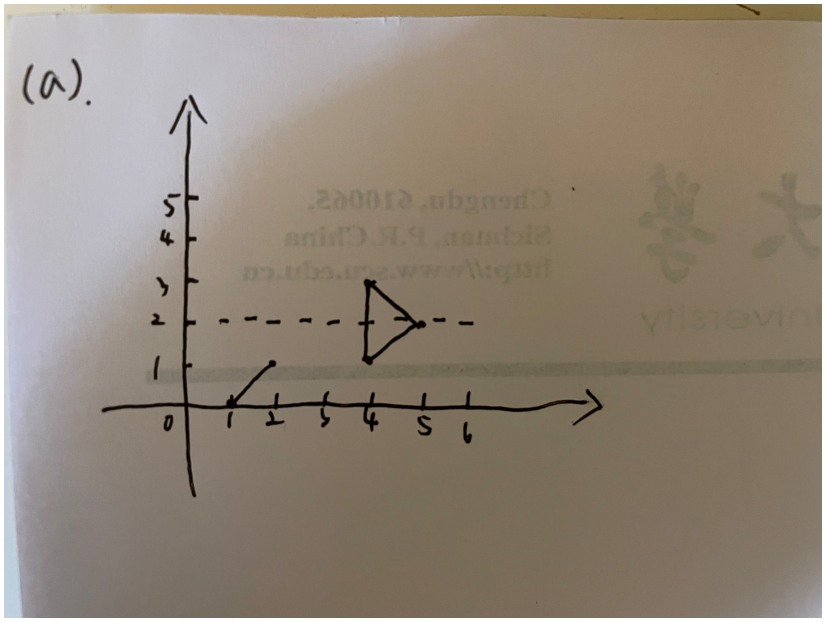
Name: Zhang Tong
Student ID: z5235242

Question 5
(a)



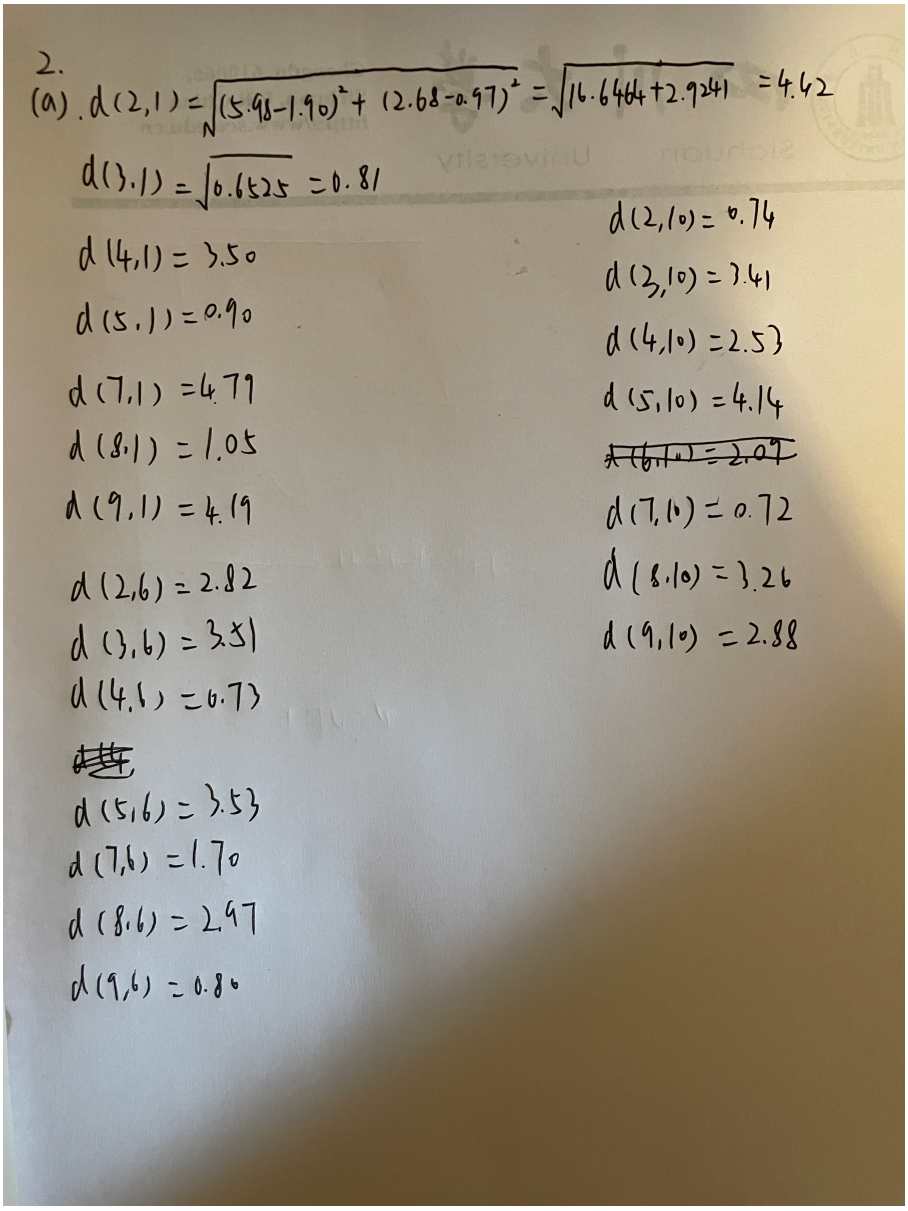
If the training examples are linearly separable, there would be no intersection of the convex hull of the positive examples and negative examples. By observing the picture above, it is obvious that the range of h is 0 <= h < 4.

