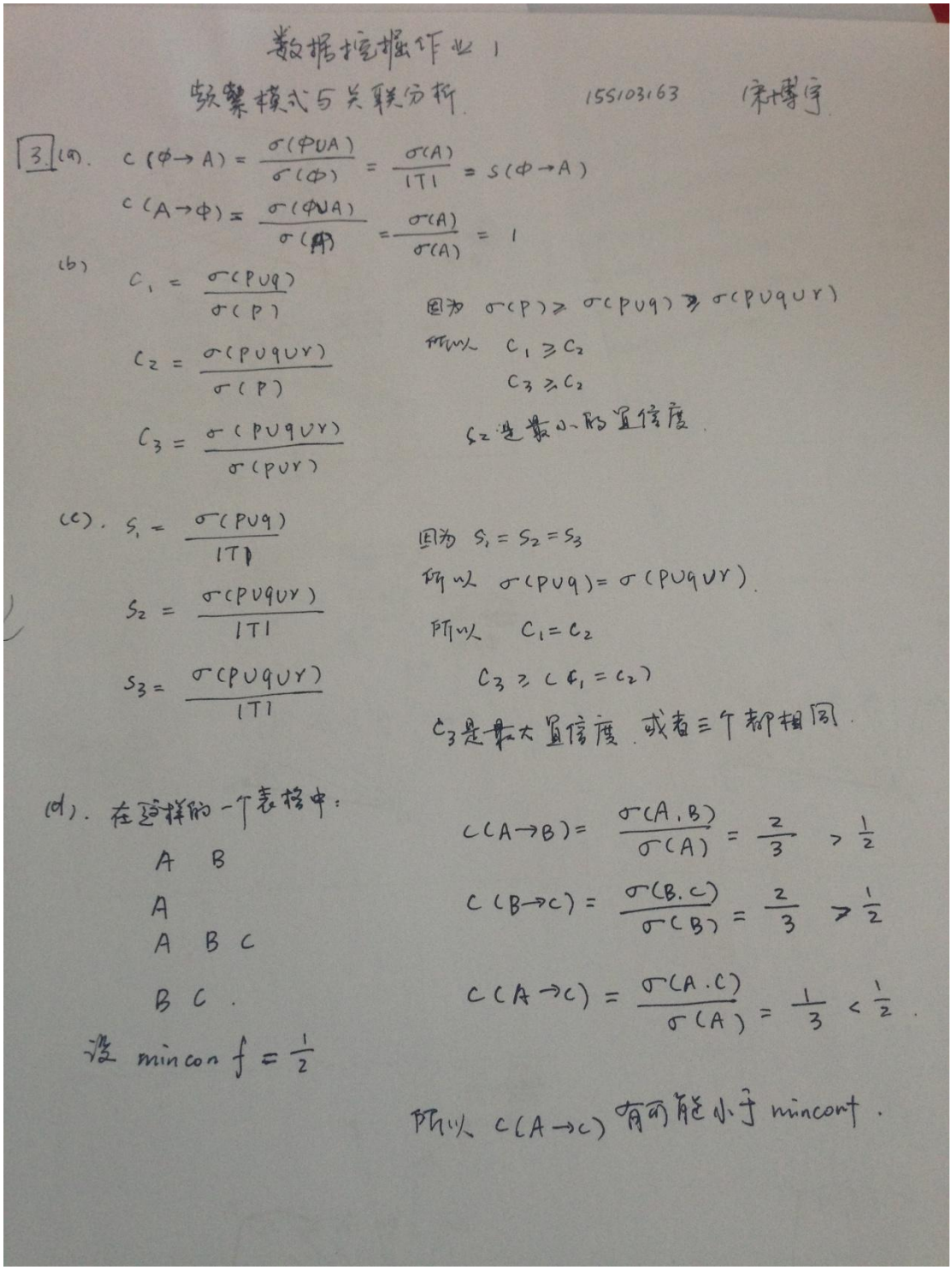
Assignment 1
Oct 12, 2015

Exercises 3, 8, 9, 11, 14, 16 in Chapter 6 of the textbook "Introduction to Data Mining".

