

Tutorial 4 - Linear Algebra Basics

Chuan He

Department of Computing

The Hong Kong Polytechnic University

Vectors and Operations

1. Digitalize and vectorize these datasets

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	High	No
2	No	Married	Medium	No
3	No	Single	Low	No
4	Yes	Married	High	No
5	No	Divorced	Medium	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Medium	Yes
9	No	Married	Medium	No
10	No	Single	Medium	Yes

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Vectors and Operations

What is Data?

- Collection of data **objects** and their **attributes**
- An attribute is a property or characteristic of an object
 - Examples: eye color of a person, temperature, etc.
 - Attribute is also known as **variable**, **field**, **characteristic**, or **feature**
 - Object is also known as **record**, **point**, **case**, **sample**, **entity**, or **instance**
- A collection of attributes describe an object

Objects

Attributes



<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Size: Number of objects

Dimensionality: Number of attributes

Vectors and Operations

1. Digitalize and vectorize these datasets

Steps: 1. identify data type

2. give appropriate digit mapping for each attribute:

Bool: {no, yes} -> {0, 1}

Cat: {Low, Medium, High}

-> {[1, 0, 0],
[0, 1, 0],
[0, 0, 1],}

3. map items by above rules

Boolean Value

Categorized Value

Categorized Value

Boolean Value

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	High	No
2	No	Married	Medium	No
3	No	Single	Low	No
4	Yes	Married	High	No
5	No	Divorced	Medium	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Medium	Yes
9	No	Married	Medium	No
10	No	Single	Medium	Yes

Vectors and Operations

1. Digitalize and vectorize these datasets

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	High	No
2	No	Married	Medium	No
3	No	Single	Low	No
4	Yes	Married	High	No
5	No	Divorced	Medium	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Medium	Yes
9	No	Married	Medium	No
10	No	Single	Medium	Yes



<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	1	1, 0, 0	0, 0, 1	0
2	0	0, 1, 0	0, 1, 0	0
3	0	1, 0, 0	1, 0, 0	0
4	1	0, 1, 0	0, 0, 1	0
5	0	0, 0, 1	0, 1, 0	1
6	0	0, 1, 0	1, 0, 0	0
7	1	0, 0, 1	0, 0, 1	0
8	0	1, 0, 0	0, 1, 0	1
9	0	0, 1, 0	0, 1, 0	0
10	0	1, 0, 0	0, 1, 0	1

Vectors and Operations

- Each object (transaction) is a **set of items**.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

- A set of items can also be represented as a **binary vector**, where each attribute is an item.

Vectors and Operations

1. Digitalize and vectorize these datasets

Categorized value

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

- Steps:
1. identify data type
 2. find the item set (unique items):
{Bread, Coke, Milk, Beer, Diaper}
 3. assign Boolean value for each item
{0, 1} represent {no, yes}
 4. map items w.r.t Boolean value

<i>TID</i>	<i>Items</i>
1	1, 1, 1, 0, 0
2	1, 0, 0, 1, 0
3	0, 1, 1, 1, 1
4	1, 0, 1, 1, 1
5	0, 1, 1, 0, 1

Vectors and Operations

2. The following word vectors/frequencies are generated from 4 datasets.
What can you learn from it?

the 27514
and 14508
i 13088
a 12152
to 10672
of 8702
ramen 8518
was 8274
is 6835
it 6802
in 6402
for 6145
but 5254
that 4540
you 4366
with 4181
pork 4115
my 3841
this 3487
wait 3184
not 3016
we 2984
at 2980
on 2922

the 16710
and 9139
a 8583
i 8415
to 7003
in 5363
it 4606
of 4365
is 4340
burger 432
was 4070
for 3441
but 3284
shack 3278
shake 3172
that 3005
you 2985
my 2514
line 2389
this 2242
fries 2240
on 2204
are 2142
with 2095

the 16010
and 9504
i 7966
to 6524
a 6370
it 5169
of 5159
is 4519
sauce 4020
in 3951
this 3519
was 3453
for 3327
you 3220
that 2769
but 2590
food 2497
on 2350
my 2311
cart 2236
chicken 2220
with 2195
rice 2049
so 1825

the 14241
and 8237
a 8182
i 7001
to 6727
of 4874
you 4515
it 4308
is 4016
was 3791
pastrami 3748
in 3508
for 3424
sandwich 2928
that 2728
but 2715
on 2247
this 2099
my 2064
with 2040
not 1655
your 1622
so 1610
have 1585