(b)
If h = 1, the minimum distance of the positive examples and negative examples would be 2 (x5 and x1). Thus the functional margin of the training dataset would be 2/2 = 1.

(c)
We just need to ensure that the minimum distance of the positive examples and the negative examples is still 2.
According to the picture in (a), we have 0<=h<=2.

(d)
x5 and x1. If the other points are removed, the boundary would not change since the functional margin would not change. But if x1 or x5 is removed, the functional margin would change because the minimum distance of the positive and negative examples would change.

Question 2
(a)
Thus clusterA : {1, 3, 5, 8}



ClusterB: {4, 6, 9}
ClusterC; {2, 7, 10}

(b)
Center of A:
X1 = (1.9 + 2.68 + 1.54 + 2.46) / 4 = 2.145
X2 = (0.97 + 1.18 + 1.80 + 1.86) / 4 = 1.4525
(2.145, 1.4525)
Center of B:
X1 = (3.14 + 3.82 + 3.17) / 3 = 3.3767
X2 = (4.24 + 4.50 + 4.96) / 3 = 4.5667
(3.3767, 4.5667)
Center of C:
X1 = (5.98 + 5.74 + 5.44) / 3 = 5.72
X2 = (2.68 + 3.84 + 3.18) / 3 = 3.2333

(5.72, 3.2333)

(c)
Yes we should. Because that we use Manhattan distance, the lower distance does not always represents that the point is more close to the center.

Question 3
(a)

Q3.

(a) (P_1, P_2) has the lowest value (0.1)

So we have

	P_{12}	P_3	P_4	P_5
P_{12}	0	0.39	0.37	0.48
P_3		0	0.44	0.85
P_4			0	0.76
P_5				0

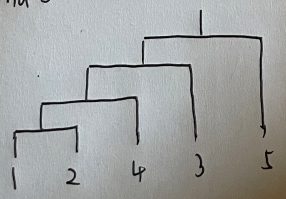
min = (P_{12}, P_4).