Students who don't have the textbook can download the pdf file named "ch6.pdf" from the QQ group for this course (Group#: 324742036).

Please complete the assignment using the following MS Word template and print on A4 papers. Please make sure your name and ID are correctly filled in your assignment.

Assignment due date: Oct 26, 2015, on class in classroom A12.



数据挖掘作业1

频繁模式与关联分析　　155103163　宋博宇

3. (a). $c(\phi \to A) = \frac{\sigma(\phi \cup A)}{\sigma(\phi)} = \frac{\sigma(A)}{|T|} = s(\phi \to A)$

$c(A \to \phi) = \frac{\sigma(\phi \cup A)}{\sigma(A)} = \frac{\sigma(A)}{\sigma(A)} = 1$

(b)

$c_1 = \frac{\sigma(P \cup q)}{\sigma(P)}$

$c_2 = \frac{\sigma(P \cup q \cup Y)}{\sigma(P)}$

$c_3 = \frac{\sigma(P \cup q \cup Y)}{\sigma(P \cup Y)}$

因为 $\sigma(P) \geq \sigma(P \cup q) \geq \sigma(P \cup q \cup Y)$

所以 $c_1 \geq c_2$

$c_3 \geq c_2$

c_2 是最小的置信度.

(c). $s_1 = \frac{\sigma(P \cup q)}{|T|}$

$s_2 = \frac{\sigma(P \cup q \cup Y)}{|T|}$

$s_3 = \frac{\sigma(P \cup q \cup Y)}{|T|}$

因为 $s_1 = s_2 = s_3$

所以 $\sigma(P \cup q) = \sigma(P \cup q \cup Y)$.

所以 $c_1 = c_2$

$c_3 \geq (c_1 = c_2)$

c_3 是最大置信度. 或者三个都相同.

(d). 在这样的一个表格中：

A B

A

A B C

B C.

