แบบฝึกหัด

1. กำหนดแอทริบิวต์ดังต่อไปนี้ จงจำแนกประเภทของแอทริบิวต์ว่าเป็นแอทริบิวต์เชิงคุณภาพ (qualitative) ชนิดใด (nominal, ordinal) หรือเป็นแอทริบิวต์เชิงปริมาณ (quantitative) ชนิดใด (interval, ratio) นอกจากนี้ให้จำแนกด้วยว่าแอทริบิวต์แต่ละตัวมีชนิดเป็น Binary, Discrete, หรือ Continuous
   (ก) เวลาในรูปแบบ AM หรือ PM
   (ข) ความสว่างที่วัดโดยมิเตอร์วัดแสง
   (ค) ความสว่างที่วัดจากความรู้สึกของมนุษย์

(ง) มุมที่วัดเป็นองศาระหว่าง 0 ถึง 360

(จ) เหรียญทอง เหรียญเงิน เหรียญทองแดง ของกีฬาโอลิมปิก

(ฉ) ความสูงจากระดับน้ำทะเล

(ช) จำนวนผู้ป่วยในโรงพยาบาลแห่งหนึ่ง

(ซ) เลข ISBN ของหนังสือ

(ฌ) ความหนาแน่นของสสารหน่วยเป็น กรัมต่อลูกบาศก์เซนติเมตร

(ญ) ระยะทางจากจุดศูนย์กลางของวิทยาเขตหน่วยเป็นเมตร

(ฎ) ชั้นยศของกองทัพ

(ฏ) ความสามารถในการส่งผ่านคลื่นแสง: opaque translucent transparent

2. จงยกตัวอย่างสถาณการณ์ที่ identification numbers (รหัสประจำตัว) น่าจะมีประโยชน์สำหรับการทำนาย
3. ปริมาณใดต่อไปนี้ที่มีคุณสมบัติ spatial autocorrelation : daily rainfall หรือ daily temperature และทำไมจึงเป็นเช่นนั้น
4. โปรแกรมเมอร์คนหนึ่งได้ออกแบบอัลกอริทึม k-nearest neighbors ดังนี้

Algorithm 2.1 Algorithm for finding K nearest neighbors.

1: **for** $i = 1$ to *number of data objects* **do**
2: Find the distances of the i^{th} object to all other objects.
3: Sort these distances in decreasing order.
(Keep track of which object is associated with each distance.)
4: **return** the objects associated with the first K distances of the sorted list
5: **end for**

(ก) จงอภิปรายว่าจะมีปัญหาอะไรเกิดขึ้นได้บ้างกับอัลกอริทึมนี้ ถ้าดาต้าเซตมีข้อมูลซ้ำ (duplicates)

(ข) จงเสนอวิธีการแก้ไขปัญหาที่เกิดจากการมีข้อมูลซ้ำซ้อนในดาต้าเซต

5. คำนวณค่า cosine, correlation, Jaccard และ Euclidean distance ของ ดาต้าอ็อบเจ็กต์ x และ y ดังต่อไปนี้

(ก) x = (1, 1, 1, 1), y = (2, 2, 2, 2)

(ข) x = (0, 1, 0, 1), y = (1, 0, 1, 0)

(ค) x = (0, -1, 0, 1), y = (1, 0, -1, 0)

6. คำนวณค่า Mutual information ของดาต้าอ็อบเจ็กต์ x และ y ดังต่อไปนี้

(ก) x = (-7, -2, 1, 0, 1, 2), y = (9, 4, 1, 0, 4, 1)

(ข) x = (1,1,1,1), y = (2,2,2,2)

เอกสารอ้างอิง

[1] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar. “Introduction to Data Mining”. Pearson, 2nd edition, 2018.

1. ก.) เชิงคุณภาพ ชนิด nominal แบบ binary

ข.) เชิงปริมาณ interval แบบ continuous

ค.) เชิงคุณภาพ ordinal แบบ discrete

ง.) เชิงปริมาณ ratio แบบ continuous

จ.) เชิงคุณภาพ ordinal แบบ discrete

ฉ.) เชิงปริมาณ interval แบบ continuous

ช.) เชิงปริมาณ interval แบบ discrete

ซ.) เชิงคุณภาพ nominal แบบ discrete

ฌ.) เชิงปริมาณ ratio แบบ continuous

ญ.) เชิงปริมาณ interval แบบ continuous

ฎ.) เชิงคุณภาพ ordinal แบบ discrete

ฏ.) เชิงปริมาณ ratio แบบ continuous

2. - ต้องการรู้ช่วงอายุ

- ต้องการรู้ช่วงปีจบ

- ต้องการรู้วิชาที่น่าจะเรียนในปัจจุบัน

3. daily temperature เพราะ ข้อมูลจากพื้นที่หนึ่งที่ได้นั้นจะมีค่าใกล้เคียงกลับพื้นที่ ๆ อยู่ใกล้ ๆ