	P_{124}	P_3	P_5
P_{124}	0	0.44	0.54
P_3		0	0.85
P_5			0

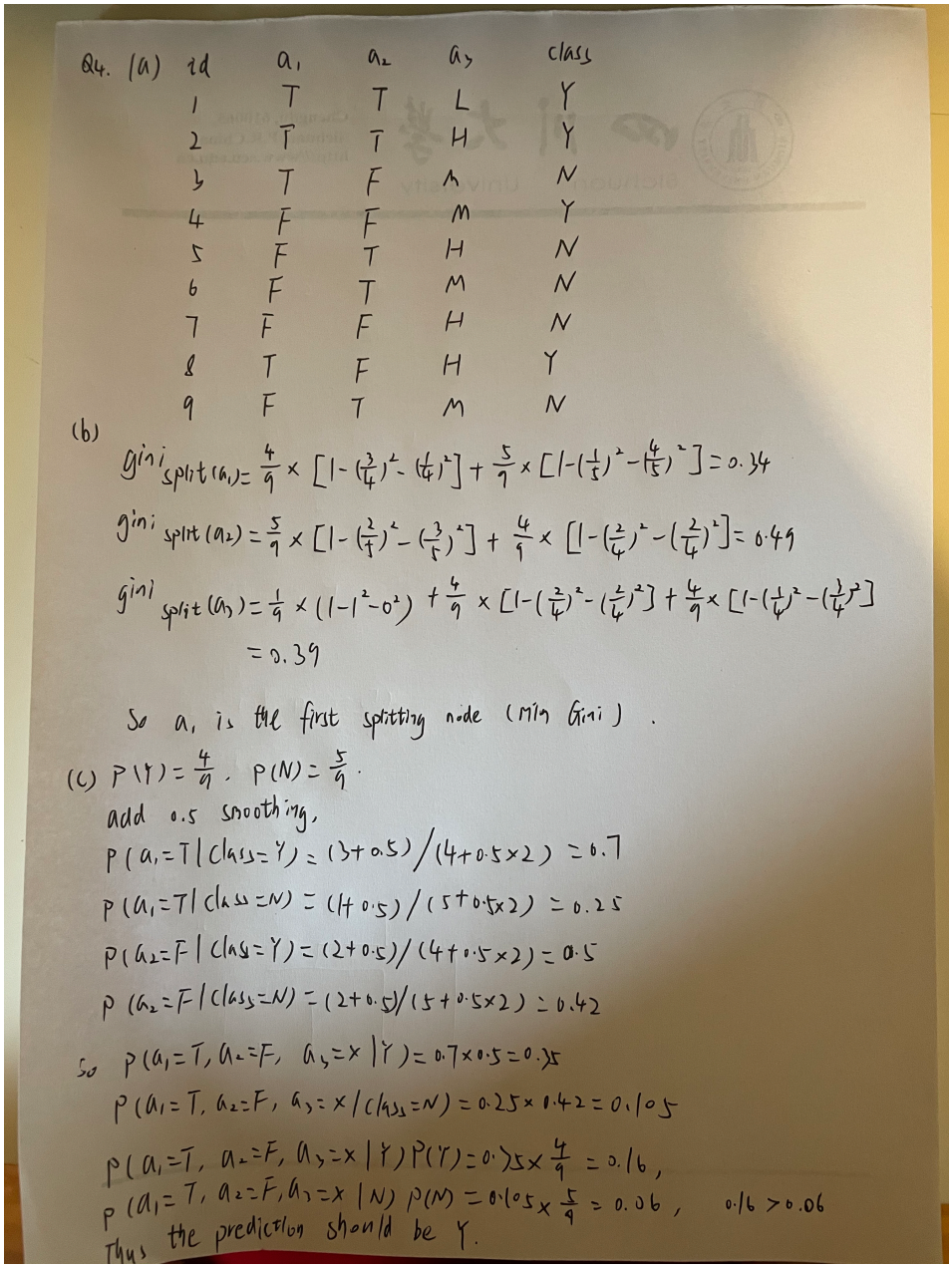
min = (P_{124}, P_3)

	P_{1243}	P_5
P_{1243}	0	0.56
P_5	0.56	0

Then we have



Question 4



Q4. (a)

id	a_1	a_2	a_3	class
1	T	T	L	Y
2	T	T	H	Y
3	T	F	M	N
4	F	F	M	Y
5	F	T	H	N
6	F	T	M	N
7	F	F	H	N
8	T	F	H	Y
9	F	T	M	N

(b)

$gini_{split}(a_1) = \frac{4}{9} \times [1-(\frac{3}{4})^2-(\frac{1}{4})^2] + \frac{5}{9} \times [1-(\frac{1}{5})^2-(\frac{4}{5})^2] = 0.34$

$gini_{split}(a_2) = \frac{5}{9} \times [1-(\frac{2}{5})^2-(\frac{3}{5})^2] + \frac{4}{9} \times [1-(\frac{2}{4})^2-(\frac{2}{4})^2] = 0.49$

$gini_{split}(a_3) = \frac{1}{9} \times (1-1^2-0^2) + \frac{4}{9} \times [1-(\frac{2}{4})^2-(\frac{2}{4})^2] + \frac{4}{9} \times [1-(\frac{1}{4})^2-(\frac{1}{4})^2]$

$= 0.39$

So a_1 is the first splitting node (min Gini).