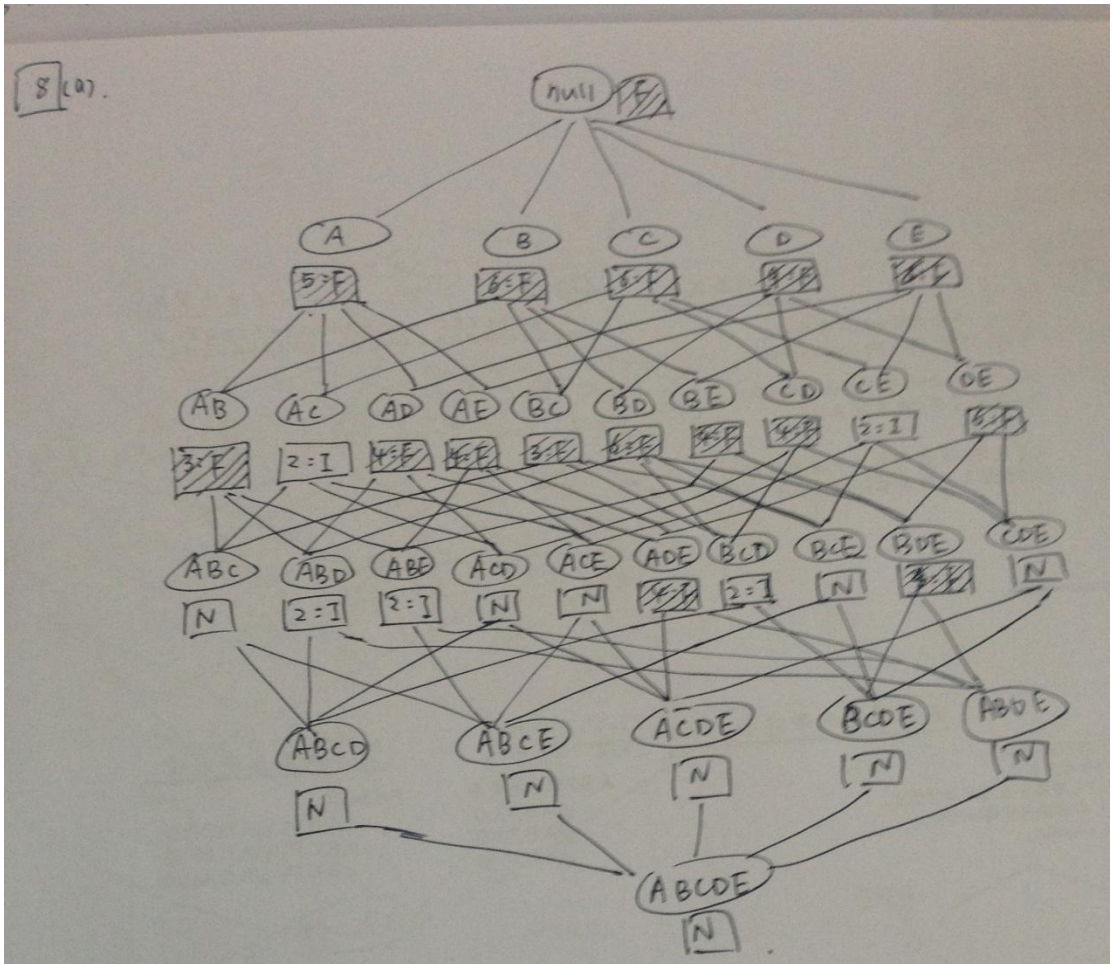
设 $minconf = \frac{1}{2}$

$c(A \to B) = \frac{\sigma(A, B)}{\sigma(A)} = \frac{2}{3} > \frac{1}{2}$

$c(B \to C) = \frac{\sigma(B, C)}{\sigma(B)} = \frac{2}{3} > \frac{1}{2}$

$c(A \to C) = \frac{\sigma(A, C)}{\sigma(A)} = \frac{1}{3} < \frac{1}{2}$.

所以 $c(A \to C)$ 有可能小于 minconf.



(b) $\frac{F}{T} = \frac{16}{32} = 50\%$ (包括N).

(c). $\frac{N}{T} = \frac{11}{32} = 34.4\%$

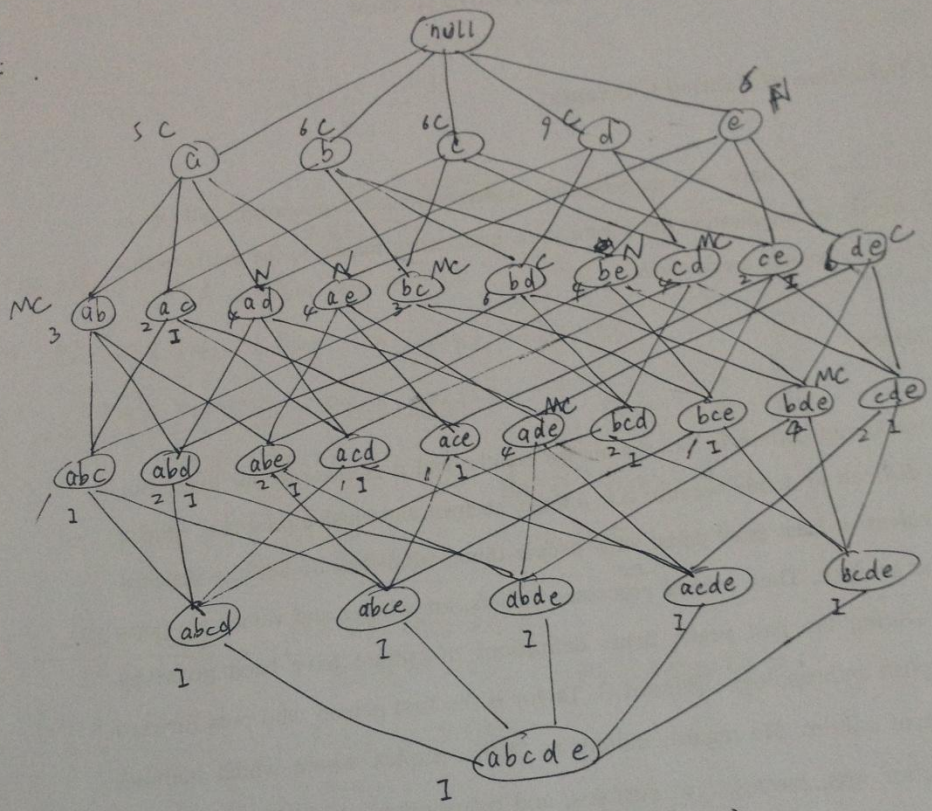
(d). $\frac{I}{T} = \frac{5}{32} = 15.6\%$.

