1. (9 pt) Consider the dataset in Fig 1, with points belonging to two classes, blue squares and red circles.

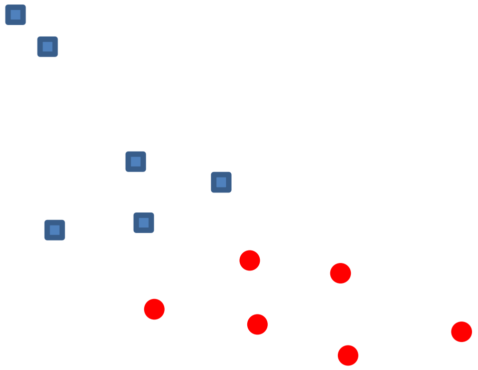
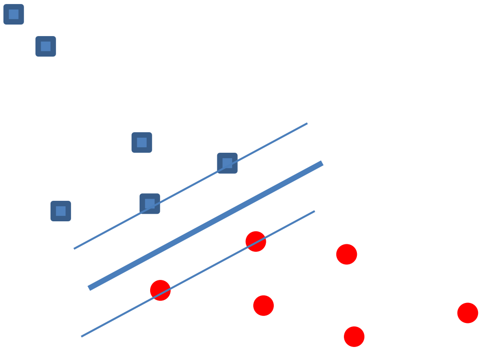


Fig. 1

(a) [1 pt] Draw (approximately) the SVM line separator.



(b) [1 pt] Suppose we find (1/2)***w**^{2} to be 2 in the SVM optimization. What is the margin, i.e. the distance of closest points to the line?

||w||=2
Margin = 1/||w|| = 1/ 2

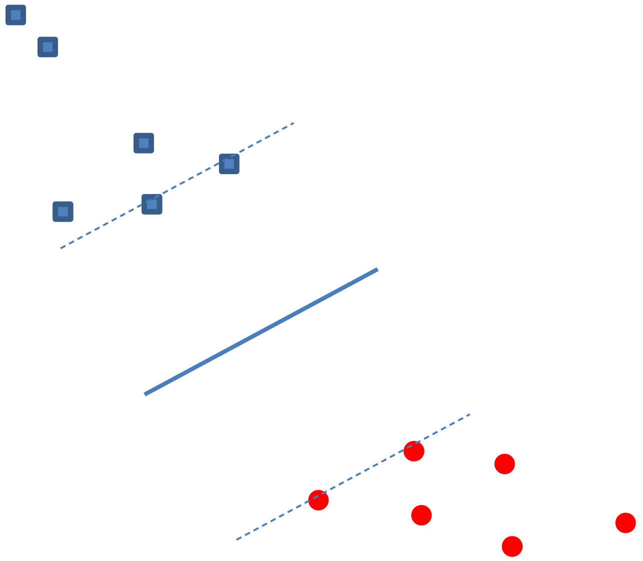


Fig. 2

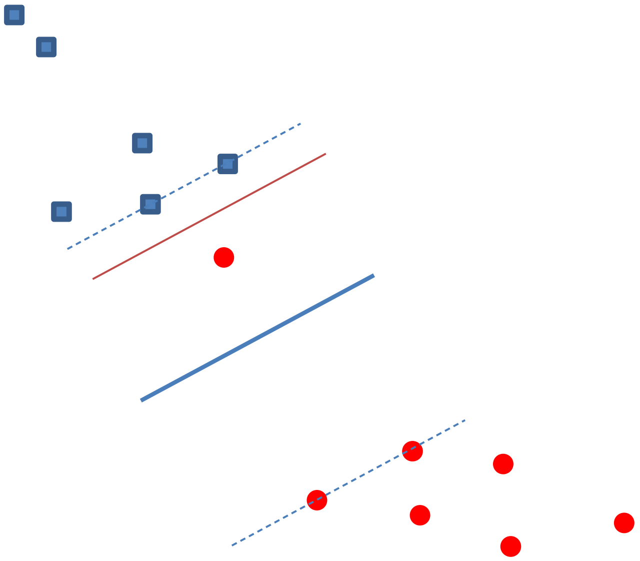


Fig. 3

(c) [1 pt] Now consider the dataset in Fig 2 (the red points are shifted below). Will (1/2)***w**^{2} be smaller or greater than previously? Explain.

Margin will be greater, so (1/2)*w^2 be will be smaller (since maximizing margin means minimizing w)

(d) [2 pt] Using a ruler, and the fact that (1/2)***w**^{2} was 2 previously, find (approximately) the magnitude of the new line coefficient vector, **w**'.

See lines in Fig 2.

1/||w'|| = 4*(1/||w||) = 4*(1/2) = 2 so ||w'|| = 1/2

(e) [3 pt] Consider the dataset in Fig 3 (with one additional red circle quite close to the blue squares). Assuming optimization using slack variables and C=1, draw a line that does not perfectly separate the points, but which is nonetheless better than the line that perfectly separates the points. (Draw it in the figure, and explain why).

See (bold) blue line in Fig. 3.

The red line that perfectly separates the points will have a higher cost than the blue line because the red line has a smaller margin than the black.

To see this, observe that ξ for the new red circle will be equal to 1.5 (or slightly less).

The cost of the blue line is: (1/2)*w'^2+ξ because it does not perfectly separate the points. The margin error ξ for the red circle is between 1 and 2 . The other ξ's will be zero because we have only one margin error (one red circle on the wrong side).

The cost of the red line is: (1/2)*w^2 because it perfectly separates the points (the red circles from the blue squares).

All the ξ's will be zero because every red circle and blue square are separated correctly even though the margin is smaller than the blue line.

