Homework 2

COSC 522: Machine Learning

Harshvardhan
Student Id: 609162

15 September 2021

① (a) For class I:

X	Y
0.8	1.2
0.9	1.4
1.2	1.4
1.1	1.5

So, $\underset{\sim}{\bar{X}_1} = \begin{bmatrix} \bar{X} \\ \bar{Y} \end{bmatrix} = \begin{bmatrix} 1 \\ 1.375 \end{bmatrix}$ [Mean matrix]

$$\bar{X}_1 = \frac{0.8 + 0.9 + 1.2 + 1.1}{4} = 1.$$

$$\bar{Y}_1 = \frac{1.2 + 1.4 + 1.4 + 1.5}{4} = 1.375$$

- $\mathrm{Var}(X_1) = \frac{1}{n-1}\left\{(0.8-1)^2 + (0.9-1)^2 + (1.2-1)^2 + (1.1-1)^2\right\}$

$$= \frac{1}{3}\left\{0.04 + 0.01 + 0.04 + 0.01\right\}$$

$$= \frac{1}{3} \times 0.1 = 0.033.$$

- $\mathrm{Var}(Y_1) = \frac{1}{3}\left\{(1.2-1.375)^2 + (1.4-1.375)^2 + (1.4-1.375)^2 + (1.5-1.375)^2\right\}$

$$= \frac{1}{3}\left(0.0306 + 0.0006 + 0.0006 + 0.0156\right)$$

$$= 0.0158.$$

$$\mathrm{Cov}(X_1, Y_1) = \frac{1}{3}\Big\{(0.8-1)(1.2-1.375) + (0.9-1)(1.4-1.375)$$

$$+ (1.2-1)(1.4-1.375) + (1.1-1)(1.5-1.375)\Big\}$$

$$= \frac{1}{3}\Big\{0.2 \times 0.175 \overset{-}{\cancel{+}} 0.1 \times 0.025 + 0.2 \times 0.025 + 0.1 \times 0.125\Big\}$$

$$= \frac{0.05}{3} = 0.016$$

$$\text{So, } \mathrm{Cov}(\underset{\sim}{X}_1) = \cancel{\begin{bmatrix} 0.33 & 0.0183 \\ 0.0183 & 0.0158 \end{bmatrix}} \begin{bmatrix} 0.33 & 0.016 \\ 0.016 & 0.0158 \end{bmatrix}$$

<u>for Class 2:</u>

$$\bar{X}_2 = \frac{0.8 + 0.6 + 0.65 + 0.75}{4}$$

$$= 0.7$$

$$\bar{Y}_2 = \frac{1.1 + 1 + 1.1 + 0.9}{4} = \frac{4.2}{4} = 1.025$$

$$\mathrm{Var}(X_2) = \frac{(0.7-0.8)^2 + (0.6-0.7)^2 + (0.65-0.7)^2 + (0.75-0.7)^2}{4-1}$$

$$= 0.00833$$

$$\mathrm{Var}(Y_2) = \frac{(1.1-1.025)^2 + (1-1.025)^2 + (1.1-1.025)^2 + (0.9-1.025)^2}{4-1}$$

$$= \cancel{\frac{(0.05)^2 + (0.05)^2 + (0.025)^2 + (0.125)^2}{3}}$$

$$= 0.00916 \quad \text{[Calculated using calculator]}$$

$$\mathrm{Cor}(X_2, Y_2) = \frac{(0.8-0.7)(0.075) + 0.1 \times 0.025 + 0.05 \times 0.075 - 0.05 \times 0.125}{4-1}$$