4. ก. - มีประสิทธิภาพต่ำจากการที่ต้องนำตัวซ้ำมาประมวลผลร่วมด้วย

- ได้เพื่อนบ้านที่เป็นตัวซ้ำกับตัวต้น

- หากมีตัวใกล้ ๆ เป็นตัวซ้ำอาจทำให้เพื่อนบ้านที่ได้มีแต่ตัวนั้น ๆ

ข. จัดรวมกลุ่มตัวซ้ำเป็นกลุ่มเดียวกันแล้วให้นับกลุ่มนั้นเป็น data object ตัวหนึ่ง

5.)

ก. **cosine**

$\langle x,y \rangle = 1*2 + 1*2 + 1*2 + 1*2 = 8$

$||x|| = (1^2+1^2+1^2+1^2)^{1/2} = 2$ $||y|| = (2^2+2^2+2^2+2^2)^{1/2} = 4$

Cosine = 8/2*4 = 1

Correlation mean(x) = (1+1+1+1)/4 = 1 mean(y) = (2+2+2+2)/4 = 2

$S_x = (1/4-1*((1-1)^2+(1-1)^2+(1-1)^2+(1-1)^2))^{1/2} = 0$ $S_y = (1/4-1*((2-2)^2+(2-2)^2+(2-2)^2+(2-2)^2))^{1/2} = 0$

$S_{xy} = (1/4-1*((1-1)(2-2)+(1-1)(2-2)+(1-1)(2-2)+(1-1)(2-2)))^{1/2} = 0$

Correlation = 0/0*0 = 0

Jaccard

$f_{1,2} = 4$

J = 0/4 = 0

Euclidean distance

$d(x,y) = ((1-2)^2+(1-2)^2+(1-2)^2+(1-2)^2)^{1/2} = 2$

ข.) **cosine**

<x,y> = 0*1 + 1*0 + 0*1+ 1*0 = 0

$||x|| = (0^2+1^2+0^2+1^2)^{1/2} = 2^{1/2}$ $||y|| = (1^2+0^2+1^2+0^2)^{1/2} = 2^{1/2}$

$Cosine = 0/2^{1/2}*2^{1/2} = 0$

Correlation mean(x) = (1+0+1+0)/4 = 1/2 mean(y) = (0+1+0+1)/4 = 1/2

$S_x = (1/4-1*((1-1/2)^2+(0-1/2)^2+(1-1/2)^2+(0-1/2)^2))^{1/2} = 1/3^{1/2}$ $S_y = (1/4-1*((0-1/2)^2+(1-1/2)^2+(0-1/2)^2+(1-1/2)^2))^{1/2} = 1/3^{1/2}$

$S_{xy} = (1/4-1*((1-1/2)(0-1/2)+(0-1/2)(1-1/2)+(1-1/2)(0-1/2)+(0-1/2)(1-1/2)))^{1/2} = (-1/3)^{1/2}$

$Correlation = (-1/3)^{1/2}/(1/3)^{1/2}*(1/3)^{1/2} = 1/(-1/3)^{1/2}$

Jaccard

$f_{1,0} = 2$ $f_{0,1} = 2$

J = 0/2+2 = 0

Euclidean distance

$d(x,y) = ((1-0)^2+(0-1)^2+(1-0)^2+(0-1)^2)^{1/2} = 2$

ค. **cosine**

<x,y> = 0*1 + -1*0 + 0*-1 + 1*0 = 0

$||x|| = (0^2+1^2+0^2+(-1)^2)^{1/2} = 2^{1/2}$ $||y|| = (1^2+0^2+(-1)^2+0^2)^{1/2} = 2^{1/2}$

$Cosine = 0/2^{1/2}*2^{1/2} = 0$

Correlation mean(x) = (0+1+0-1)/4 = 0 mean(y) = (1+0-1+0)/4 = 0

$S_x = (1/4-1*((0-0)^2+(1-0)^2+(0-0)^2+(-1-0)^2))^{1/2} = 2/3^{1/2}$ $S_y = (1/4-1*((1-0)^2+(0-0)^2+(-1-0)^2+(0-0)^2))^{1/2} = 1/3^{1/2}$

$S_{xy} = (1/4-1*((0-0)(1-0)+(1-0)(0-0)+(0-0)(-1-0)+(1-0)(0-0)))^{1/2} = 0$

$Correlation = 0/1/3^{1/2}*1/3^{1/2} = 0$

Jaccard

$F_{0,1} = 1$ $f_{-1,0} = 1$ $f_{0,-1} = 1$ $f_{-1,0} = 1$

J = 0/(1+1+1+1) = 0

Euclidean distance

$d(x,y) = ((0-1)^2+(1-0)^2+(0-(-1))^2+(-1-0)^2)^{1/2} = 2$

6. ก.