Now let's plug in the values for ||**w**'||, ||**w**||, and ξ:

||w'||= 1/2, ||w||=2, and ξ=1.5:

Blue line, (1/2)*$w'^2+\xi$ = (1/2)*$(1/2)^2$ +(1.5) = 1.625

Red line, (1/2)*w^2 = (1/2)*$(2)^2$ = 2

(f) [1 pt] Why would we rather prefer the line in (e) to the line that perfectly separates the points?

Because the blue line has a smaller ||w|| (||w||=1.625) than the red line (||w||=2)

Thus, the blue has greater margin than the red line.

Question 2

(4 pt) Consider the task of building a classifier from random data, where the attribute values are generated randomly irrespective of the class labels. Assume the data set contains records from two classes, "+" and "−." Half of the data set is used for training while the remaining half is used for testing.

(a) (1 pt) Suppose there are an equal number of positive and negative records in the data and a classifier predicts every test record to be positive. What is the expected error rate of the classifier on the test data?

Suppose we have 100 test instances, 50 positive and 50 negative. The classifier will classify all 50 positive instances correctly and will classify all 50 negative instances incorrectly. Hence, the error rate is 50/100 = 50%

(b) Repeat the previous analysis assuming that the classifier predicts each test record to be positive class with probability 0.8 and negative class with probability 0.2.

Suppose we have 100 test instances, 50 positive and 50 negative. the error rate is 50/100 = 50% because since the probability to predict positive class is 0.8 > 0.5, so it will classify all 50 positive instances correctly. And since the probability to predict negative class is 0.2 < 0.5, it will classify all 50 negative instances incorrectly.

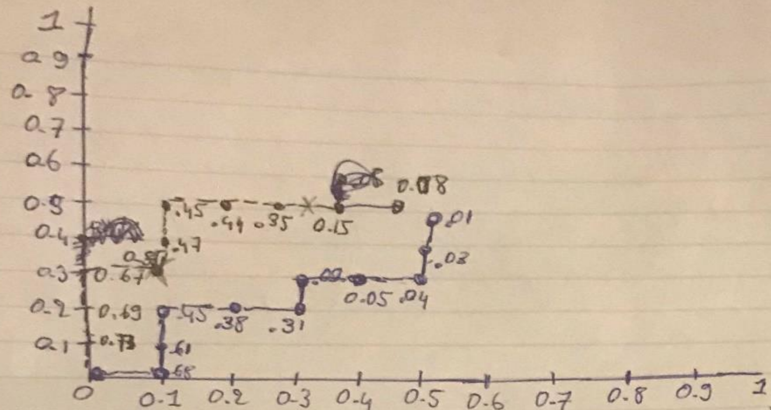
(c) Suppose two-thirds of the data belong to the positive class and the remaining one-third belong to the negative class. What is the expected error of a classifier that predicts every test record to be positive?

Suppose we have 100 test instances, 67 positive and 33 negative. the error rate is 33/100 = 33% because since every negative instance will be classified incorrectly. (There are 33 negative instances, and there are incorrectly classified positive)

(d) Repeat the previous analysis assuming that the classifier predicts each test record to be positive class with probability 2/3 and negative class with probability 1/3.

The error rate is (2/3)*(2/3) = 44.4% Because the probability for each record to be positive class is 2/3

Question 3



A is better since it is above B, and more of its points are on the north west.

b) Confusion matrix

	+	−
+	3	2
−	1	4

$$\text{precision} = \frac{TP}{TP+FP} = \frac{3}{3+1} = 75\%$$

$$\text{re call} = \frac{TP}{P} = \frac{3}{5} = 60\%$$

$$\text{F-mesure} = \frac{2\times(0.75)(0.6)}{0.75+0.6} = 67\%$$

c)

	+	−
+	1	4
−	1	4

$$\text{precision} = \frac{1}{2} = 50\%$$

$$\text{re call} = \frac{1}{5} = 20\%$$

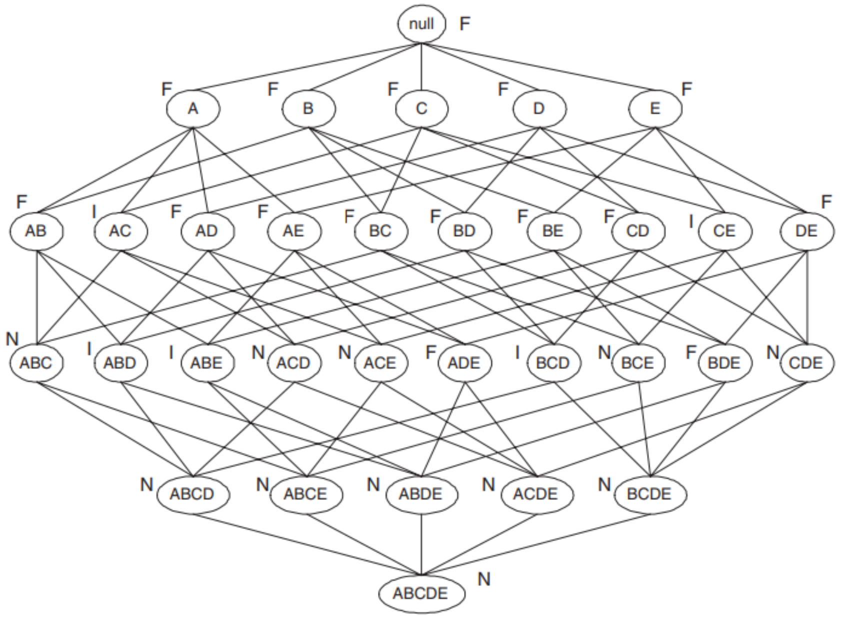
$$\text{F-mesure} = \frac{2\times(0.5)(0.2)}{0.5+0.2} = 0.2857 \simeq 29\%$$

A is better since it F. mesure
67% is greater than B F- mesure
29%. Yes the results are consistant
with what I expect from ROC curve.

d) see the graph
t = 0.5, One application could be predicted
whether its sunny or not and t = 0.1 can
be used for serious and consequently things
as diseases like a patient has a cancer or
not.