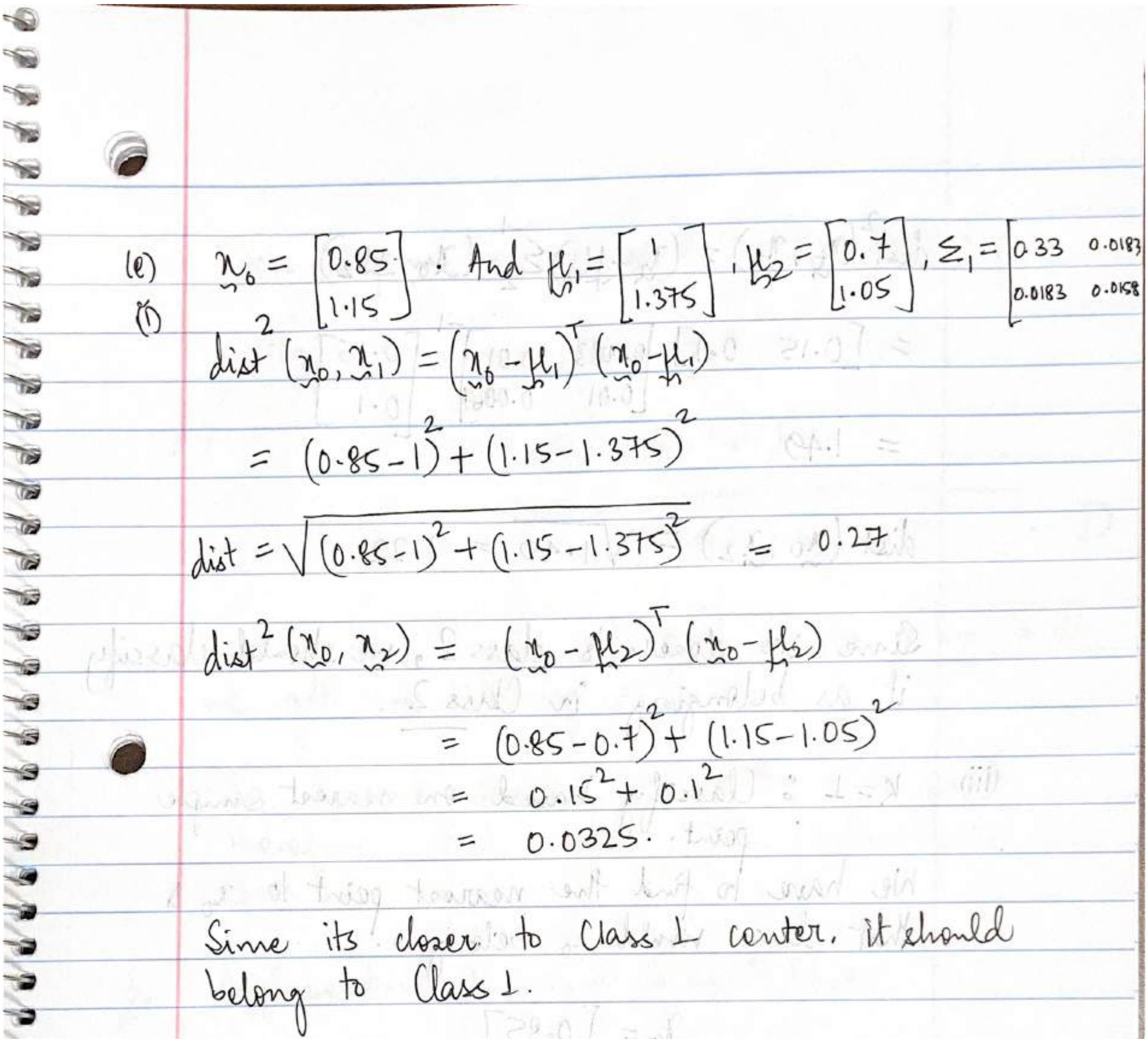
$$= 0/3 = 0.$$

$$\therefore \mathrm{Cor}(\underset{\sim}{X}_2) = \cancel{\begin{bmatrix} 0.0083 & 0.01 \\ 0.01 & 0.0067 \end{bmatrix}} \begin{bmatrix} 0.0083 & 0 \\ 0 & 0.0091 \end{bmatrix}$$

(c) Mahalanobis Distance $= \left((\underset{\sim}{x} - \underset{\sim}{\mu})^T \Sigma^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu})\right)^{1/2}$

Euclidean Distance $= \left((\underset{\sim}{x} - \underset{\sim}{\mu})^T (\underset{\sim}{x} - \underset{\sim}{\mu})\right)^{1/2}$.

(d) Mahalanobis distance accounts for the correlation between the variables. It assumes the correlatation to be same between classes. Euclidean distance assumes that the features are independent of each other and have equal variance throughout



(e) (i) $\underset{\sim}{x}_0 = \begin{bmatrix} 0.85 \\ 1.15 \end{bmatrix}$. And $\underset{\sim}{\mu}_1 = \begin{bmatrix} 1 \\ 1.375 \end{bmatrix}$, $\underset{\sim}{\mu}_2 = \begin{bmatrix} 0.7 \\ 1.05 \end{bmatrix}$, $\Sigma_1 = \begin{bmatrix} 0.33 & 0.0183 \\ 0.0183 & 0.0158 \end{bmatrix}$

$$\text{dist}^2(\underset{\sim}{x}_0, \underset{\sim}{x}_1) = (\underset{\sim}{x}_0 - \underset{\sim}{\mu}_1)^T (\underset{\sim}{x}_0 - \underset{\sim}{\mu}_1)$$

