drawback

- does not understand similar meaning words.
- no sumantic similarities between words.

Pearson correlation

	I_1	I_2	I_3	I_4	I_5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

$$Sim(a,b) = \frac{\sum_{p \in P} (r_{a,p} - \bar{r}_a)(r_{b,p} - \bar{r}_b)}{\sqrt{\sum_{p \in P} (r_{a,p} - \bar{r}_a)^2} \sqrt{\sum_{p \in P} (r_{b,p} - \bar{r}_b)^2}}$$

$$Sim(Alice, user1) = \frac{(5-4)\times(3-2.4) + (3-4)\times(1-2.4) + (4-4)\times(2-2.4) + (4-4)(3-2.4)}{\sqrt{(5-4)^2 + (3-4)^2 + (4-4)^2 + (4-4)^2}\sqrt{(3-2.4)^2 + (1-2.4)^2 + (2-2.4)^2 + (3-2.4)^2}}$$

$$= \frac{2}{\sqrt{2}\sqrt{2}}$$

$$= 0.839$$

$$Sim(Alice, user2) = \frac{(5-4)\times(4-3.8) + (3-4)\times(3-3.8) + (4-4)(4-3.8) + (4-4)\times(3-3.8)}{\sqrt{(5-4)^2 + (3-4)^2 + (4-4)^2 + (4-4)^2}\sqrt{(4-3.8)^2 + (3-3.8)^2 + (4-3.8)^2 + (3-3.8)^2}}$$

$$= \frac{1}{\sqrt{2}\sqrt{1.36}}$$

$$= \frac{1}{1.649}$$

$$= 0.606$$

$$\text{Sim}(\text{Alice}, \text{user 3}) = \frac{(5-4)(3-3.2) + (3-4)(3-3.2) + (4-4)(1-3.2) + (4-4)(5-3.2)}{\sqrt{(5-4)^2 + (3-4)^2 + (4-4)^2 + (4-4)^2}\ \sqrt{(3-3.2)^2 + (3-3.2)^2 + (1-3.2)^2 + (5-3.2)^2}}$$

$$= \frac{-0.2 + 0.2 + 0 + 0}{\sqrt{\quad}\ \sqrt{\quad}}$$

$$= 0$$

$$\text{Sim}(\text{Alice}, \text{user 4}) = \frac{(5-4)(1-2.8) + (3-4)(5-2.8) + (4-4)(5-2.8) + (4-4)(2-2.8)}{\sqrt{(5-4)^2 + (3-4)^2 + (4-4)^2 + (4-4)^2}\ \sqrt{(1-2.8)^2 + (5-2.8)^2 + (5-2.8)^2 + (2-2.8)^2}}$$

$$= \frac{-1.8 - 2.2 + 0 + 0}{\sqrt{2}\ \sqrt{13.56}}$$

$$= \frac{-3.96}{5.208}$$

$$= 0.760$$

most similar users are user 1, user 4

Finding ~~similarity between Alice~~ unknown rating

$$r_{\text{Alice}, I_5} = \frac{4 + 0.839(5 - 2.8) + 0.760(1 - 2.8)}{0.839 + 0.760}$$

$$= 4 + \frac{0.503 - 1.368}{1.599}$$

$$= 4 - 0.541$$

$$= 3.459$$

Association rule mining

$X \rightarrow Y$

Antecident consequent

$$\text{confidence} = \frac{\text{support}(X, Y)}{\text{support}(X)}$$

$$\text{Lift} = \frac{\text{support}(X, Y)}{\text{support}(X) \times \text{support}(Y)}$$

Inference

Lift = 1 ⇒ X, Y are independ

> 1 ⇒ +ve correlatio

< 1 ⇒ - ve correlatio

$$\text{Leverage} = \text{support}(X, Y) - \text{support}(X) * \text{support}(Y$$

Assuming 1000 transactions with milk, egg appearing in 300 of them, milk 500, eggs 40 bread 400, milk, bread 400

Find lift (milk → egg) = ?

lift (milk → bread) = ?

If lift value is 1 X and Y are independent of each other

- Lift $(X \rightarrow Y)$ > 1 indicates that there is a positive relationship
- Larger value of lift indicates the stronger association relationship between X and Y

Leverage

If leverage is 0, X and Y statistically independent of each other.

