook reached 2 billion monthly active users.
- Google 3.5 billion searches per day[4].

[3] https://en.wikipedia.org/wiki/Big_data#Case_studies

[4] http://www.internetlivestats.com/google-search-statistics/

What is data mining?

I won't define this term (and it probably isn't very important). My own perspective on data mining is that you are given a (big) dataset and your problem is to image what can you (ethically) do with your data to drive your business.

What is data mining?

I won't define this term (and it probably isn't very important). My own perspective on data mining is that you are given a (big) dataset and your problem is to image what can you (ethically) do with your data to drive your business. In this class, we will learn several principles from techniques that we use to "mine" our data. These technique will be sampled from:

- Statistics.
- Machine Learning.
- Computational approach.

Statistics

- Data mining as the construction of a *statistical model*.

Example

Find a model for:

$$x = [-0.13, -0.12, 0.95, 0.12, -0.61, -0.47, -0.21, 0.24, -0.50, 0.11]$$

- Find sample mean: $\mu = \frac{\sum_{i=1}^{10} x_i}{10} = -0.062$.
- Find sample variance: $\sigma^2 = \frac{1}{10} \sum_{i=1}^{10} (x[i] - \mu)^2 = 0.1866$.

Machine Learning

- Useful when you DON'T have an idea of what you are looking for in the data.
  - Your data is too complicated to discover patterns.
  - Train a ML model and let them predict the outcome.

Machine Learning

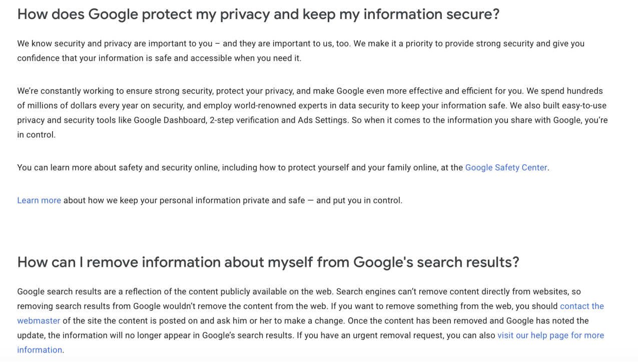
- Useful when you DON'T have an idea of what you are looking for in the data.
  - Your data is too complicated to discover patterns.
  - Train a ML model and let them predict the outcome.
- Typically NOT useful when you can describe your goal concretely.

Example:

- WhizBang! Labs: use ML to locate people resumes on the Web. They can't compete with a simple algorithm designed by hand that looks for a particular words or phrases.

Machine Learning

QnA Maker (personal experience).



How does Google protect my privacy and keep my information secure?

We know security and privacy are important to you – and they are important to us, too. We make it a priority to provide strong security and give you confidence that your information is safe and accessible when you need it.

We're constantly working to ensure strong security, protect your privacy, and make Google even more effective and efficient for you. We spend hundreds of millions of dollars every year on security, and employ world-renowned experts in data security to keep your information safe. We also built easy-to-use privacy and security tools like Google Dashboard, 2-step verification and Ads Settings. So when it comes to the information you share with Google, you're in control.

You can learn more about safety and security online, including how to protect yourself and your family online, at the Google Safety Center.

Learn more about how we keep your personal information private and safe — and put you in control.

How can I remove information about myself from Google's search results?

Google search results are a reflection of the content publicly available on the web. Search engines can't remove content directly from websites, so removing search results from Google wouldn't remove the content from the web. If you want to remove something from the web, you should contact the webmaster of the site the content is posted on and ask him or her to make a change. Once the content has been removed and Google has noted the update, the information will no longer appear in Google's search results. If you have an urgent removal request, you can also visit our help page for more information.

Computational approach

- Summarization: summarize the data succintly and approximately.