$$= (0.85-1)^2 + (1.15-1.375)^2$$

$$\text{dist} = \sqrt{(0.85-1)^2 + (1.15-1.375)^2} = 0.27$$

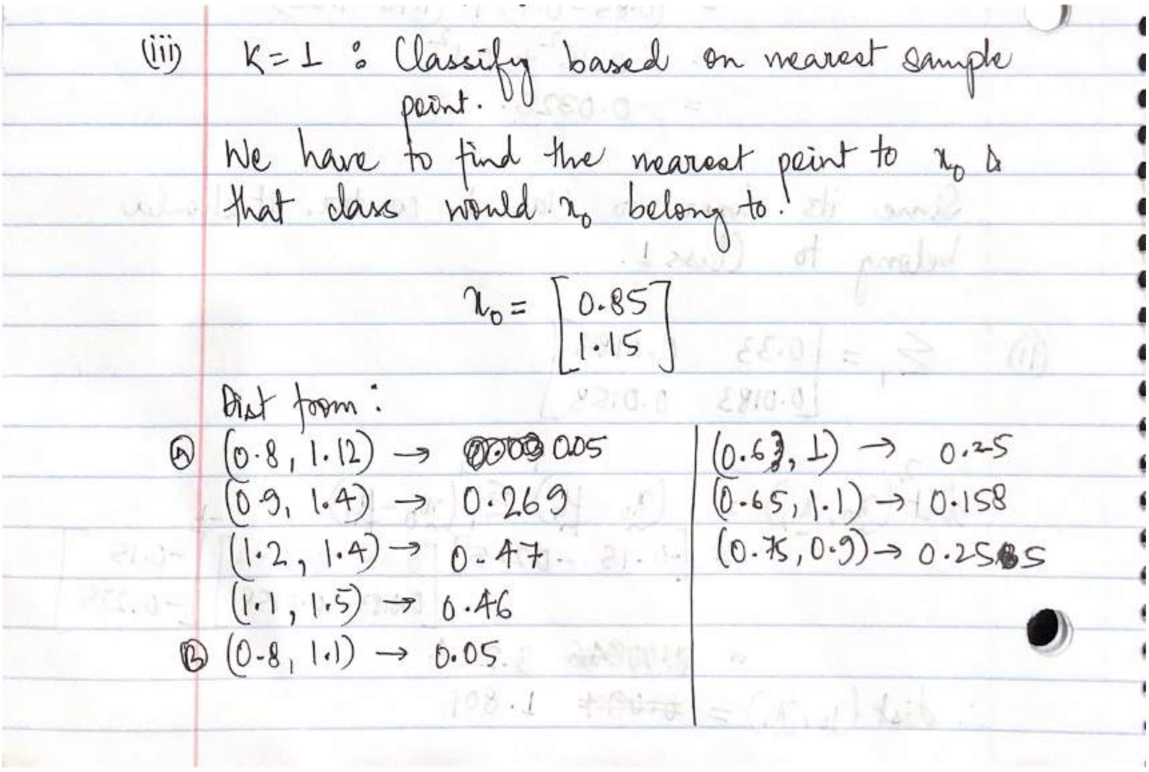
$$\text{dist}^2(\underset{\sim}{x}_0, \underset{\sim}{x}_2) = (\underset{\sim}{x}_0 - \underset{\sim}{\mu}_2)^T (\underset{\sim}{x}_0 - \underset{\sim}{\mu}_2)$$

$$= (0.85-0.7)^2 + (1.15-1.05)^2$$

$$= 0.15^2 + 0.1^2$$

$$= 0.0325.$$

Since its closer to Class 1 center, it should belong to Class 1.



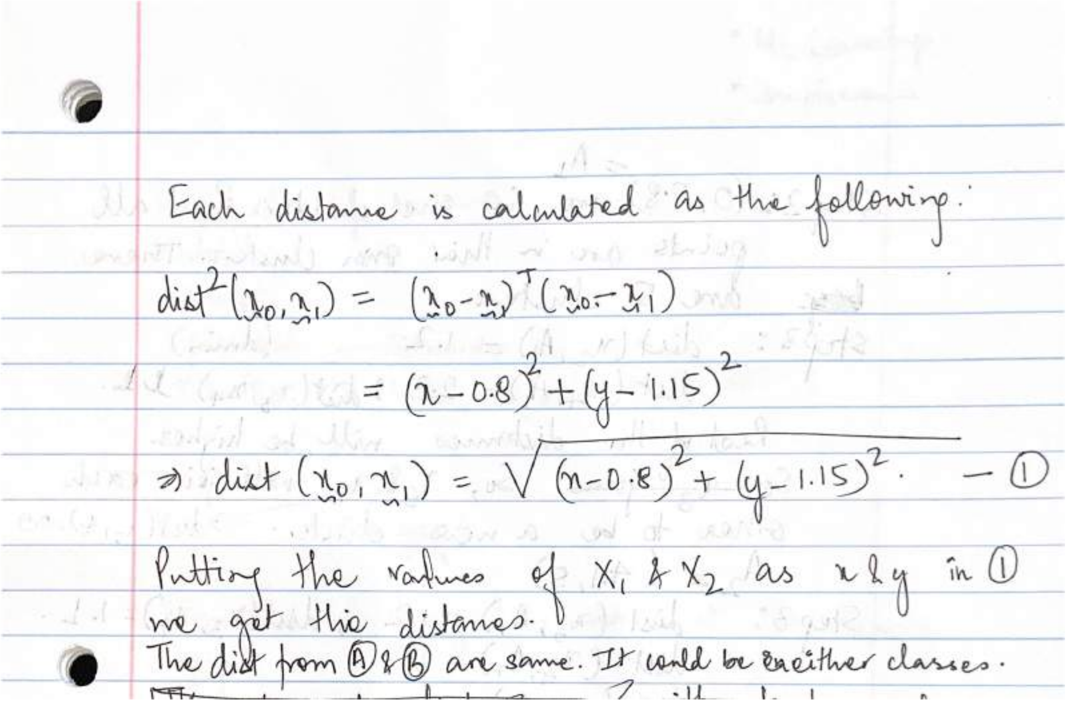
(iii) K = 1 : Classify based on nearest sample point.