x_j	$P(X=x_j)$	$-P(X=x_j)\log_2 P(X=x_j)$
-7	1/6	0.43082
-2	1/6	0.43082
0	1/6	0.43082
1	2/6	0.52832
2	1/6	0.43082
H(x)		2.2516

Y_k	$P(Y=y_k)$	$-P(Y=y_k)\log_2 P(Y=y_k)$
0	1/6	0.43082
1	2/6	0.52832
4	2/6	0.52832
9	1/6	0.43082
H(y)		1.91828

x_j	Y_k	$P(X=x_j, Y=y_k)$	$-P(X=x_j, Y=y_k)\log_2 P(X=x_j,Y=y_k)$
-7	9	1/6	0.43082
-2	4	1/6	0.43082
1	1	1/6	0.43082
0	0	1/6	0.43082
1	4	1/6	0.43082
2	1	1/6	0.43082
H(x,y)			2.58492

I(x,y) = 2.2516+1.91828-2.58492 = 1.58496

ข.

x_j	$P(X=x_j)$	$-P(X=x_j)\log_2 P(X=x_j)$
1	4/4	0
H(x)		0

Y_k	$P(Y=y_k)$	$-P(Y=y_k)\log_2 P(Y=y_k)$
2	4/4	0
H(y)		0

x_j	Y_k	$P(X=x_j, Y=y_k)$	$-P(X=x_j, Y=y_k)\log_2 P(X=x_j,Y=y_k)$
1	2	4/4	0
H(x,y)			0

I(x,y) = 0+0-0 = 0

สรุป

ประเภทของข้อมูล

- **ชุดข้อมูล** คือ กลุ่มของ data objects ประกอบด้วย Attributes ที่บอกลักษณะ

ตารางที่ 2.1. ชุดข้อมูลนิสิต (Student Information Data Set)

Attributes

Data objects

ลำดับที่	รหัสนิสิต	ชั้นปี	เกรดเฉลี่ยสะสม
1	1034261	2	2.75
2	1034262	3	3.24
3	1034263	2	3.51
4	1034265	1	2.99
5	1034266	3	3.12

Attributes ของ Data objects สามารถ เปลี่ยน แปลงได้ตลอด
เช่น อุณหภูมิพื้นผิวโลก-มหาสมุทร

การแบ่งประเภทของ Attributes โดย Operation ของระบบจำนวน

ตารางที่ 2.2. ชนิดของแอทริบิวต์

ชนิดของแอทริบิวต์		คำอธิบาย	ตัวอย่าง	โอเปอเรชั่น
Categorical (เชิงคุณภาพ)	Nominal	ค่าของ Nominal attribute สามารถใช้ในการแยกแยะดาต้าอ็อบเจ็กต์ได้ ด้วยโอเปอเรชั่น Distinctness (=, ≠)	รหัสไปรษณีย์ รหัสพนักงาน สีตา เพศ	ฐานนิยม, entropy, contingency correlation, Chi-squared test
	Ordinal ก. binary ค. Discete จ.	มีคุณสมบัติและโอเปอเรชั่น Distinctness เช่นเดียวกันกับ Nominal attributes และค่าของ Ordinal attribute สามารถใช้ในการเรียงลำดับดาต้าอ็อบเจ็กต์ได้ ด้วยโอเปอเรชั่น Order (<, >)	ความแข็งของแร่ธาตุ, เกรด {A, B+, B, C+, C, D+, D, F},	มัธยฐาน, เปอร์เซ็นต์ไทล์, rank correlation, run tests, sign tests
Numeric (เชิงปริมาณ)	Interval	มีคุณสมบัติและโอเปอเรชั่น Distinctness และ Order เช่นเดียวกันกับ Nominal attributes และ Ordinal attributes นอกจากนี้ ความแตกต่างระหว่าง interval attributes สองค่า คำนวณได้ด้วยโอเปอเรชั่น Addition (+, -) สามารถตีความได้ กล่าวคือ interval attributes จะมีหน่วยของการวัด	อุณหภูมิในหน่วยองศาเซลเซียส หรือองศาฟาเรนไฮต์, วันที่ตามปฏิทิน	ค่าเฉลี่ย, ส่วนเบี่ยงเบนมาตรฐาน, Pearson's correlation, t-test, F-test
	Ratio ง. Continuous	มีคุณสมบัติและโอเปอเรชั่น Distinctness, Order, และ Interval เช่นเดียวกันกับ Nominal attributes, Ordinal attributes, และ Interval attributes นอกจากนี้ อัตราส่วนของ ratio attributes ซึ่งคำนวณได้โดยใช้โอเปอเรชั่น Multiplication (x, /) สามารถตีความได้	อุณหภูมิในหน่วยเคลวิน (Kelvin), อายุ, มวล, ความยาว, กระแสไฟฟ้า	ค่าเฉลี่ยเรขาคณิต, ค่าเฉลี่ยฮาร์มอนิค, เปอร์เซ็นต์ความผันแปร