Question 4

(a) Draw an itemset lattice representing the data set



(b) What is the percentage of frequent itemsets (with respect to all itemsets in the lattice

16/32 = 50 %

(c) What is the pruning ratio of the Apriori algorithm on this data set?

11/32 = 34.40 %

(d) What is the false alarm rate (i.e, percentage of candidate itemsets that are found to be infrequent after performing support counting)?

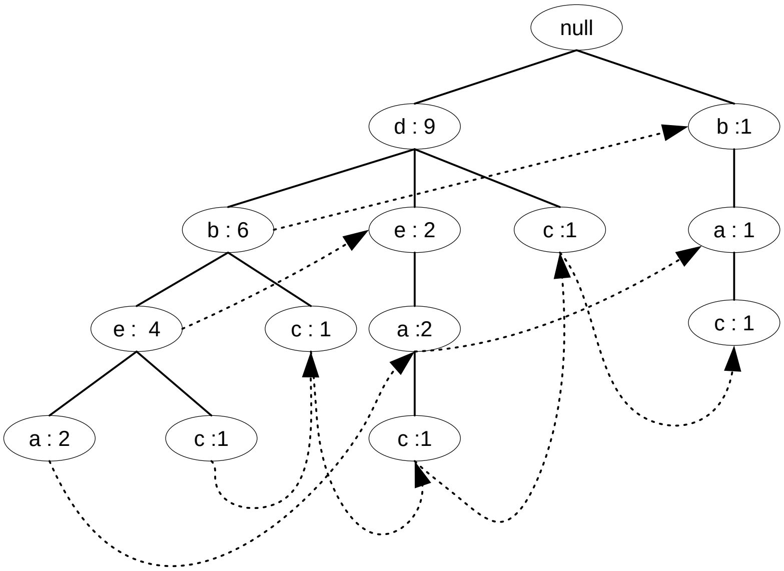
5/32 = 15.60 %

Question 5

Using the data in the above table, build an FP-Tree, and then mine the frequent itemsets using FP-Growth with minsup=30%.

Frequent items	Count
d	9
b	7
e	6
a	5
c	5

TID	Filtered ordered transactions
1	{d,b,e,a}
2	{d,b,c}
3	{d,b,e,a}
4	{d,e,a,c}
5	{d,b,e,c}
6	{d,b,e}
7	{d,c}
8	{b,a,c}
9	{d,e,a}
10	{d,b}



Suffix c: => Frequent itemset (FI) = {c}

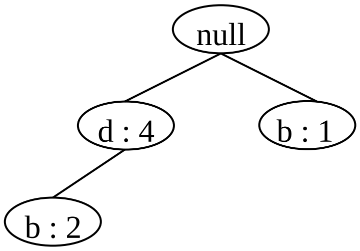
Conditional pattern base for “c” i.e. the fragment of this tree containing transactions with c (without drawing c):

d , b , e : 1
d , b : 1
d , e , a : 1
d : 1
b , a : 1

Frequent items:

d	4
b	3

Conditional FP-tree for “c”:



Suffix bc: => Frequent itemset (FI) = {b,c}

Conditional pattern base for “bc”, i.e. the fragment of this conditional tree containing transactions with b (without drawing b):

d : 2 infrequent

Suffix dc: => Frequent itemset (FI) = {d,c}

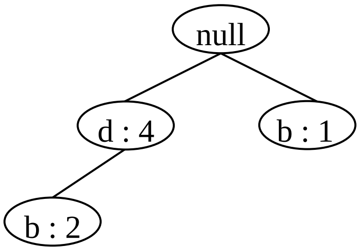
Conditional pattern base for “dc”:

Nothing

FIs so far = {c} : 5 , {b,c} : 3 , {d,c} : 4

Suffix a: => Frequent itemset (FI) = {a} = 5

d,b,e: 2
d,e,c: 1
b,c:1
d,e:1



Suffix ba: => Frequent itemset (FI) = {b,a} = 3

d:2

e:2

infrequent

Suffix ea: => Frequent itemset (FI) = {e,a}

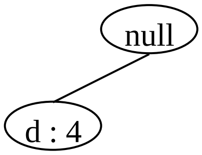
Conditional pattern base for “ea”:

d :
4

Frequent items:
for “ea”:

d	4

Conditional FP-tree



One path FP-tree. Frequent itemset (FI): {d,e,a}

Suffix da: => Frequent itemset (FI) = {d,a} = 4

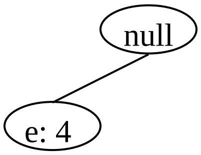
b,e:2

e,a:1

e:1

frequent items:

e:4



Suffix e: => Frequent itemset (FI) = {e} = 6

d,b,a:2

d,a,c:1

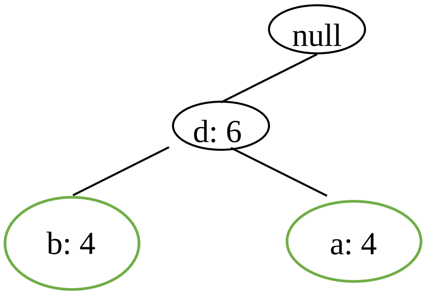
d,b,c:1

d,b:1

d,a:1

frequent items:

d:6 ; b:4, a:4



Suffix b: => Frequent itemset (FI) = {b} = 7

d,e,a:2

d,c:1

d,e,c:1

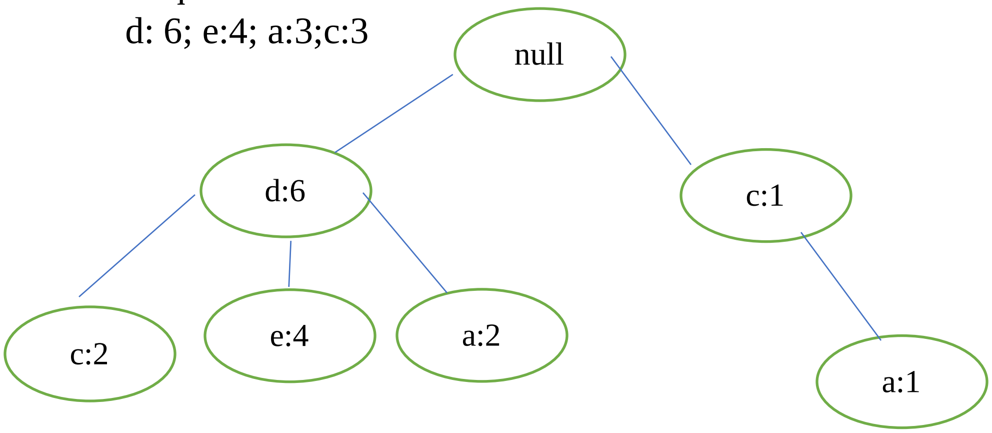
d,e:1

a,c:1

d:1

frequent items

d: 6; e:4; a:3;c:3



Suffix d: => Frequent itemset (FI) = {d} = 9

b,e,a:2

b,c:1

e,a,c:1

b,e,c:1

b,e:1

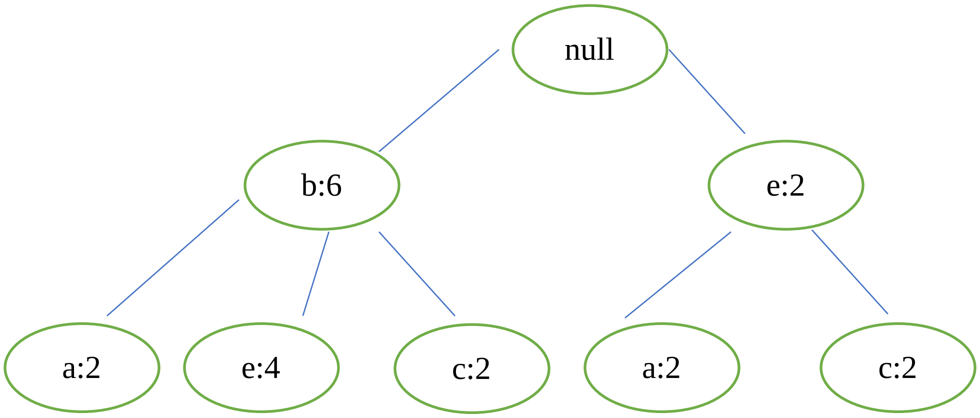
c:1

e,a:1

b:1

frequents items

b:6, e:6, c: 4, a: 4

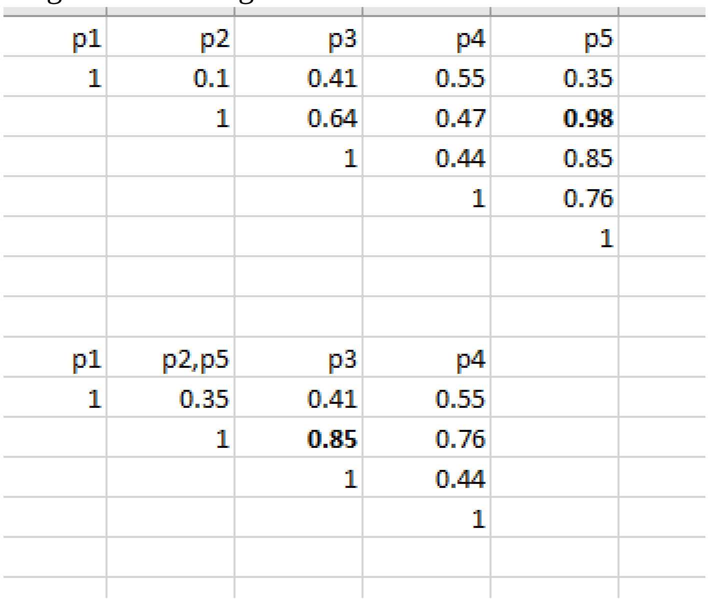


All Frequent itemsets

{c} : 5 , {b,c} : 3 , {d,c} : 4
{a} : 5 , {b,a} : 3 , {e,a} : 4 , {d,e,a} : 4 , {d,a} : 4
{e} : 6 , {b,e} : 4 , {d,e} : 6 , {d,b,e} : 4
{b} : 7 , {d,b} : 6
{d} : 9