We have to find the nearest point to x_0 & that class would x_0 belong to.

$$x_0 = \begin{bmatrix} 0.85 \\ 1.15 \end{bmatrix}$$

Dist from :

Ⓐ	$(0.8, 1.12)$	→ 0.05
	$(0.9, 1.4)$	→ 0.269
	$(1.2, 1.4)$	→ 0.47
	$(1.1, 1.5)$	→ 0.46
Ⓑ	$(0.8, 1.1)$	→ 0.05
	$(0.63, 1)$	→ 0.25
	$(0.65, 1.1)$	→ 0.158
	$(0.75, 0.9)$	→ 0.255



Each distance is calculated as the following:

$$\text{dist}^2(\underset{\sim}{x}_0, \underset{\sim}{x}_1) = (\underset{\sim}{x}_0 - \underset{\sim}{x}_1)^T (\underset{\sim}{x}_0 - \underset{\sim}{x}_1)$$

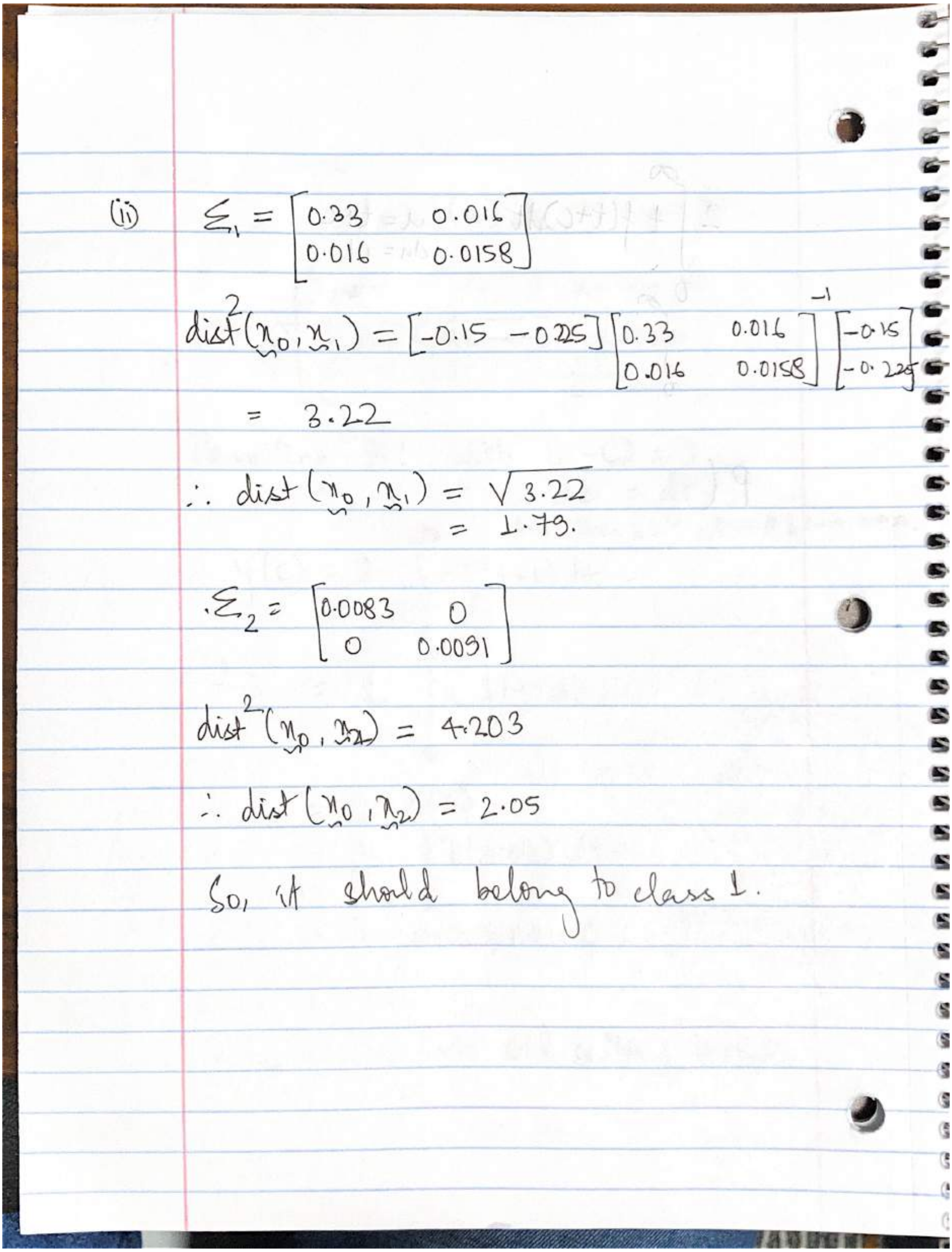
$$= (x - 0.8)^2 + (y - 1.15)^2$$

$$\Rightarrow \text{dist}(\underset{\sim}{x}_0, \underset{\sim}{x}_1) = \sqrt{(x-0.8)^2 + (y-1.15)^2}. \quad — ①$$