姓名：宋博宇　　学号：15S103163　　日期： 2015-10-24

9. (a). L_1 L_3 L_5 L_9 L_{11}.

(b). {1 4 5} L_1　{1 5 8} L_3　{4 5 8} L_3

11.



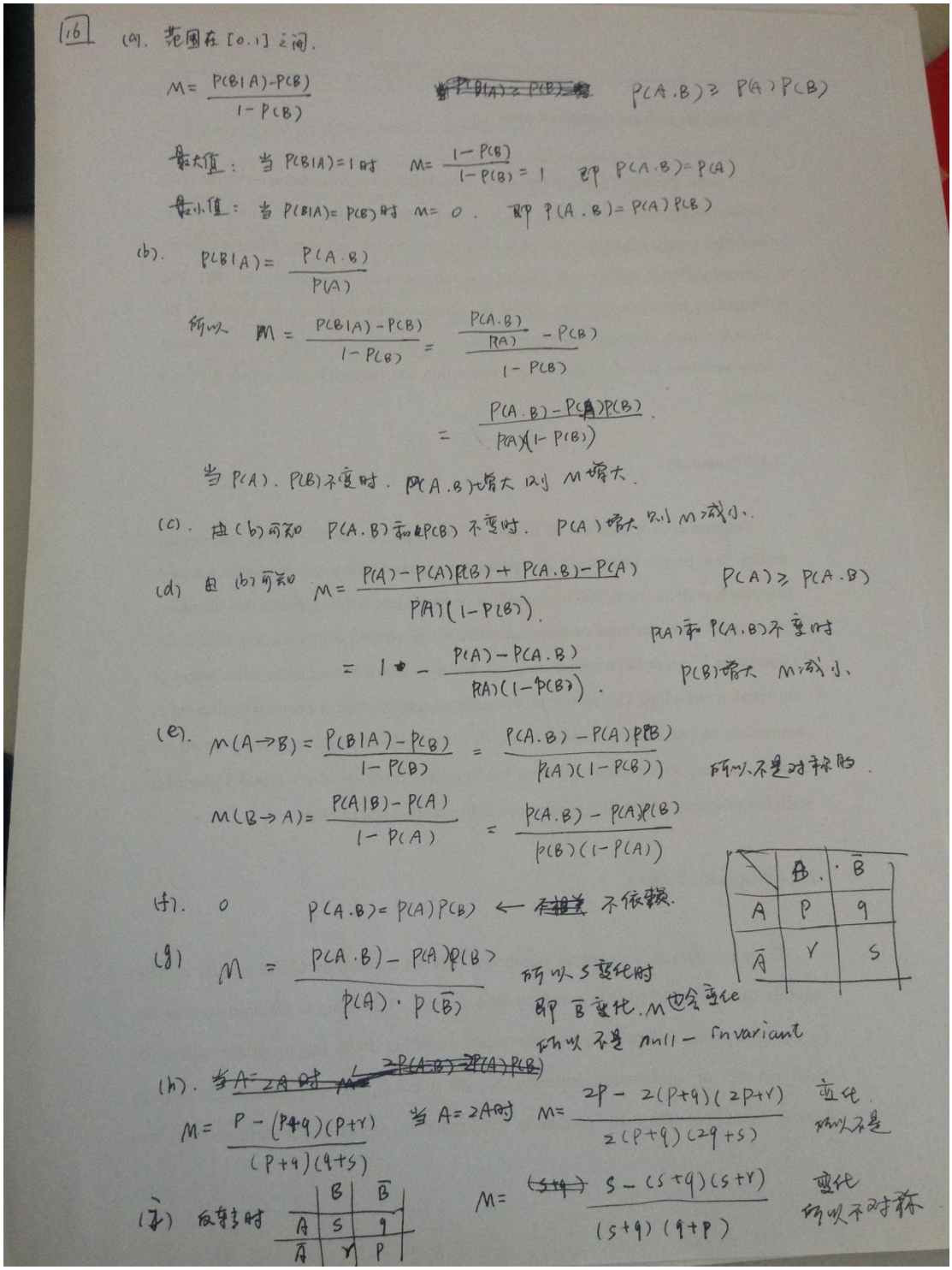
14. (a). e. 有频繁项集最长，所以会产生最多频繁项集

(b). d. 没有项集个数大于1000，所以频繁项集个数为0.

(c). e

(d). b. 最左侧长条表示有最大支持度.

(e). e. 支持度变化最广.



16 (a). 范围在 [0,1] 之间.

$M = \frac{P(B|A)-P(B)}{1-P(B)}$　　$P(A,B) \geq P(A)P(B)$

最大值：当 $P(B|A)=1$ 时　$M = \frac{1-P(B)}{1-P(B)} = 1$　即 $P(A,B)=P(A)$

最小值：当 $P(B|A)=P(B)$ 时　$M=0$.　即 $P(A,B)=P(A)P(B)$

(b). $P(B|A) = \frac{P(A,B)}{P(A)}$

所以 $M = \frac{P(B|A)-P(B)}{1-P(B)} = \frac{\frac{P(A,B)}{P(A)} - P(B)}{1-P(B)}$

$= \frac{P(A,B)-P(A)P(B)}{P(A)(1-P(B))}$.

当 $P(A)$, $P(B)$ 不变时，$P(A,B)$ 越大以时 M 越大.

(c). 由 (b) 可知 $P(A,B)$ 和 $P(B)$ 不变时，$P(A)$ 增大以时 M 减小.

