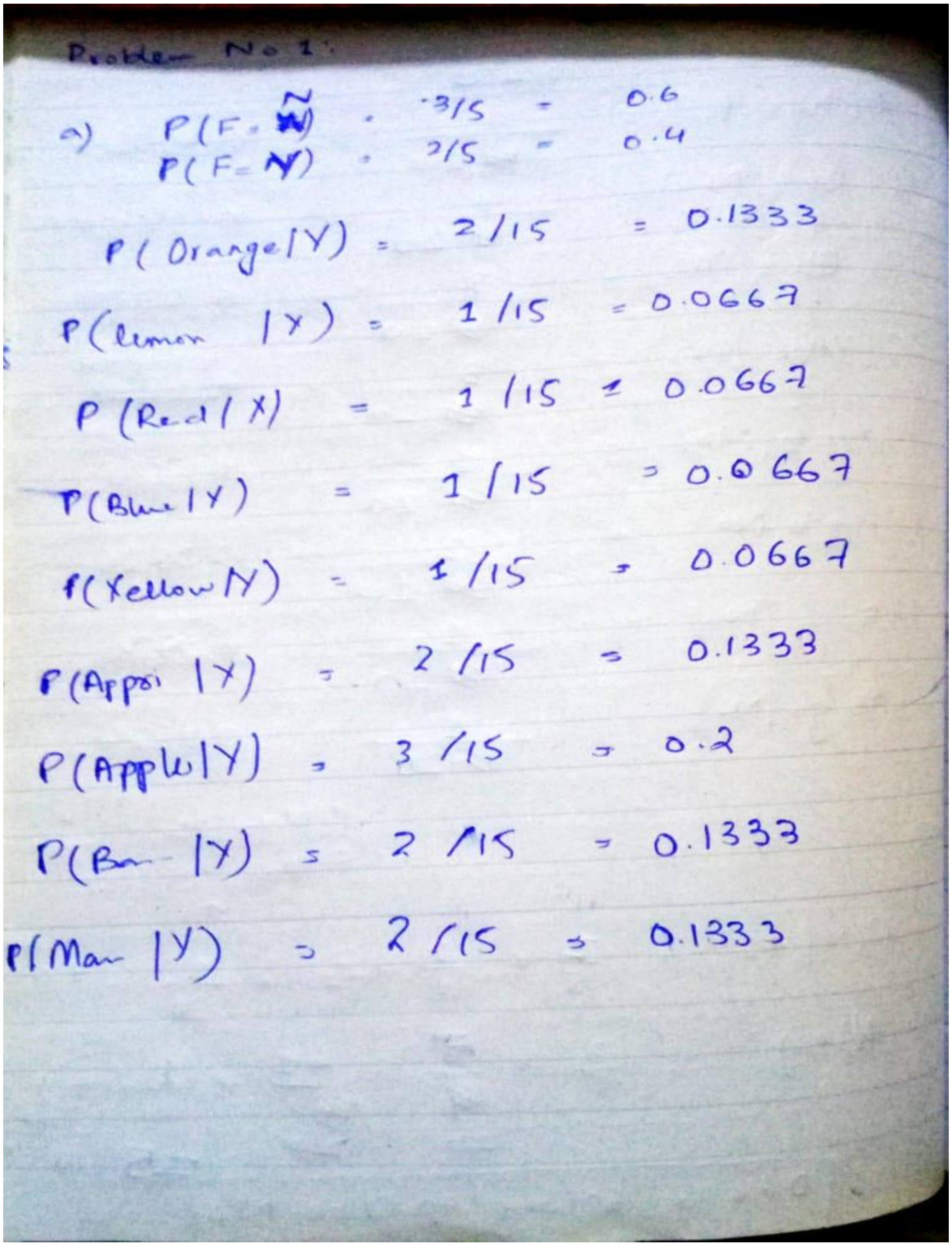
IR CR-2

K173795 Muhammad Mustafa Manga

Problem 1:



Problem No 1:

a) P(F = ~~W~~ N) = 3/5 = 0.6
P(F = Y) = 2/5 = 0.4

P(Orange|Y) = 2/15 = 0.1333

P(lemon |Y) = 1/15 = 0.0667

P (Red| X) = 1/15 = 0.0667

P(Blue|Y) = 1/15 = 0.0667

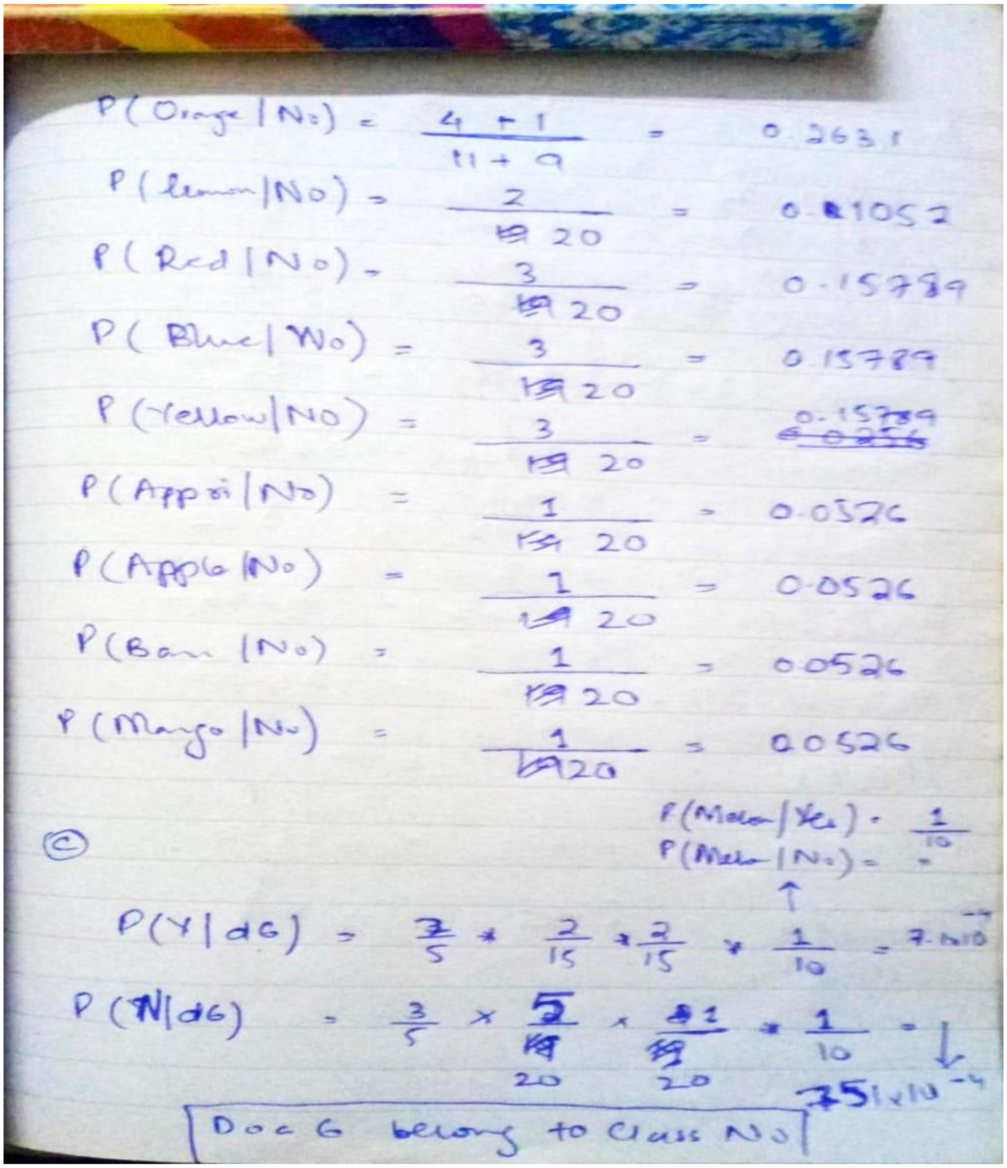
P(Yellow|Y) = 1/15 = 0.0667

P(Appoi |Y) = 2/15 = 0.1333

P(Apple|Y) = 3/15 = 0.2

P(Ban |Y) = 2/15 = 0.1333

P(Man |Y) = 2/15 = 0.1333



$P(\text{Orange}|\text{No}) = \frac{4+1}{11+9} = 0.2631$

$P(\text{lemon}|\text{No}) = \frac{2}{\cancel{19}\ 20} = 0.\cancel{2}1052$

$P(\text{Red}|\text{No}) = \frac{3}{\cancel{19}\ 20} = 0.15789$

$P(\text{Blue}|\text{No}) = \frac{3}{\cancel{19}\ 20} = 0.15789$

$P(\text{Yellow}|\text{No}) = \frac{3}{\cancel{19}\ 20} = 0.15789$ ~~0.0256~~