Putting the values of X_1 & X_2 as x & y in ① we get the distances.

The dist from Ⓐ & Ⓑ are same. It could be eaither classes.

No, kNN with k=1 is not equivalent to minimum distance classifier.



(ii) $\Sigma_1 = \begin{bmatrix} 0.33 & 0.016 \\ 0.016 & 0.0158 \end{bmatrix}$

$$dist^2(\underset{\sim}{x}_0, \underset{\sim}{x}_1) = \begin{bmatrix} -0.15 & -0.225 \end{bmatrix} \begin{bmatrix} 0.33 & 0.016 \\ 0.016 & 0.0158 \end{bmatrix}^{-1} \begin{bmatrix} -0.15 \\ -0.225 \end{bmatrix}$$

$= 3.22$

$\therefore dist(\underset{\sim}{x}_0, \underset{\sim}{x}_1) = \sqrt{3.22}$
$= 1.79.$

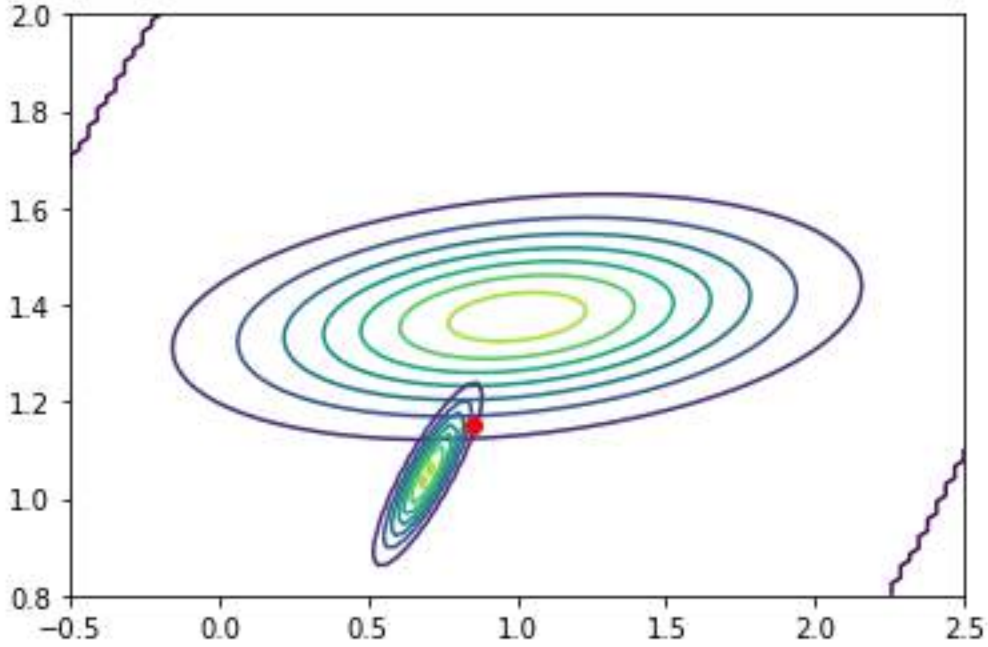
$\Sigma_2 = \begin{bmatrix} 0.0083 & 0 \\ 0 & 0.0091 \end{bmatrix}$