(c) $P(Y) = \frac{4}{9}$, $P(N) = \frac{5}{9}$.

add 0.5 smoothing,

$P(a_1=T \mid class=Y) = (3+0.5)/(4+0.5\times2) = 0.7$

$P(a_1=T \mid class=N) = (1+0.5)/(5+0.5\times2) = 0.25$

$P(a_2=F \mid class=Y) = (2+0.5)/(4+0.5\times2) = 0.5$

$P(a_2=F \mid class=N) = (2+0.5)/(5+0.5\times2) = 0.42$

So $P(a_1=T, a_2=F, a_3=x \mid Y) = 0.7\times0.5 = 0.35$

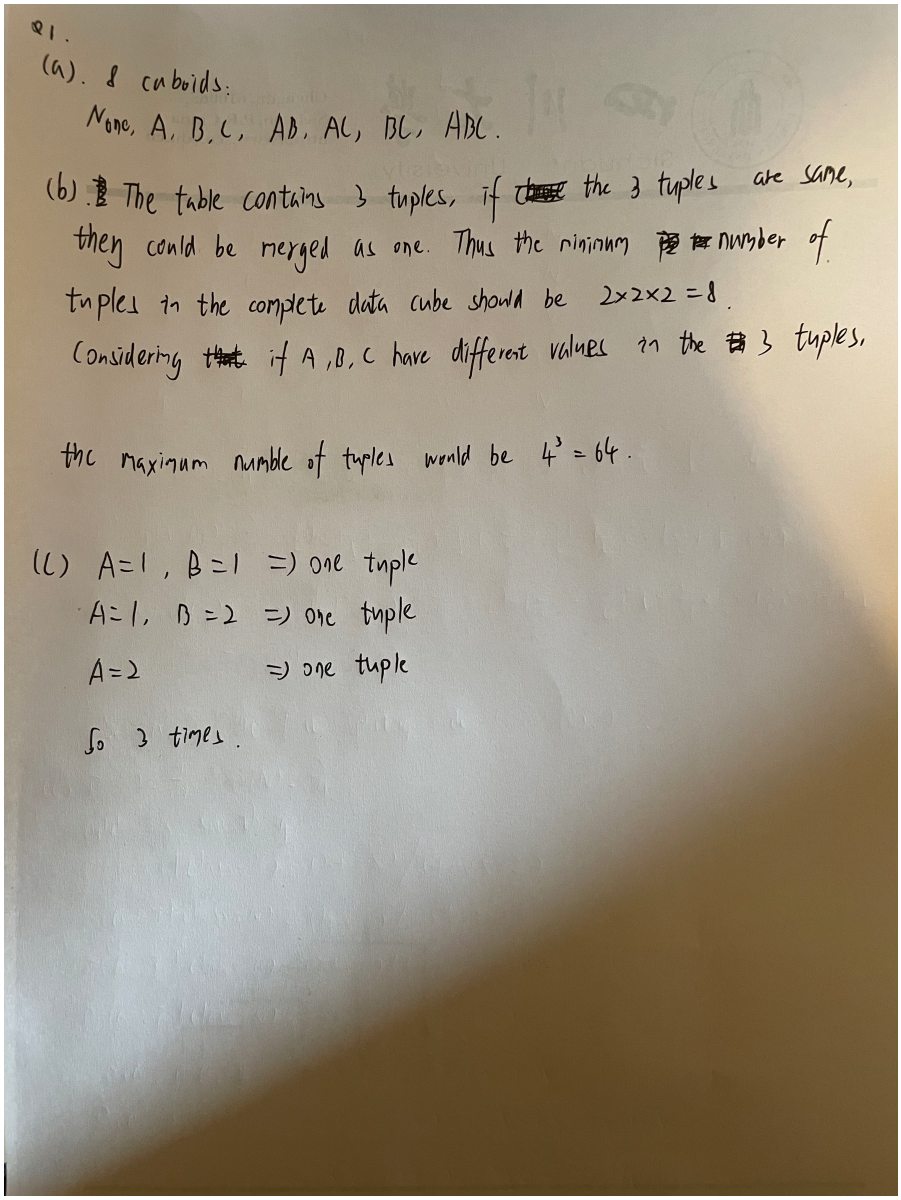
$P(a_1=T, a_2=F, a_3=x \mid class=N) = 0.25\times0.42 = 0.105$

$P(a_1=T, a_2=F, a_3=x \mid Y)P(Y) = 0.35\times\frac{4}{9} = 0.16$,

$P(a_1=T, a_2=F, a_3=x \mid N)P(N) = 0.105\times\frac{5}{9} = 0.06$, $0.16 > 0.06$

Thus the prediction should be Y.

Question 1



Q1.

(a). 8 cuboids:

None, A, B, C, AB, AC, BC, ABC.

(b). The table contains 3 tuples, if the 3 tuples are same, they could be merged as one. Thus the minimum number of tuples in the complete data cube should be $2 \times 2 \times 2 = 8$.

Considering if A, B, C have different values in the 3 tuples, the maximum number of tuples would be $4^3 = 64$.

(c) A=1, B=1 => one tuple

A=1, B=2 => one tuple

A=2 => one tuple

So 3 times.