$P(\text{Appri}|\text{No}) = \frac{1}{\cancel{19}\ 20} = 0.0526$

$P(\text{Apple}|\text{No}) = \frac{1}{\cancel{19}\ 20} = 0.0526$

$P(\text{Ban}|\text{No}) = \frac{1}{\cancel{19}\ 20} = 0.0526$

$P(\text{Mango}|\text{No}) = \frac{1}{\cancel{19}\ 20} = 0.0526$

ⓒ

$P(\text{Melon}|\text{Yes}) = \frac{1}{10}$

$P(\text{Melon}|\text{No}) = "$

↑

$P(Y|d6) = \frac{2}{5} * \frac{2}{15} * \frac{2}{15} * \frac{1}{10} = 7.1 \times 10^{-4}$

$P(N|d6) = \frac{3}{5} \times \frac{5}{\cancel{19}\ 20} \times \frac{\cancel{4}1}{\cancel{19}\ 20} * \frac{1}{10} = \downarrow$

7.5×10^{-4}

Doc 6 belong to Class No

Correction:

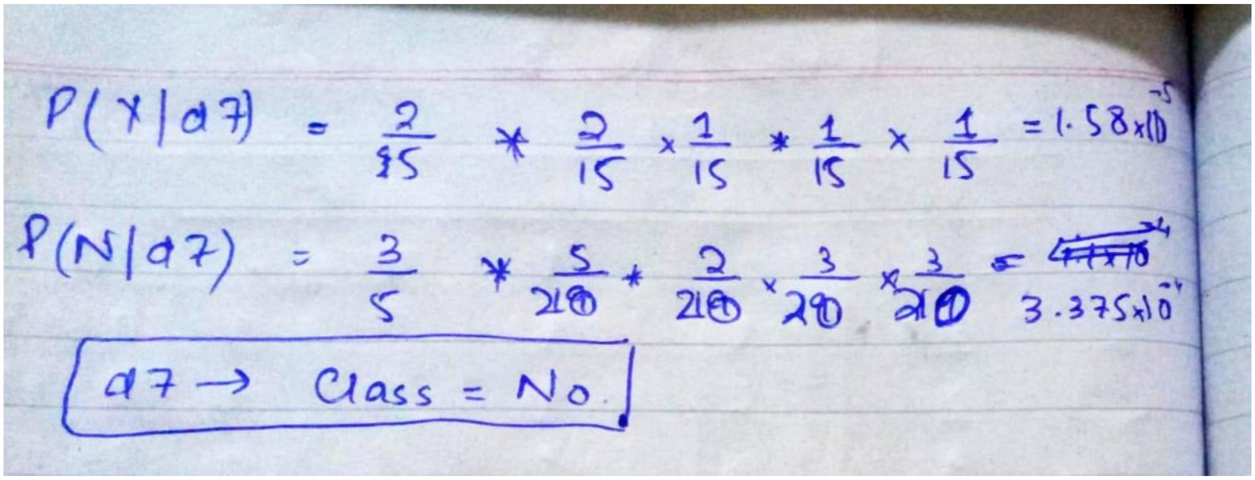
For P (Mellon | Yes): 0 + 1 / (8 + 1) (9 + 1) = 1/17