Example: PageRank

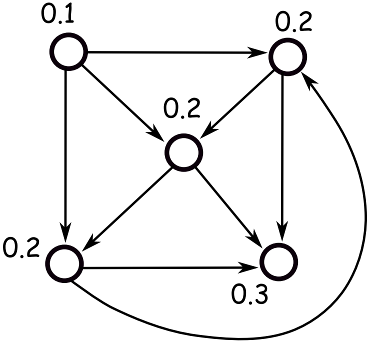
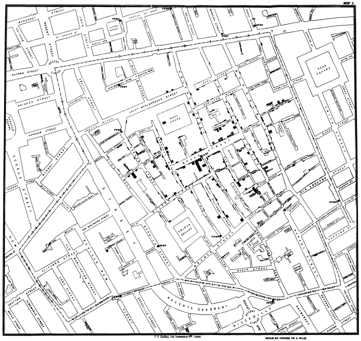


Figure: A Google FAQ page.

Computational approach

- Summarization: summarize the data succintly and approximately.

Example: Clustering. Cholera outbreak in 1854, London.



The physician John Snow plotted Cholera case on the street map. He observed that "nearly all the deaths had taken place within a short distance of the [Broad Street] pump" and deduced that contaminated water is linked to the outbreak.

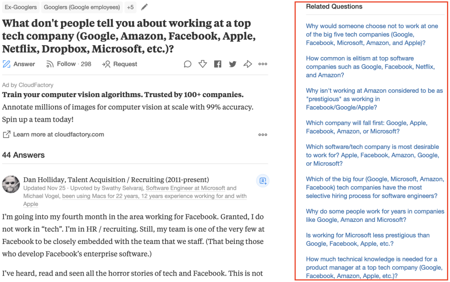
Computational approach

- Summarization: summarize the data succintly and approximately.
- Feature extraction: look for the most extreme examples in the data.
  - Frequent Itemsets. You are given many "baskets" of items and you want to find a group of items that appear together in many baskets.

Baskets	Items
1	{Bread,Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread,Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

Computational approach

- Summarization: summarize the data succintly and approximately.
- Feature extraction: look for the most extreme examples in the data.
  - Frequent Itemsets.
  - Similar items. You are given a collection of sets and you want to find pairs of sets that are similar to each other.



What don't people tell you about working at a top tech company (Google, Amazon, Facebook, Apple, Netflix, Dropbox, Microsoft, etc.)?

Ad by CloudFactory

Train your computer vision algorithms. Trusted by 100+ companies.

Annotate millions of images for computer vision at scale with 99% accuracy. Spin up a team today!

44 Answers

Dan Holliday, Talent Acquisition / Recruiting (2011-present)
Updated Nov 25 · Upvoted by Swathy Selvaraj, Software Engineer at Microsoft and Michael Vogel, been using Macs for 22 years, 12 years experience working for and with Apple

I'm going into my fourth month in the area working for Facebook. Granted, I do not work in "tech". I'm in HR / recruiting. Still, my team is one of the very few at Facebook to be closely embedded with the team that we staff. (That being those who develop Facebook's enterprise software.)

I've heard, read and seen all the horror stories of tech and Facebook. This is not

Related Questions

Why would someone choose not to work at one of the big five tech companies (Google, Facebook, Microsoft, Amazon, and Apple)?

How common is elitism at top software companies such as Google, Facebook, Netflix, and Amazon?

Why isn't working at Amazon considered to be as "prestigious" as working in Facebook/Google/Apple?

Which company will fall first: Google, Apple, Facebook, Amazon, or Microsoft?

Which software/tech company is most desirable to work for? Apple, Facebook, Amazon, Google, or Microsoft?

Which of the big four (Google, Microsoft, Amazon, Facebook) tech companies have the most selective hiring process for software engineers?

Why do some people work for years in companies like Google, Amazon and Microsoft?

Is working for Microsoft less prestigious than Google, Facebook, Apple, etc.?