> 0, X and Y have a relationship

a larger value of leverage indicates stronger association rules relationship between X and Y.

$$\text{Lift}(\text{milk} \rightarrow \text{egg}) = \frac{\text{support}(\text{milk}, \text{egg})}{\text{support}(\text{milk}) \times \text{support}(\text{egg})}$$

$$= \frac{300}{400 \times 500} = \frac{0.3}{0.5 \times 0.4}$$

$$= 0.0015 = 1.5$$

$$\text{Lift}(\text{milk} \rightarrow \text{bread}) = \frac{\text{support}(\text{milk}, \text{bread})}{\text{support}(\text{milk})\ \text{support}(\text{bread})}$$

$$= \frac{400}{500 \times 400} = \frac{0.4}{0.5 \times 0.4}$$

$$= 0.002 = 2$$

$$\text{Support}(x) = \frac{\text{no. of times } x \text{ occurs}}{\text{no. of transactions}}$$

$$\text{Leverage}(\text{milk} \rightarrow \text{egg}) = \text{support}(\text{milk}, \text{egg}) - \text{support}(\text{milk})\,\text{support}(\text{egg})$$

$$= 0.3 - 0.5 \times 0.4$$

$$= 0.1$$

$$\text{Leverage}(\text{milk} \rightarrow \text{bread}) = \text{support}(\text{milk}, \text{bread}) - \text{support}(\text{milk})\,\text{support}(\text{bread})$$

$$= 0.4 - 0.5 \times 0.4$$

$$= 0.2$$

$\Rightarrow$ milk $\rightarrow$ bread is strong association.

Streaming Data

26/5/22

eg: Google search Queries in given period of time

- Query data arrives continously and rapidly
- user don't have control
- not uniformly distributed.
- cannot be stored in server
  - sampling operations
  - analytics is performed and summary is stored.

outputs
- approximation / estimation
- probability of max value.

Sliding window

Applications

- counting the distinct elements present in the stream (Flajolet Martin algorithm)
- filtering / selecting data (Bloom's filter)
- count the number of one's in the window (DGIM Algorithm)

Flajolet Martin Algorithm (FM)

- Selecting a Hash function h each element in the set is mapped to a string to atleast $\log_2 n$ bits.
- For each element $r(x)$ = record (x) = length of trailing zeros in $h(x)$.

$$R = \max(r(x))$$

distinct no. of elements $= 2^R$

Find no. of distinct elements

1, 4, 2, 1, 2, 4, 4, 4, 1, 2, 4, 1, 7

Hash function $h(x) = 3x + 1 \bmod 5$

$1 \Rightarrow h(x) = 3(1) + 1 \bmod 5$
$= 4 \bmod 5$
$= 4$

$4 \Rightarrow h(4) = 3(4) + 1 \bmod 5$
$= 13 \bmod 5$
$= 3$

$2 \Rightarrow h(2) = 3(2) + 1 \bmod 5$
$= 7 \bmod 5$
$= 2$

$1 \Rightarrow h(1) = 4$
$2 \Rightarrow h(2) = 2$
$4 \Rightarrow 3$

$4 \Rightarrow 3$
$4 \Rightarrow 3$
$1 \Rightarrow 4$
$2 \Rightarrow 2$
$4 \Rightarrow 3$
$1 \Rightarrow 4$
$7 \Rightarrow h(7) = 3(7) + 1 \bmod 5 = 22 \bmod 5 = 2$

$\therefore$ 4, 3, 2, 4, 2, 3, 3, 3, 4, 2, 3, 4, 2

Convert to binary

100, 011, 010, 100, 010, 011, 011, 011, 100, 010, 011, 100, 010

Find the trailing no. of zeroes

2, 0, 1, 2, 1, 0, 0, 0, 2, 1, 0, 2, 1

record maximum value = 2

distinct number of elements $= 2^R = 2^2 = 4$

DGIM algorithm

- provides 50% approximate output.
- Datar Gionis Indyak Motwani
- count the number of 1's in the window.
- rules that must be followed when representing a stream by buckets
- right end of a bucket is always a position with 1
- every position with a one is in some bucket
- no position is in more than one bucket
- There are one or two buckets of any given size upto some maximum size.
- all sizes must be a power of 2.
- buckets cannot decrease in size as we move to the left.
- estimate the no. of 1's in the window with an error of no more than 50%.