For P (Mellon | Yes): 0 + 1 / (11 + 1) (9 + 1) = 1/22

P (Yes | doc6): 2/5 * 2/15 * 2/15 * 1/17 = 0.0004183

P (Yes | doc6): 3/5 * 5/20 * 1/20 * 1/22 = 0.00034

Doc 6 belongs to Class Yes



$$P(Y|a7) = \frac{2}{15} * \frac{2}{15} \times \frac{1}{15} * \frac{1}{15} \times \frac{1}{15} = 1.58\times10^{-5}$$

$$P(N|a7) = \frac{3}{5} * \frac{5}{20} * \frac{2}{20} \times \frac{3}{20} \times \frac{3}{20} = \cancel{4.1\times10^{-4}}\ 3.375\times10^{-4}$$

$$\boxed{a7 \rightarrow \text{Class} = \text{No.}}$$

Problem 2:

A) Concept drift refers to the gradual change over time of the concept underlying a class. The concept drift means that the statistical properties of the target variable, which the model is trying to predict, change over time in unpredicted ways. This causes problems because the predictions become less accurate as time passes. The Bernoulli model is robust with respect to concept drift. It gives good performance when the feature size is low. The most important indicators for a class are less likely to change. Thus, a model that only relies on these features is more likely to maintain a certain level of accuracy in concept drift.

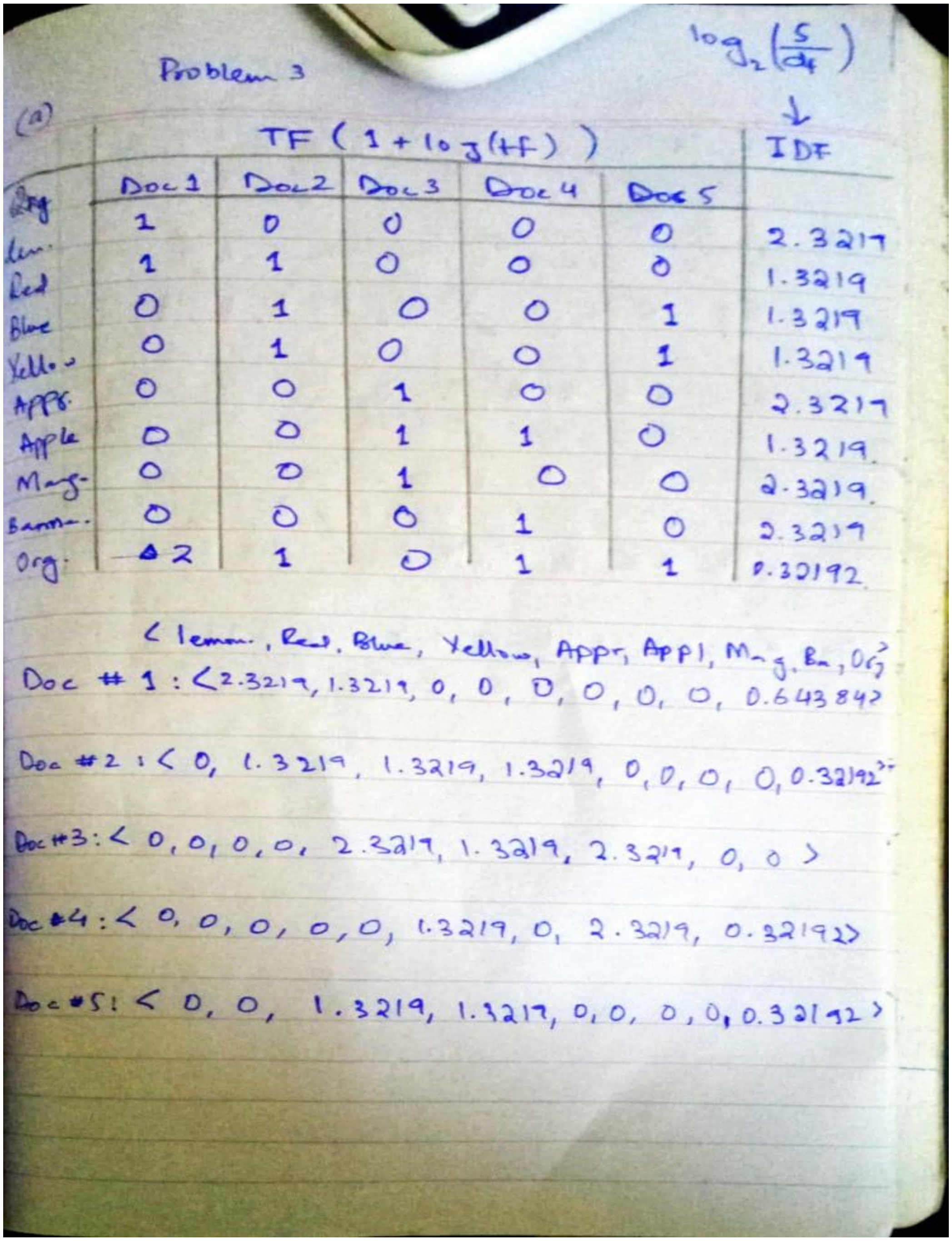
B) The NB classifier are highly biased towards a particular class and hence other class have very low probability. The NB Classifier makes the conditional independence assumption to reduce the number of parameters. That is, the presence of one particular feature (term) does not affect others. This assumption is very naive for a model of natural language. That's why the NB classifier is called ‘naive’.