How much technical knowledge is needed for a product manager at a top tech company (Google, Facebook, Amazon, Apple, etc.)?

Figure: A screenshot from Quora

Statistical Limits on Data Mining

In many occasions, you may want to find rare events in your data. You need to be aware of randomness in your data.

Statistical Limits on Data Mining

Bonferroni's Principle:

1. Calculate the expectation of the rare event, say $\mathbb{E}[\text{rare event}]$, given that the data is random.
2. If the number of rare events you hope to find is much less than $\mathbb{E}[\text{rare event}]$, then whatever you found in the data is likely bogus.

The significance is that you need to redefine your rare event so that it is unlikely to occur in random data.

Statistical Limits on Data Mining

Example

Suppose you have a data of 1B people going to 10^5 hotels in 1000 days, and you hope to find "evil doers" in your data. To find "evil doers", you will find pairs of people who went to the same hotel in two different days.

Two facts from your data:

- Everyone gets to a hotel in 100 days.
- Each hotel can accommodate 100 people in the same day.

Statistical Limits on Data Mining

Example

Suppose you have a data of 1B people going to 10^5 hotels in 1000 days, and you hope to find "evil doers" in your data. To find "evil doers", you will find pairs of people who went to the same hotel in two different days.

Two facts from your data:

- Everyone gets to a hotel in 100 days.
- Each hotel can accommodate 100 people in the same day.

So if people behave completely random, each person, with probability 0.01 will visit a particular hotel in each day.

- There will be about 250000 pairs that look like "evil doers" (see the board calculation).

So if you hope to find 10 pair of evil doers in your data, you won't able to find them with this hypothesis.

Statistical Limits on Data Mining

What can you do? Change your hypothesis.

Example

Suppose you have a data of 1B people going to 10^5 hotels in 1000 days, and you hope to find "evil doers" in your data. To find "evil doers", you will find pairs of people who went to the same hotel in three different days.

Useful things to know

$e = \lim_{x \to \infty}(1 + \frac{1}{x})^x$ and $1/e = \lim_{x \to \infty}(1 - \frac{1}{x})^x$

Useful things to know

$$e = \lim_{x\to\infty}(1+\frac{1}{x})^x \text{ and } 1/e = \lim_{x\to\infty}(1-\frac{1}{x})^x$$

Why do we care?

$$(1-a)^b \sim ((1-a)^{1/a})^{ab} \sim e^{-ab} \tag{1}$$

when $a \ll 1$.

Useful things to know

$$e = \lim_{x\to\infty}(1+\tfrac{1}{x})^x \text{ and } 1/e = \lim_{x\to\infty}(1-\tfrac{1}{x})^x$$

Why do we care?

$$(1-a)^b \sim ((1-a)^{1/a})^{ab} \sim e^{-ab} \qquad (1)$$

when $a \ll 1$.

Example (Birthday paradox)

Suppose that there are 23 random people in the same room. The probability that at least two of them have the same birthday is more than 50%.

Useful things to know

$$e = \lim_{x\to\infty}(1 + \tfrac{1}{x})^x \text{ and } 1/e = \lim_{x\to\infty}(1 - \tfrac{1}{x})^x$$

Why do we care?

$$(1-a)^b \sim ((1-a)^{1/a})^{ab} \sim e^{-ab} \tag{1}$$

when $a \ll 1$.

Example (Birthday paradox)

Suppose that there are 23 random people in the same room. The probability that at least two of them have the same birthday is more than 50%.