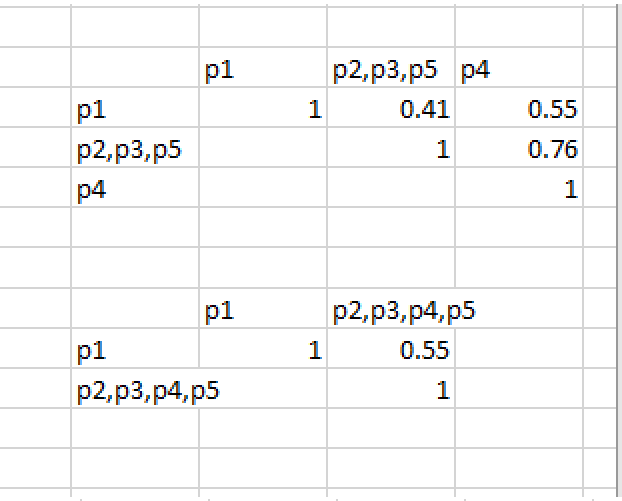
Q 6. (6 points) Use the following similarity matrix to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged. (“Single link clustering” means clustering using MIN, while “Complete link clustering” means clustering using MAX)

Sing link clustering



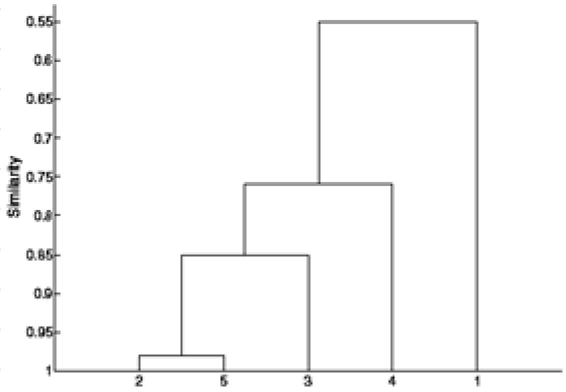
p1	p2	p3	p4	p5
1	0.1	0.41	0.55	0.35
	1	0.64	0.47	**0.98**
		1	0.44	0.85
			1	0.76
				1