C) The strategy NB classifier use to handle out of training set vocabulary for test instances is Laplace correction by adding artificial observation of every feature for every class. This is also known as uniform prior. Due to this we avoid 0 propagations and do effect of this feature negligible.

D) The above three docs have identical bag of word for Bernoulli model because model focus on doc containing term and consider binary model and don’t focus on multi occurrence.

The above D1, D2 have identical bag of word for Multinomial model because the term w1 is occurs twice in D1, D2 and D3 is different because w1occur only once.

Problem 3:



Problem 3

$\log_2\left(\frac{5}{d_f}\right)$

↓

(a)

	TF (1 + log(tf))					IDF
~~Org~~	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	
lem.	1	0	0	0	0	2.3217
Red	1	1	0	0	0	1.3219
Blue	0	1	0	0	1	1.3219
Yellow	0	1	0	0	1	1.3219
Appr.	0	0	1	0	0	2.3217
Apple	0	0	1	1	0	1.3219.
Mang.	0	0	1	0	0	2.3219.
Banana.	0	0	0	1	0	2.3219
Org.	2	1	0	1	1	0.32192

< lemon, Red, Blue, Yellow, Appr, Appl, Mang, Ba, Org >

Doc # 1 : <2.3219, 1.3219, 0, 0, 0, 0, 0, 0, 0.64384>

Doc #2 : < 0, 1.3219, 1.3219, 1.3219, 0, 0, 0, 0, 0.32192>

Doc #3 : < 0, 0, 0, 0, 2.3217, 1.3219, 2.3217, 0, 0 >

Doc #4 : < 0, 0, 0, 0, 0, 1.3219, 0, 2.3219, 0.32192>

Doc #5 : < 0, 0, 1.3219, 1.3217, 0, 0, 0, 0, 0.32192 >

(b): $\text{Centroid}_{F=Y} = = \frac{1}{|\text{Class}_{D=Y}|} \sum_{d \in c} \vec{x}$