Vectors and Operations

- Do simple processing to “**normalize**” the data (remove punctuation, make into lower case, clear white spaces, etc.)
- Break into words, **keep the most popular words**

the 27514
and 14508
i 13088
a 12152
to 10672
of 8702
ramen 8518
was 8274
is 6835
it 6802
in 6402
for 6145
but 5254
that 4540
you 4366
with 4181
pork 4115
my 3841
this 3487
wait 3184
not 3016
we 2984
at 2980
on 2922

the 16710
and 9139
a 8583
i 8415
to 7003
in 5363
it 4606
of 4365
is 4340
burger 432
was 4070
for 3441
but 3284
shack 3278
shake 3172
that 3005
you 2985
my 2514
line 2389
this 2242
fries 2240
on 2204
are 2142
with 2095

the 16010
and 9504
i 7966
to 6524
a 6370
it 5169
of 5159
is 4519
sauce 4020
in 3951
this 3519
was 3453
for 3327
you 3220
that 2769
but 2590

the 14241
and 8237
a 8182
i 7001
to 6727
of 4874
you 4515
it 4308
is 4016
was 3791
pastrami 3748
in 3508
for 3424
sandwich 2928
that 2728
but 2715

Most frequent words are **stop words**

cart 2236
chicken 2220
with 2195
rice 2049
so 1825

with 2040
not 1655
your 1622
so 1610
have 1585

Vectors and Operations

Possible findings:

1. Most frequent words are stop words
2. Commonly used words in reviews, less interesting
3. Notional words indicate these four datasets may come from four countries/regions (Japan, U.S., Mid-east, Israel)

the 27514	the 16710	the 16010	the 14241
and 14508	and 9139	and 9504	and 8237
i 13088	a 8583	i 7966	a 8182
a 12152	i 8415	to 6524	i 7001
to 10672	to 7003	a 6370	to 6727
of 8702	in 5363	it 5169	of 4874
ramen 8518	it 4606	of 5159	you 4515
was 8274	of 4365	is 4519	it 4308
is 6835	is 4340	sauce 4020	is 4016
it 6802	was 4070	in 3951	was 3791
in 6402	for 3441	this 3519	pastrami 3748
for 6145	but 3284	was 3453	in 3508
but 5254	shack 3278	for 3327	for 3424
that 4540	shake 3172	you 3220	sandwich 2928
you 4366	that 3005	that 2769	that 2728
with 4181	you 2985	but 2590	but 2715
pork 4115	my 2514	food 2497	on 2247
my 3841	line 2389	on 2350	this 2099
this 3487	this 2242	my 2311	my 2064
wait 3184	fries 2240	cart 2236	with 2040
not 3016	on 2204	chicken 2220	not 1655
we 2984	are 2142	with 2195	your 1622
at 2980	with 2095	rice 2049	so 1610
on 2922		so 1825	have 1585

Norm and Distance of Vectors

3. The topic ratios of reports in 3 websites is shown in the table below. Find the Euclidean distance of each two websites. What can you learn from it?

document	Apple	Microsoft	Obama	Election
website1	0.35	0.5	0.1	0.05
website2	0.4	0.4	0.1	0.1
website2	0.05	0.05	0.6	0.3

Norm and Distance of Vectors

3. The topic ratios of reports in 3 websites is shown in the table below. Find the Euclidean distance of each two websites. What can you learn from it?

document	Apple	Microsoft	Obama	Election
website1	0.35	0.5	0.1	0.05
website2	0.4	0.4	0.1	0.1
website3	0.05	0.05	0.6	0.3

L₂ norm: Euclidean distance:

$$L_2(x, y) = \sqrt{|x_1 - y_1|^2 + \dots + |x_d - y_d|^2}$$

Shown in distance table, website1 and website2 are more similar (may be technology websites), website3 could be political website.

	W1	W2	W3
W1	0	0.122	0.778
W2	0.122	0	0.731
W3	0.778	0.731	0

Clustering

4. This table shows the result of k-means clustering for some dataset. Can you judge its performance?

K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports
1	3	5	40	506	96	27
2	4	7	280	29	39	2
3	1	1	1	7	4	671
4	10	162	3	119	73	2
5	331	22	5	70	13	23
6	5	358	12	212	48	13
Total	354	555	341	943	273	738

Clustering

4. This table shows the result of k-means clustering for some dataset. Can you judge its performance?

Observe the majority class of each cluster

K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports
1	3	5	40	506	96	27
2	4	7	280	29	39	2
3	1	1	1	7	4	671
4	10	162	3	119	73	2
5	331	22	5	70	13	23
6	5	358	12	212	48	13
Total	354	555	341	943	273	738

Remark

1. Data preparation (digitalize and vectorize)
2. Data types (Boolean and categorized)
3. Word vector & frequency
4. Measurement (distance)