การแบ่งประเภท Attributes ด้วยจำนวนของค่าที่เป็นไปได้

1. Discrete Attribute ที่มีค่าเป็นไปได้ไม่กี่ค่าหรือจำกัด แต่สามารถหนึ่งแสนแสนได้ มักใช้กับ Attribute เชิงคุณภาพ
2. Binary Attribute มีแค่ 2 ค่า 0/1, จริง/เท็จ, ใช่/ไม่ใช่, ชาย/หญิง
3. Continuous Attribute มีค่าเป็นจำนวนจริง มักใช้กับ Attributes เชิงปริมาณ

ชนิดของ Data sets มี 3 ประเภท

1. ข้อมูล Record data

ข้อมูลที่ใช้ในการทำเหมืองข้อมูล มักอยู่ในรูปแบบของกลุ่ม Record data ซึ่งประกอบด้วย Attributes หรือ ฟิลด์ข้อมูลต่างๆ

Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
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

(a) Record data.

TID	ITEMS
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diapers, Milk
4	Beer, Bread, Diapers, Milk
5	Soda, Diapers, Milk

(b) Transaction data.

Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

(c) Data matrix.

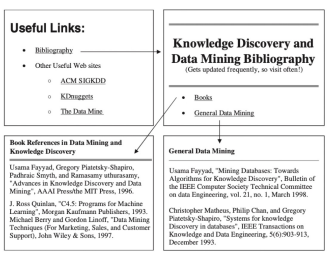
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

(d) Document-term matrix.

2. ข้อมูลกราฟ

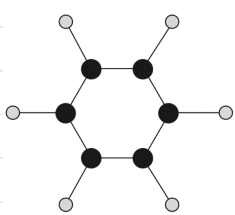
กราฟเป็นโครงสร้างข้อมูลสำหรับแสดงความสัมพันธ์ระหว่าง Data object เรามักใช้ทำกราฟ 2 กรณี

1. แสดงความสัมพันธ์ระหว่าง Data object



(a) Linked web pages.

2. Data object มีโครงสร้างแบบกราฟ



(b) Benzene molecule.

3. ข้อมูลแบบมีลำดับ

ข้อมูลบางชนิดประกอบด้วย Attributes ที่มีความสัมพันธ์กันเชิงเวลาหรือเชิงพื้นที่

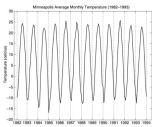
Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

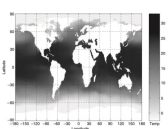
(a) Sequential transaction data.

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

(b) Genomic sequence data.



(c) Temperature time series.



(d) Spatial temperature data.

คุณภาพข้อมูล

การแก้ปัญหาคุณภาพข้อมูล เทคนิคการจัดการกับคุณภาพของข้อมูล แบ่งได้ 2 กลุ่ม

1. เทคนิคสำหรับการตรวจจับและการแก้ไขปัญหาคุณภาพข้อมูล
2. การใช้อัลกอริทึมการทำแบบจำลองข้อมูลที่ทนทานต่อข้อมูลคุณภาพต่ำ

ความผิดพลาดจากการวัดและการเก็บข้อมูล

ปัญหาคุณภาพข้อมูลเกิดได้จากความผิดพลาดของมนุษย์ ข้อจำกัดของอุปกรณ์

เช่น ข้อมูลรบกวน, ข้อมูลเที่ยม, ความล่าช้า, ความเที่ยงตรง และความแม่นยำ

- ความผิดพลาดจากการวัด
- ความผิดพลาดจากการเก็บข้อมูล

ปัญหาคุณภาพที่เกี่ยวข้องกับการประยุกต์ใช้งาน

ข้อมูลที่มีคุณภาพสูง คือข้อมูลที่เหมาะกับการประยุกต์ใช้งาน คุณภาพข้อมูลจากมุมมองของการใช้

แต่ปัญหาทั่วไปของคุณภาพต่ำไปใช้งาน มี 3 อย่าง

1. Timelines เวลาในการเก็บ
2. Relevance ข้อมูลทุกชิ้นที่จำเป็นต้องใช้กับวัตถุประสงค์
3. Knowledge about the Data คุณภาพเอกสาร

คะแนน Present

- ได้ 8 คะแนนกลับ มีเนื้อหาเพิ่มเติมให้ พูดชัดเจน อธิบายน่าฟัง
- เค้า 2 คะแนน เพราะพูดตกตกหล่น พูดเพื่อเยอะไป