The probability that no two of them have the same birth day is:

$$(1 - \frac{1}{365})(1 - \frac{2}{365}) \cdots (1 - \frac{22}{365}) \sim e^{-\frac{1+2+\ldots+22}{356}} \sim e^{-\frac{11*23}{365}} < 0.5 \tag{2}$$

Useful things to know

$e = \lim_{x \to \infty}(1 + \frac{1}{x})^x$ and $1/e = \lim_{x \to \infty}(1 - \frac{1}{x})^x$

Useful things to know

$$e = \lim_{x \to \infty}(1 + \frac{1}{x})^x \text{ and } 1/e = \lim_{x \to \infty}(1 - \frac{1}{x})^x$$

Why do we care?

$$(1-a)^b \sim ((1-a)^{1/a})^{ab} \sim e^{-ab} \tag{3}$$

Useful things to know

$$e = \lim_{x\to\infty}(1+\tfrac{1}{x})^x \text{ and } 1/e = \lim_{x\to\infty}(1-\tfrac{1}{x})^x$$

Why do we care?

$$(1-a)^b \sim ((1-a)^{1/a})^{ab} \sim e^{-ab} \tag{3}$$

Example (Birthday paradox)

Suppose that there are 23 random people in the same room. The probability that at least two of them have the same birthday is more than 50%.

Useful things to know

$$e = \text{lim}_{x \to \infty}(1 + \tfrac{1}{x})^x \text{ and } 1/e = \text{lim}_{x \to \infty}(1 - \tfrac{1}{x})^x$$

Why do we care?

$$(1-a)^b \sim ((1-a)^{1/a})^{ab} \sim e^{-ab} \tag{3}$$

Example (Birthday paradox)

Suppose that there are 23 random people in the same room. The probability that at least two of them have the same birthday is more than 50%.

The probability that no two of them have the same birth day is:

$$(1 - \frac{1}{365})(1 - \frac{2}{365}) \cdots (1 - \frac{22}{365}) \sim e^{-\frac{1+2+\ldots+22}{356}} \sim e^{-\frac{11*23}{365}} < 0.5 \tag{4}$$

Useful things to know

We will see many applications of hash functions in this class. Good to have a thorough review.

Hash functions

Given a set of integers S and a positive integer m, a hash function is a random map $h : S \to \{0, 1, \ldots, m-1\}$ such that for every $x \neq y \in S$:

$$\Pr[h(x) = h(y)] = \frac{1}{m} \tag{5}$$

Useful things to know

We will see many applications of hash functions in this class. Good to have a thorough review.

Hash functions

Given a set of integers S and a positive integer m, a hash function is a random map $h : S \to \{0, 1, \ldots, m-1\}$ such that for every $x \neq y \in S$:

$$\Pr[h(x) = h(y)] = \frac{1}{m} \tag{5}$$

- Every digital object can be seen as an integer!
- Sometimes we use a stronger assumption, such as $\Pr[h(x) = i] = \frac{1}{m}$ for any x and $i \in \{0, 1, \ldots, m\}$.

Useful things to know

Representing documents by TF-IDF

Given a collection $\mathcal{D} = \{D_1, D_2, \ldots, D_N\}$ of documents, find a vector representation of $\mathcal{D}$.

Let $f_j[i]$ be the frequency of i-th word (in the dictionary) in document D_j. Let N_i be the number of documents contain i-th word.

- Term frequency (TF): $TF_j[i] = \frac{f_j[i]}{\max_k f_j[k]}$.
- Inverse document frequency (IDF): $IDF[i] = \log_2 \frac{N}{N_i}$.

Represent each D_i as a vector $\mathbf{w}_j$ where:

$$w_j[i] = TF_j[i] \cdot IDF[i] \tag{6}$$

(if i-th word is not in D_j, then $w_j[i] = 0$.)

Useful things to know

Power Laws

$y = cx^a$ for some constant a, c.

Some examples:

- Node Degrees in the Web Graph. y is the number if in-link degree to the x-th popula page, then $y \sim cx^{-2}$.
- Amazon Book Sale. y is the number of of sold copies of the x-th popular book, then $y \sim cx^{-2}$.
- Zipf's Law. Oder words appeared in a collection of documents by frequency. y is the number of time x-th word appears, then $y \sim cx^{-1/2}$.