Purchase of items is represented as

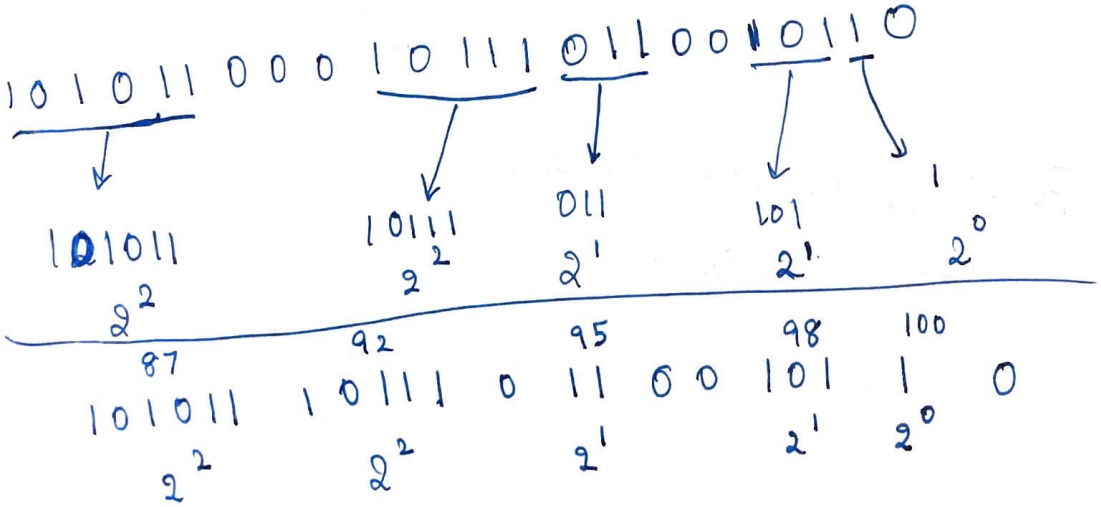
1 0 1 0 1 0 1 1 1 0 0 0 1 1 1 1

where 1 represents purchase of an item.

- we divide the window into buckets the timestamp of its rightmost mention the most recent end.
- No. of one's in the buckets, the number must be a power of 2.

31/5/22

N = 24 [Window size]



new bits

101	102	103	104
0	1	1	1

At timestamp 101,

1010 10111 0 11 00 101 10

102 103 104

new bit zero enters

101011 000 10111 0 11 00 101 1 0 0 ← 1 1 1

2^2 2^2 2^1 2^1 2^1 2^0

new bit 1 enters

101011 000 10111 0 11 00 101 1 001 1

2^2 2^2 2^1 2^1 2^0 2^0

104
1

101011 000 10111 0 11 00 101

100	101	102	103
1	0	0	1
2^0		2^0	2^0

2^2 2^2 2^1 2^1

103

101011 000 10111 0

11	00	101	1001
2^1		2^1	1
			2^1 2^0

2^2 2^2

Filtering / selecting stream

Bloom filtering

- always False negatives are not given but there are probability for false positive.

- works based on hashing technique.

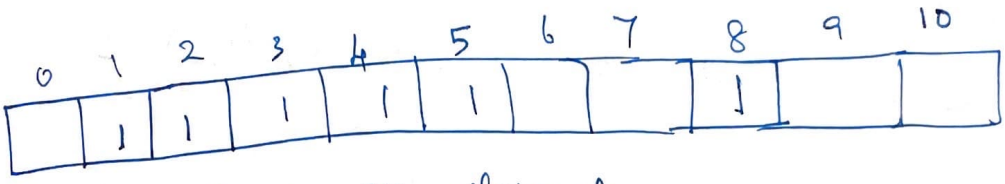
username → [hashfunction] → unique value = 5

- to avoid collision, no. of hash functions is increased.