$dist^2(\underset{\sim}{x}_0, \underset{\sim}{x}_2) = 4.203$

$\therefore dist(\underset{\sim}{x}_0, \underset{\sim}{x}_2) = 2.05$

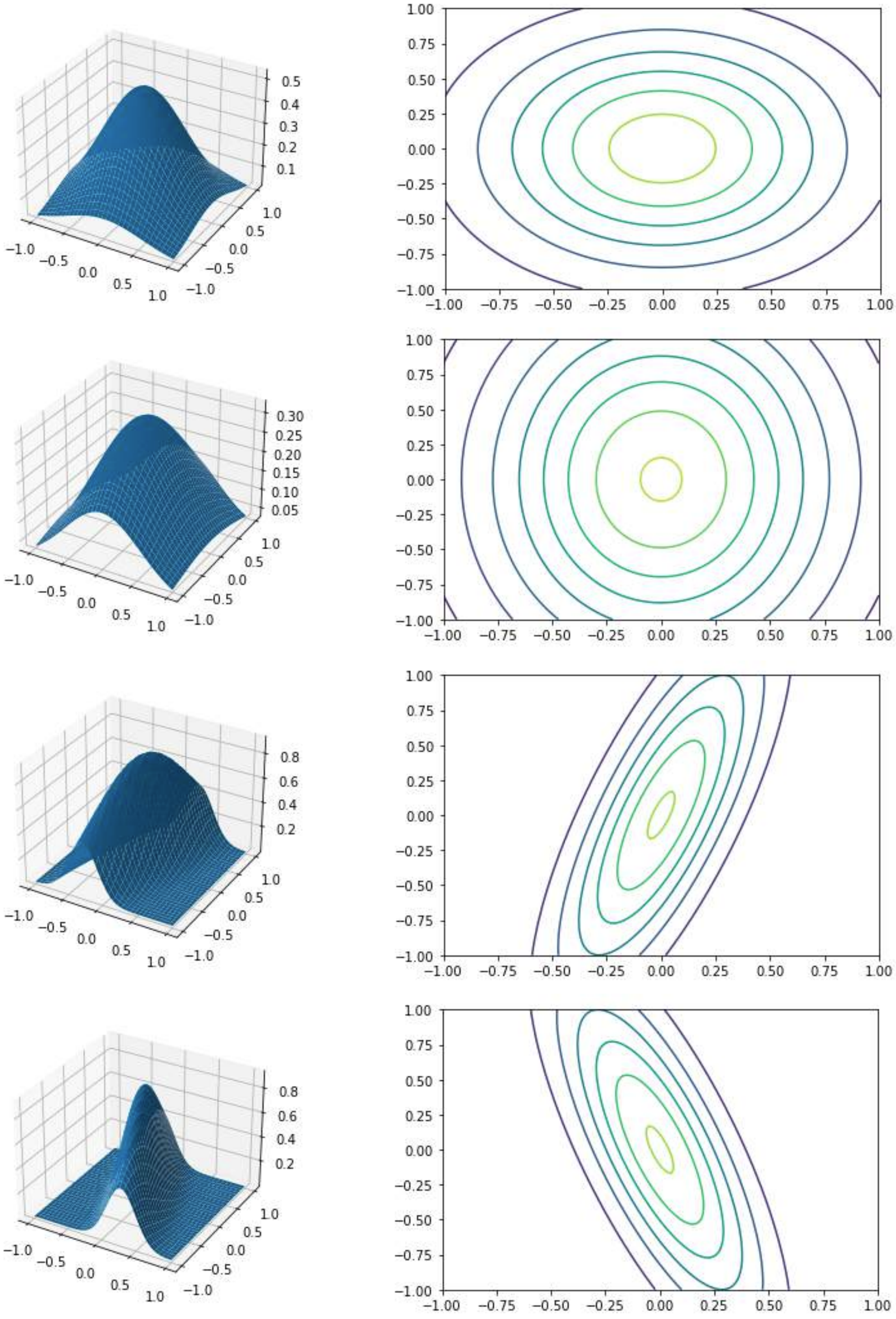
So, it should belong to class 1.

Question 1 (b)



From the plot, it looks like the point belongs in between the last two contours for both the curves, i.e. Gaussian curves obtained for both the classes. Thus, it could belong in either of the classes. However, for accurate judgement, we should only use the analytical method and not rely on visual inspection method.

Question 2


0.5
0.4
0.3
0.2
0.1
−1.0
−0.5
0.0
0.5
1.0
1.0
0.5
0.0
−0.5
−1.0
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
0.30
0.25
0.20
0.15
0.10
0.05
0.8
0.6
0.4
0.2

I have assumed the following matrices:

$A = \begin{bmatrix} 0.3 & 0 \\ 0 & 0.3 \end{bmatrix}$, $B = \begin{bmatrix} 0.3 & 0 \\ 0 & 0.8 \end{bmatrix}$, $C = \begin{bmatrix} 0.1 & 0.2 \\ 0.2 & 0.7 \end{bmatrix}$ and $D = \begin{bmatrix} 0.1 & -0.2 \\ -0.2 & 0.7 \end{bmatrix}$.

Insights on each element:

1. Diagonal elements decide the width and height of ellipse of the contour plot. They decide the stretch of the normal curve in different directions.
2. The variance decides what will be the stretch of the normal curve. For example, if Var(X1) > Var(X2), then the curve will be an ellipse with the major axis along X1.
3. Consequently, if the diagonal elements are equal, they would result in circular contour plots and thus normal curve with circular cross-sections.
4. The off-diagonal elements decide the tilt of Gaussian curves. The higher the covariance, greater will be the tilt.
5. The absolute value of covariance decides the magnitude of tilting. The sign of covariance decides the direction of tilting.

Question 3 (Comments)

The shapes vary because of the distance measurement methods. It is observed that when dmin is used, clusters are extended. Only first and second element are in first cluster and successively each element is added to the cluster. In the final step, they all agglomerate to one cluster.