(d) 由 (b) 可知 $M = \frac{P(A)-P(A)P(B)+P(A,B)-P(A)}{P(A)(1-P(B))}$.　$P(A) \geq P(A,B)$

$= 1 - \frac{P(A)-P(A,B)}{P(A)(1-P(B))}$.　$P(A)$ 和 $P(A,B)$ 不变时，$P(B)$ 增大 M 减小.

(e). $M(A \to B) = \frac{P(B|A)-P(B)}{1-P(B)} = \frac{P(A,B)-P(A)P(B)}{P(A)(1-P(B))}$　所以不是对称的.

$M(B \to A) = \frac{P(A|B)-P(A)}{1-P(A)} = \frac{P(A,B)-P(A)P(B)}{P(B)(1-P(A))}$

	B	$\bar{B}$
A	p	q
$\bar{A}$	r	s

(f). 0　　$P(A,B)=P(A)P(B)$ ← 不依赖.

(g) $M = \frac{P(A,B)-P(A)P(B)}{P(A)\cdot P(\bar{B})}$　所以 s 变化时，即 $\bar{B}$ 变化，M 也会变化，所以不是 null-invariant

(h). 当 $A=2A$ 时　$M = \frac{2p-2(p+q)(2p+r)}{2(p+q)(2q+s)}$　变化，所以不是

当 $A=2A$ 时　$M = \frac{p-(p+q)(p+r)}{(p+q)(q+s)}$

(i) 反转时

	B	$\bar{B}$
A	s	q
$\bar{A}$	r	p

$M = \frac{s-(s+q)(s+r)}{(s+q)(q+p)}$　变化，所以不对称