Insertion

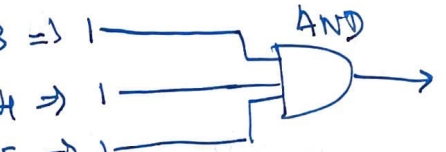
1. ceg ⇒
   - $h_1(ceg) = 3$
   - $h_2(ceg) = 4$
   - $h_3(ceg) = 5$

2. mit ⇒
   - $h_1(mit) = 8$
   - $h_2(mit) = 2$
   - $h_3(mit) = 1$

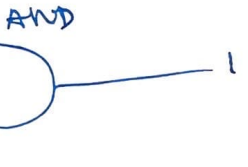


Searching an element

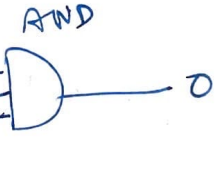
ceg
- $h_1(ceg) = 3 \Rightarrow 1$
- $h_2(ceg) = 4 \Rightarrow 1$
- $h_3(ceg) = 5 \Rightarrow 1$



mit
- $h_1(mit) \Rightarrow 8 = 1$
- $h_2(mit) = 2 \Rightarrow 1$
- $h_3(mit) = 1 \Rightarrow 1$



Anna
- $h_1(anna) = 8 \Rightarrow 1$
- $h_2(anna) = 9 \Rightarrow 0$
- $h_3(anna) = 5 \Rightarrow 1$



Estimating moments.

- A stream consists of elements chosen from a universal set. assume that your universal set is order. Let m_i be the number of occurences of i^{th} element

for any i then the k^{th} order moment of the stream is $\text{Sum}(m_i)^k$

- 0^{th} moment is the count of no. of distinct element in a stream.
- 1^{st} moment is the sum of m_i's must be the length of the stream
- 2^{nd} moment is the sum of squares of m_i's called the surprise number which measures how uneven is the distribution of elements in the stream

Example

Stream S1: a, b, a, c, a, d, a, c

0^{th} order moment $= 4^0 + 1^0 + 2^0 + 1^0$

$= 4$ (no. of distinct elements)

first order moment $= 4^1 + 1^1 + 2^1 + 1$

$= 8$ (length of the stream)

Second order moment $= 4^2 + 1^2 + 2^2 + 1^2$

$= 22$

Find the 1^{st}, 0^{th}, 2^{nd} order moment for

10, 9, 9, 9, 9, 9, 9, 9, 9, 9

0^{th} order moment $= 1^0 + 10^0$

$= 2$

1^{st} order moment $= 1^1 + 10^1$

$= 11$

2^{nd} order moment $= 1^2 + 10^2$

$= 101$

Consider if the length of the stream is 11

One of the 11 elements appear 90 times and the other 10 appears one time each. Find the surprise number.

$$\text{Surprise number} = 90^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2$$

$$= 8100 + 10$$

$$= 8110$$

Alon - Matias Szegedy Algorithm for Second moments. (AMS algorithm)

- a particular element of the universal set which is referred to as x dock element
- An integer x dot value is a value of variable to ~~dord~~ determine the value of variable we choose a position in between 1 to n uniformly and at random.

  Set x dot element = element and initialize x dot value = 1

  as we read the stream add 1 to x dot value each time we encounter another occurance of x dot element.

- Estimate of 2nd moment

$$n(2 \times \text{value} - 1)$$

↓ length of the stream

Take the average for the estimation

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
a	b	c	b	d	a	c	d	a	b	d	c	a	a	b

~~X1 . b = 4~~
~~X1 . d = 2~~

let random positions be 2, 8, 12

x_1 . value = 4 (b) (no. of times b occurs)

x_2 . value = 2 (d)

x_3 . value = 1

x_1 . element = b
x_2 . element = d
x_3 . element = c

estimation $n[2 \times \text{value} - 1]$

$x_1 = 15[2 \times 4 - 1]$
$= 105$

$x_2 = 15[2 \times 2 - 1]$
$= 45$

$x_3 = 15[2 \times 1 - 1]$
$= 15$

$\frac{105 + 45 + 15}{3} = 55$ → approximately equal to 2nd order moment.

Suppose we keep 3 variables x_1, x_2, x_3, assume that at random we pick 3rd, 8th, 13th position to define these variables. Compute 2nd moment of the stream, estimate 2nd moment of the stream and compute average of it.

a, b, c, b, d, a, c, d, a, d, b, d, c, a
a, b .