p1	p2,p5	p3	p4
1	0.35	0.41	0.55
	1	**0.85**	0.76
		1	0.44
			1

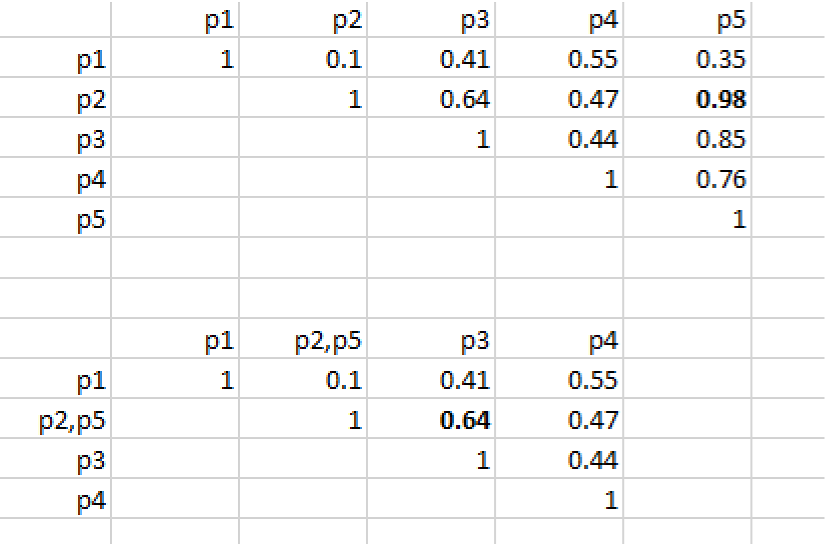


	p1	p2,p3,p5	p4
p1	1	0.41	0.55
p2,p3,p5		1	0.76
p4			1

	p1	p2,p3,p4,p5
p1	1	0.55
p2,p3,p4,p5		1

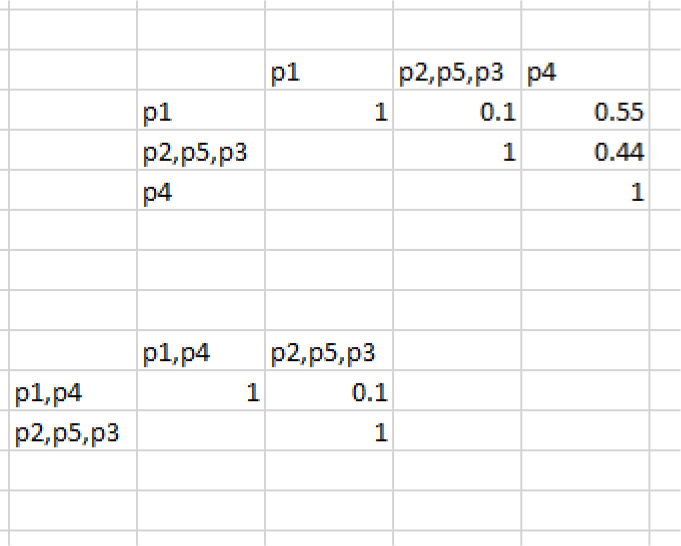


Complete link clustering



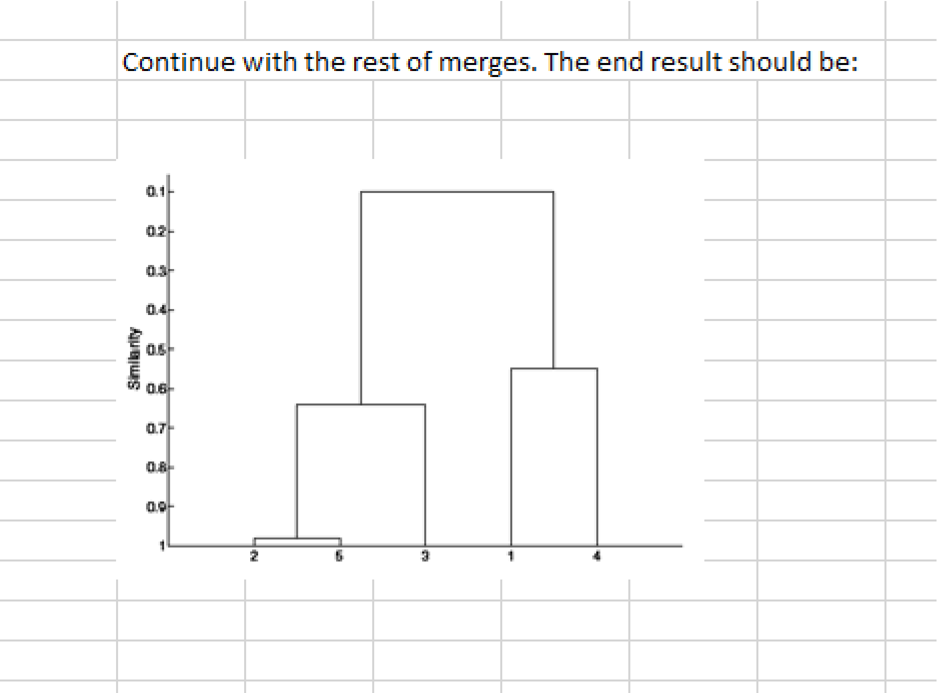
	p1	p2	p3	p4	p5
p1	1	0.1	0.41	0.55	0.35
p2		1	0.64	0.47	**0.98**
p3			1	0.44	0.85
p4				1	0.76
p5					1

	p1	p2,p5	p3	p4
p1	1	0.1	0.41	0.55
p2,p5		1	**0.64**	0.47
p3			1	0.44
p4				1



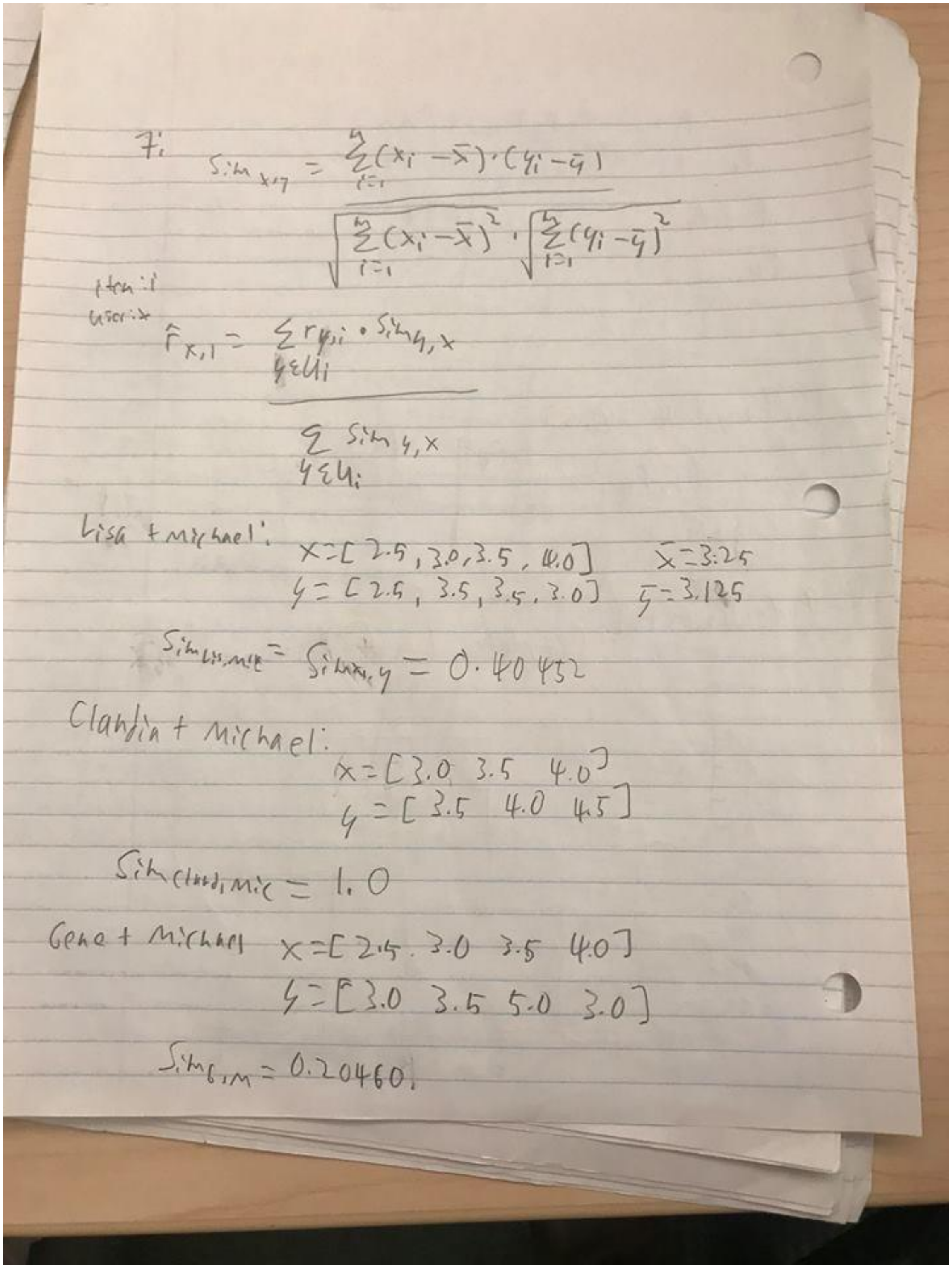
	p1	p2,p5,p3	p4
p1	1	0.1	0.55
p2,p5,p3		1	0.44
p4			1

	p1,p4	p2,p5,p3
p1,p4	1	0.1
p2,p5,p3		1



Continue with the rest of merges. The end result should be:

7-



7.

$$sim_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

item: i

user: x

$$\hat{r}_{x,i} = \frac{\sum_{y \in U_i} r_{y,i} \cdot sim_{y,x}}{\sum_{y \in U_i} sim_{y,x}}$$

Lisa + Michael: $x = [2.5, 3.0, 3.5, 4.0]$ $\bar{x} = 3.25$

$y = [2.5, 3.5, 3.5, 3.0]$ $\bar{y} = 3.125$

$sim_{lis,mic} = sim_{x,y} = 0.40452$

Clandia + Michael:

$x = [3.0\ 3.5\ 4.0]$

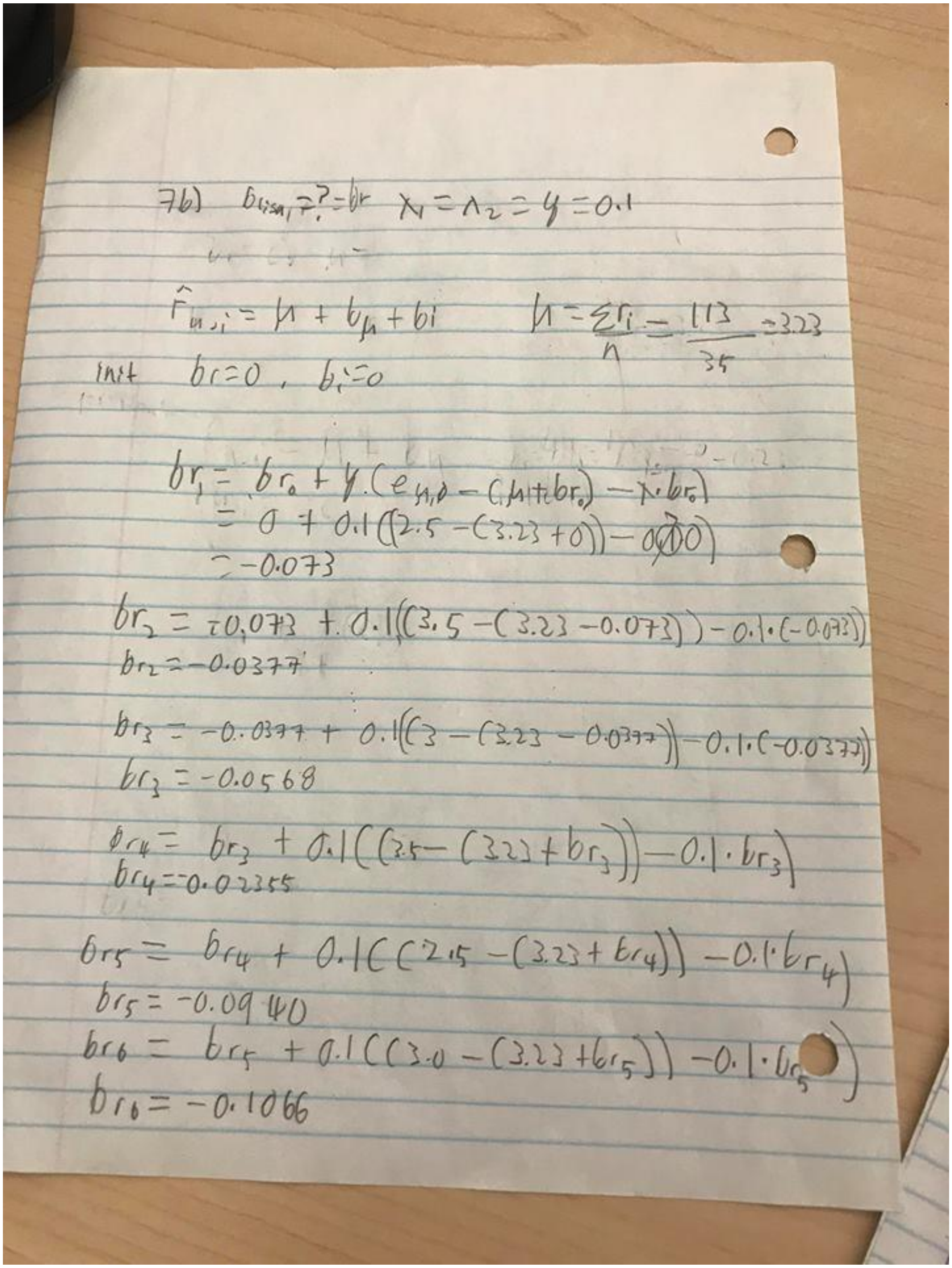
$y = [3.5\ 4.0\ 4.5]$

$sim_{clard,mic} = 1.0$

Gene + Michael $x = [2.5.\ 3.0\ 3.5\ 4.0]$

$y = [3.0\ 3.5\ 5.0\ 3.0]$

$sim_{G,M} = 0.20460.$



7b) $b_{Lisa,?} = ? = br$ $\lambda_1 = \lambda_2 = \gamma = 0.1$

$\hat{r}_{u,i} = \mu + b_\mu + bi$ $\mu = \frac{\sum r_i}{n} = \frac{113}{35} = 3.23$