However, when dmax is used there is a different pattern in how the clusters emerge. The last two elements are in one cluster. Then the second and third join to become the second cluster. Finally, the last point joins the cluster. Thus dmax produces well spaced cluster. Finally, we merge them all together.

Thus, depending on how many clusters we are looking for, both the methods will give us different or same results. If we are looking for two clusters, the result will be the same. First two elements will be in the first cluster and the rest will be in cluster two. If we are looking for three clusters, the result will be different. First two elements will be in one cluster, third element will be in a cluster of its own and the last three elements will be in one cluster. If we are looking for four clusters, we will have the same results from both methods. The first two elements will be in one cluster, last two will be in one cluster and the middle two will be two clusters of their own. If we look for five clusters, again the results will be the same. The first two elements will be in one cluster and rest will be in clusters of their own. If we are looking for six clusters, the result will be the same and all elements will be in clusters of their own.

4 Assuming PDF as 1-D gaussian:

We know that likelihood is the product of probabilities. Thus.

$$L(\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(\frac{-(x_1-\mu)^2}{2\sigma^2}\right)\right)\times\left(\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(\frac{-(x_2-\mu)^2}{2\sigma^2}\right)\right)\times\cdots\times\left(\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(\frac{-(x_n-\mu)^2}{2\sigma^2}\right)\right)$$

for $x_1 \dots x_n \in \underset{\sim}{x}$.

$$= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left(-\frac{(x_1-\mu)^2}{2\sigma^2}-\frac{(x_2-\mu)^2}{2\sigma^2}-\cdots-\frac{(x_n-\mu)^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right) \quad \left[\because a^x \cdot a^y = a^{x+y}\right]$$

Taking logarithm both sides, we get.

$$l(\mu,\sigma^2) = n\cdot\log\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right) - \left[\frac{1}{2\sigma^2}\cdot\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

↳ (A)

[Natural log].

for finding the MLE estimate of μ, differentiate Ⓐ wrt μ.

$$\frac{d\, l(\mu,\sigma^2)}{d\mu} = -\frac{1}{2\sigma^2} \cdot (-2) \sum_{i=1}^{n} (x_i - \mu)$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu).$$

Setting this to zero:

$$\frac{d\, l(\mu,\sigma^2)}{d\mu} := 0 \Rightarrow \sum_{i=1}^{n} (x_i - \mu) = 0$$

$$\Rightarrow \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu = 0$$

$$\Rightarrow \sum_{i=1}^{n} X_i - n\mu = 0$$

$$\Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} \quad //.$$

Second order derivative:

$$\frac{d^2\, l(\mu,\sigma^2)}{d\mu^2} = -1$$

Since SOD is negative, $\frac{d\, l(\mu,\sigma^2)}{d\mu}$ is the minimum.

Thus $\hat{\mu} = \frac{\sum x_i}{n}$ is the MLE estimate of population mean.

$\frac{d}{dx}\left(\frac{1}{x}\right)$

for variance we will differentiate $\textcircled{A}$ with σ^2.

$$\frac{d\,\ell(\mu,\sigma^2)}{d\sigma^2} = \frac{d}{d\sigma^2}\left(-\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$= \frac{d}{d\sigma^2}\left(-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}A\right)$$

where $A = \sum_{i=1}^{n}(x_i-\mu)^2$.

$$= \frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\cdot A.$$

Equating this to zero, we get.

$$\frac{d}{d\sigma^2}\ell(\mu,\sigma^2) = 0 \Rightarrow \frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}A = 0$$

If $\sigma^2 \neq 0$; then:

$$\frac{n}{} = \frac{A}{\sigma^2} \Rightarrow \sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}.$$