$\mu_{F=Y} = \left(\frac{1}{2}\right) * \left(d_3 \overset{+}{\cancel{*}} d_4\right)$

$= \frac{1}{2}\Big(\langle 0, 0, 0, 0, 2.3219, 1.3219, 2.3219, 0, 0 \rangle$
$+ \langle 0, 0, 0, 0, 0, 1.3219, 0, 2.3219, 0.32192 \rangle$

$= \langle 0, 0, 0, 0, 1.16095, 1.3219, 1.16095, \blacksquare, 1.16095$
$0.16096 \rangle$

$u_{F=N} = \frac{1}{3} * (d_1 + d_2 + d_5)$

$u_{F=N} = \langle 0.773, 0.8812, 0.8812, 0.8812, 0, 0$
$0, 0, 0.4292 \rangle$

~~Prob~~ (c)

For Doc 6

<Lem, Red, Blu, Yell, Appr, Appl, Man, Ban, Org, Melon

Doc#6 : $\langle 0, 0, 0, 0, 0, 0, 2.3219, 0, 0.321923$
$, 0 \rangle$.

$\left|u_{F=No} \overset{d_6}{-}\right| = \sqrt{(0.773-0)^2 + (0.8812-0)^2 + (0.88128-0)^2}$
$(0.8812-0)^2 + 0 + 0 + (0-2.3219)^2$
$0 + (0.4292-0.32192)^2 + 0$

$|u_{F=N} - d_6| = 2.88$

$$|u_{F=Yes} - d_6| = \sqrt{0+0+0+0+(1.16095)^2 + (1.3219)^2 + (1.16095 - 2.3219)^2 + (1.16095)^2 + (0.16096 - 0.32192)^2 + 0}$$

$$|u_{F=Yes} - d_6| = 2.41$$

Doc 6 → Class fruit Yes

For Doc 7:

Doc #7: <2.3219, 1.3219, 0, 1.3219, 0, 0, 0
0, 0.32192, 0, >

$$|u_{F=No} - d_7| = \sqrt{(0.773 - 2.3219)^2 + (0.8812 - 1.3219)^2 (0.8812)^2 + (0.8812 - 1.3219)^2 . 0 + 0 + 0 + 0 + (0.4292 - 0.32192)^2 + 0}$$

$$(u_{F=No} - d_7) = 1.83$$

$$|u_{F=yes} - d_7| = \sqrt{(2.3219)^2 + (1.3219)^2 + 0 + (1.3219)^2 + (1.1609)^2 + (1.3217)^2 (1.16095)^2 + (1.16095)^2 + (0.16096 - 0.32192)^2 .}$$

$$= 3.83.$$

Hence D_7 → Class fruit = No.