init $br = 0$, $b_i = 0$

$br_1 = br_0 + \gamma\cdot(e_{u,i} - (\mu + br_0) - \lambda\cdot br_0)$

$= 0 + 0.1((2.5 - (3.23 + 0)) - 0.1\cdot 0)$

$= -0.073$

$br_2 = -0.073 + 0.1((3.5 - (3.23 - 0.073)) - 0.1\cdot(-0.073))$

$br_2 = -0.0377$

$br_3 = -0.0377 + 0.1((3 - (3.23 - 0.0377)) - 0.1\cdot(-0.0377))$

$br_3 = -0.0568$

$br_4 = br_3 + 0.1((3.5 - (3.23 + br_3)) - 0.1\cdot br_3)$

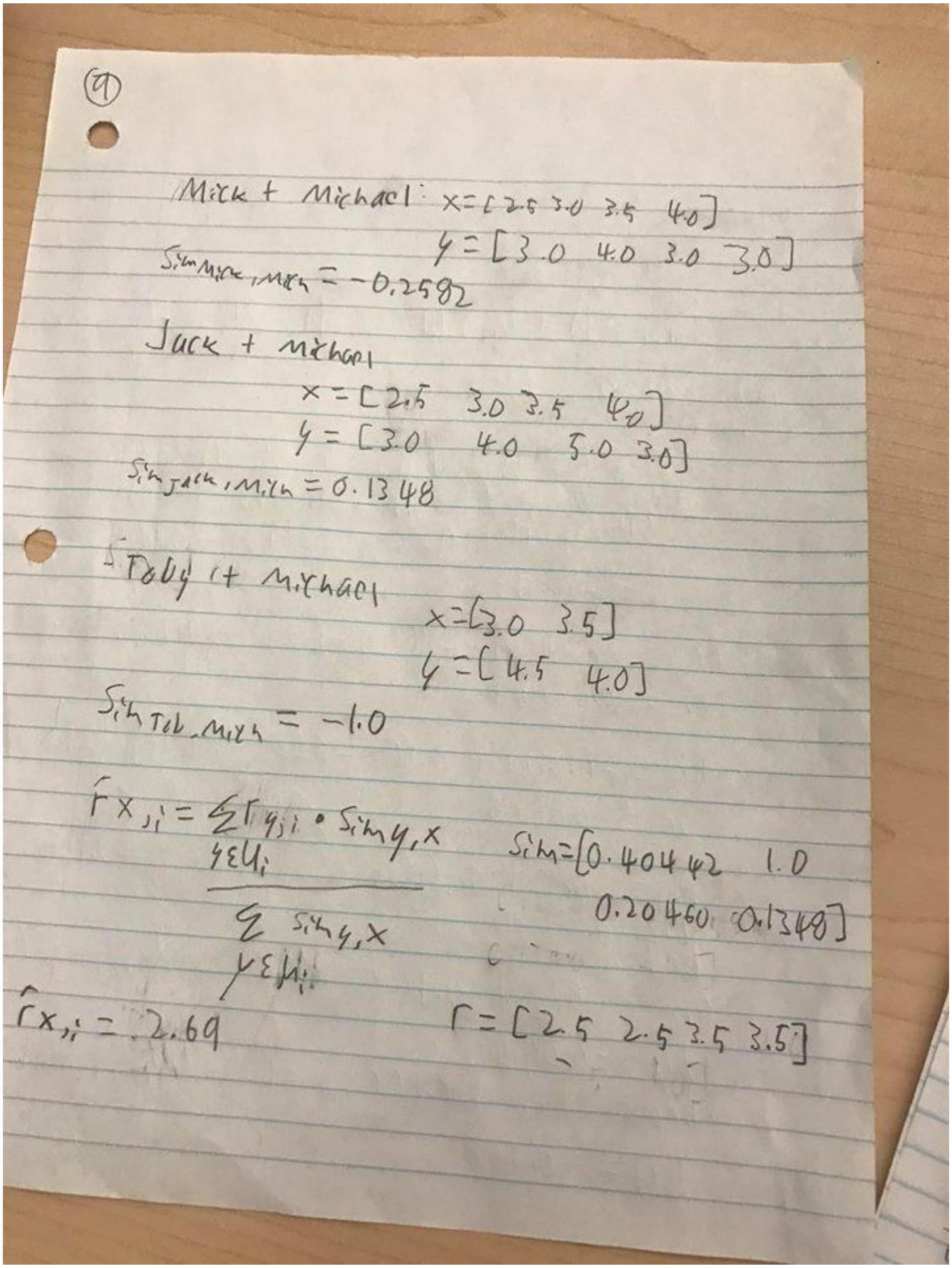
$br_4 = -0.02355$

$br_5 = br_4 + 0.1((2.5 - (3.23 + br_4)) - 0.1\cdot br_4)$

$br_5 = -0.0940$

$br_6 = br_5 + 0.1((3.0 - (3.23 + br_5)) - 0.1\cdot br_5)$

$br_6 = -0.1066$



(9)

Mick + Michael: $x = [2.5\ 3.0\ 3.5\ 4.0]$

$y = [3.0\ 4.0\ 3.0\ 3.0]$

$Sim_{Mick, Mich} = -0.2582$

Jack + Michael

$x = [2.5\ 3.0\ 3.5\ 4.0]$

$y = [3.0\ 4.0\ 5.0\ 3.0]$

$Sim_{Jack, Mich} = 0.1348$

Toby + Michael

$x = [3.0\ 3.5]$

$y = [4.5\ 4.0]$

$Sim_{Tob, Mich} = -1.0$

$$\hat{r}_{x,i} = \frac{\sum_{y \in U_i} r_{y,i} \cdot Sim_{y,x}}{\sum_{y \in U_i} Sim_{y,x}}$$

$Sim = [0.40442\ 1.0\ 0.20460\ 0.1348]$

$\hat{r}_{x,i} = 2.69$

$r = [2.5\ 2.5\ 3.5\ 3.5]$