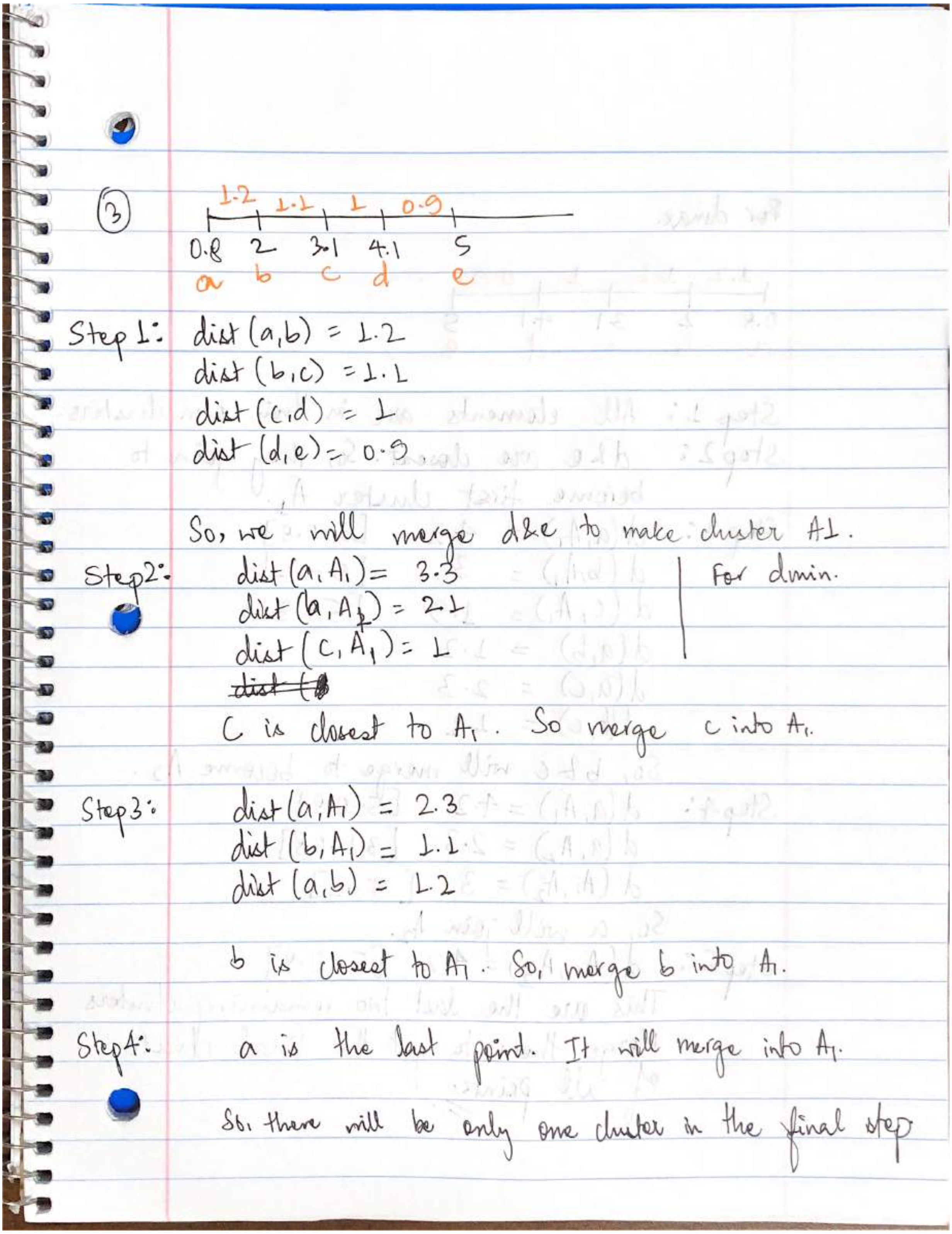
Second order condition:

$$\frac{d^2\ell(\mu,\sigma^2)}{(d\sigma^2)^2} = \frac{n}{(\sigma^2)^2} + \frac{\sum_i (x_i-\mu)^2}{(\sigma^2)^3}$$

Which will always be positive as all are squared terms.

Thus $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$

is the MLE estimate of σ^2.



(3)

1.2 1.1 1 0.9

0.8 2 3.1 4.1 5

a b c d e

Step 1: dist $(a,b) = 1.2$
dist $(b,c) = 1.1$
dist $(c,d) = 1$
dist $(d,e) = 0.9$

So, we will merge d&e to make cluster A1.

Step 2: dist $(a, A_1) = 3.3$ | For dmin.
dist $(b, A_1) = 2.1$
dist $(c, A_1) = 1$
~~dist (~~

C is closest to A_1. So merge c into A_1.

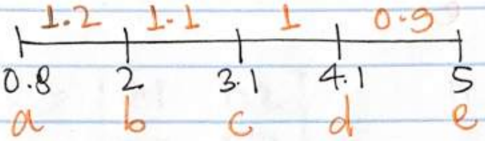
Step 3: dist $(a, A_1) = 2.3$
dist $(b, A_1) = 1.1$
dist $(a,b) = 1.2$

b is closest to A_1. So, merge b into A_1.

Step 4: a is the last point. It will merge into A_1.

So, there will be only one cluster in the final step

For dmax.



Step 1: All elements are in their own clusters.

Step 2: d & e are closest. So, they join to become first cluster A_1.

Step 3: $d(a, A_1) = 4.2$ $[5-0.8]$
$d(b, A_1) = 3$ $[5-2]$
$d(c, A_1) = 1.9$ $[5-3.1]$
$d(a,b) = 1.2$
$d(a,c) = 2.3$
$d(b,c) = 1.1$

So, b & c will merge to become A_2.

Step 4: $d(a, A_1) = 4.2$ $[5-0.8]$
$d(a, A_2) = 2.3$ $[3.1-0.8]$
$d(A_1, A_2) = 3$ $[5-2]$

So, a will join A_2.

Step 5: $d(A_1, A_2) = 4.2$ $[5-0.8]$
This are the last two remaining clusters Merge them to get the final cluster of all points.