Ams algorithm

x_1. element = c x_1. value = 3
x_2. element = d x_2. value = 2
x_3. element = a x_3. value = 2

estimation $n\,[2 \times \text{value} - 1]$

$x_1 = 15\,[2 \times 3 - 1]$
$= 75$

$x_2 = 15\,[2 \times 2 - 1]$
$= 45$

$x_3 = 15\,[2 \times 2 - 1]$
$= 45$

2^{nd} order moment $= \dfrac{75 + 45 + 45}{3} = 55$

general method

2nd order moment $= 5^2 + 4^2 + 3^2 + 3^2$
$= 25 + 16 + 9 + 9$
$= 59$

Refer NQSQL

7|6|22

	m_1	m_2	m_3	m_4	m_5
u_1	3	1	1	3	1
u_2	1	2	4	1	3
u_3	3	1	1	3	?
u_4	4	3	5	4	4

	comedy	action
m_1	3	1
m_2	1	2
m_3	1	4
m_4	3	1
m_5	1	3

	comedy	action
u_1	✓	×
u_2	×	✓
u_3	✓	×
u_4	✓	✓

matrix decomposition / Singular Value Decomposition

$m \times n = [m \times a] \times [a \times n]$

$m \times n$: m → no. of users, n → no. of movies

a → rank of the matrix (hidden features / latent features)

	m_1	m_2	m_3	m_4	m_5
u_1	3	1	1	3	1
u_2	1	2	4	1	3
u_3	3	1	1	3	1
u_4	4	3	5	4	4

$m \times n$

4×5

Model Based recommendation System

Advantages of SVD

- storage is reduced ($4 \times 5 = 20$ entries; By SVD ($5 \times 2 + 4 \times 2 = 18$ entries))

---

key value datastore

eg product details by userid

eg userid, username, Orderid, Product name, Product id

9/6/22

Document store
- eg MongoDB
- Can be stored as pdf, word, JSON, formats

FIND.id

Tabular store
- eg HBASE

HBASE does not support join operation.

HIVE
- HQL is used to write map reduce function
- time consuming, thus not suitable for real time applications.

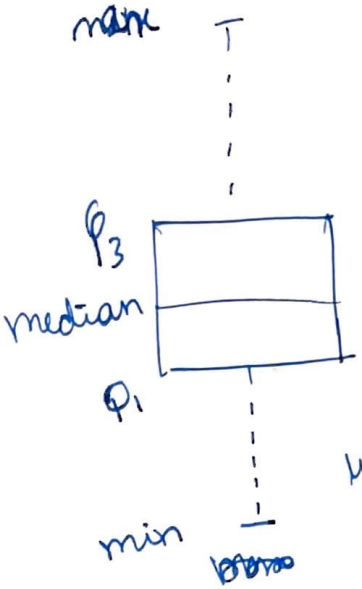
10/6/22

Module 10

Different types of visualization
* basic
* advanced

histogram (data distribution)
box plot (provides 5 number summary)
barplot (compare categorical data)
scatterplot (find relationship between two features)
lineplot (suitable for timeseries data eg stockmarket data) analyse the trend)
piechart (find the contribution of

GGplot - grammar of graphics.
4 arguments (dataset, aesthetics, geometrics, facets)
aesthetics → mapping of x, y coordinates
geometrics → pie, bar, scatter etc.



Box plot

aesthics map to values of x labels, y labels, shape, colour attributes such as

geometrics - show the data is displayed. type of data visualisation

facets - display the subset of data using columns and row

Titanic dataset

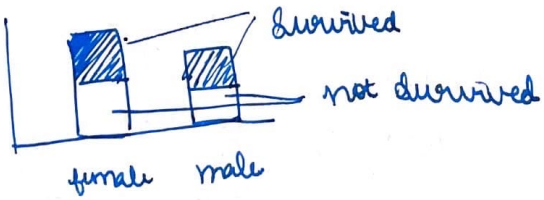
1. survival rate
2. survival by gender
3. survival rate by passenger class
4. Age distribution
5. survival rate by age

bar graph ← ggplot(titanic.aes(x = survived)) + geom_bar()

prop.table(table(titanic(survived))

gives the percentage of each unique value in survived

ggplot(titanic.aes(x = psex, fill = survived)) + geom_bar()



11/6/22

Review of data analytics using R

as factor to convert to categorical variable

exploratory data analysis

- different visualization methods

ggplot

Summary()

str() - Structure of dataset

head()

lm - linear model

Graph memory

graph

- used for social media / maps / friendship network / page rank algorithm.

influence mapping

find centrality.

Types of graph

embedded micronetworks

communication model

·Collaborative communities

influence modeling

distance modeling

important metrics

degree - no. of in connections and out connections.

betweenness -

Centrality - more influential person in the network.

Closeness

~~Fata~~ Fairness

Density