(D) KNN for Doc.6:
<0,0,0,0,0,0, 2.3219, 0, 0.32192, 0>

$\cos(d_1, d_6) = \text{dot prod}(d_1, d_6) \div |d_1||d_6|$

$= 0.20608 \div [(2.745)(2.34198)]$

$\cos(d_1, d_6) = 0.0320.$

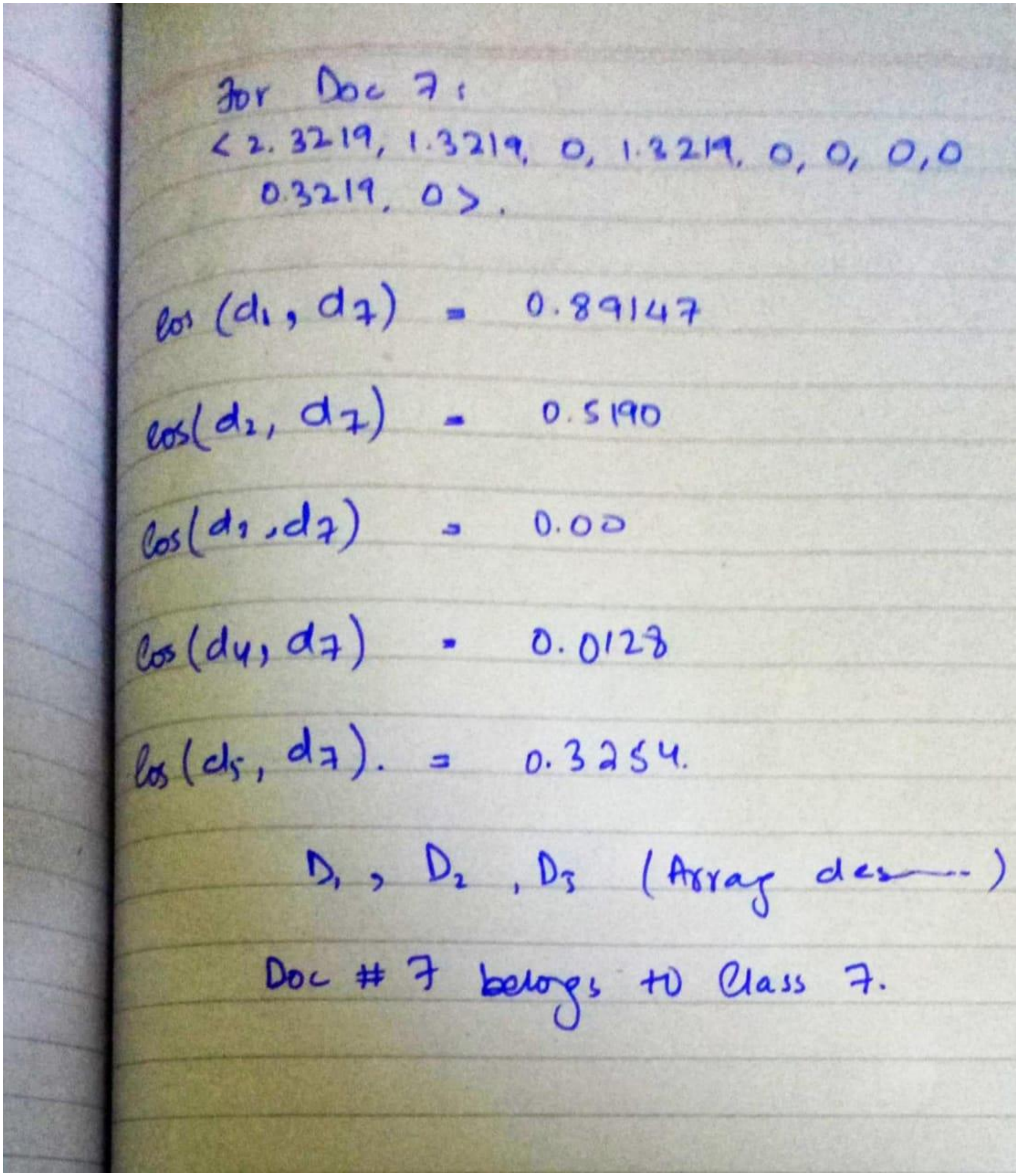
$\cos(d_2, d_6) = 0.01836$

$\cos(d_3, d_6) = 0.6498$

$\cos(d_4, d_6) = 0.01626$

$\cos(d_5, d_6) = 0.02308$

E D3, D1, D5 (Arrang in des — -)
(Doc 6 Belong to No Class)



For Doc 7 :

< 2.3219, 1.3219, 0, 1.3219, 0, 0, 0,0
0.3219, 0 >.

$\cos(d_1, d_7) = 0.89147$

$\cos(d_2, d_7) = 0.5190$

$\cos(d_3, d_7) = 0.00$

$\cos(d_4, d_7) = 0.0128$

$\cos(d_5, d_7). = 0.3254.$

D_1 , D_2 , D_3 (Array des—--)

Doc # 7 belogs to Class 7.