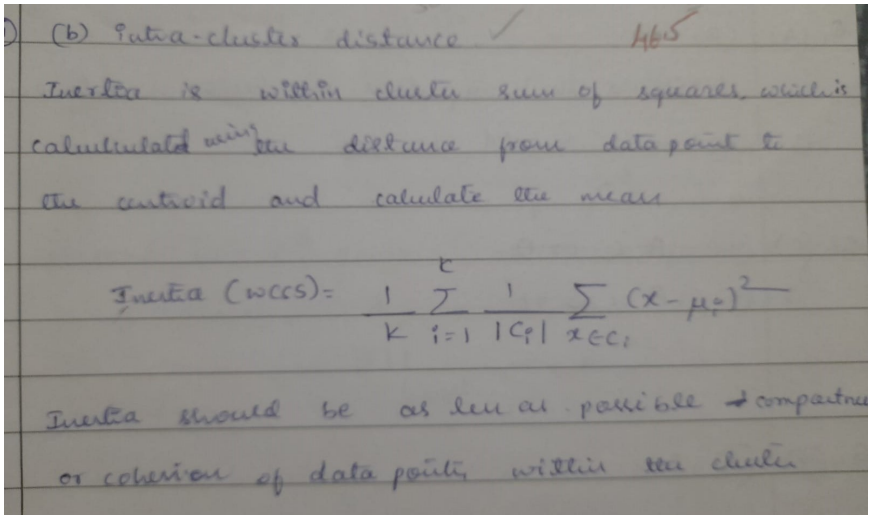
Sessional-1 AML5102 Applied Machine Learning

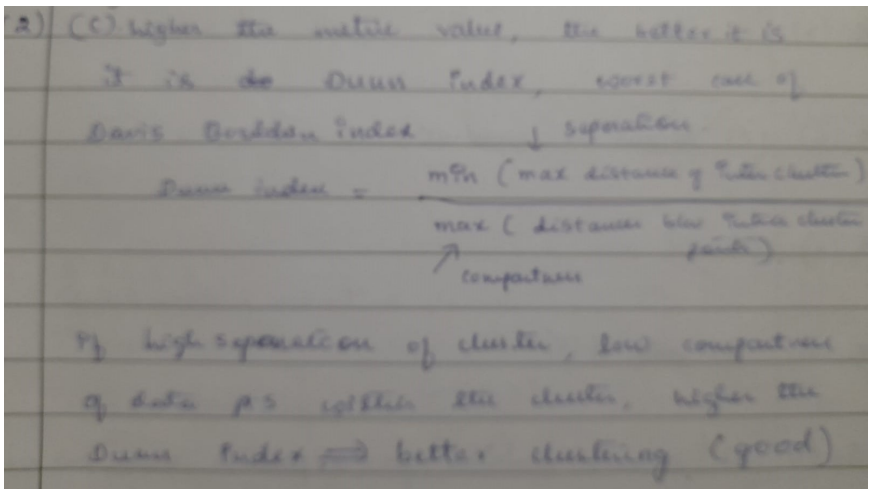
Q1: [CO 1, BT 4] 10 marks Objective Type 10 questions. 1 mark each.

Simple Selection. No reasoning is needed.

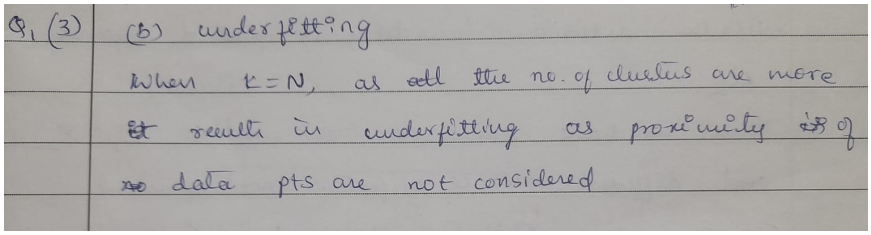
1. Inertia metric used in K-Means clustering cross validation incorporates
   (a) inter-cluster distance
   (b) intra-cluster distance
   (c) both inter-cluster distance, and inter-cluster distance equally accounted
   (d) More of intra-cluster distance. Less of inter cluster distance



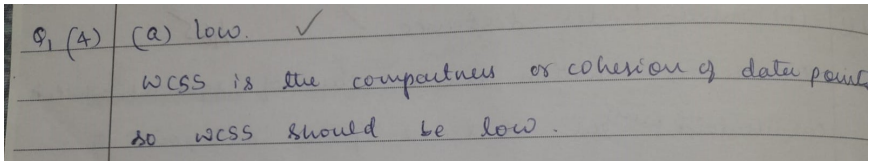
2. A metric has between cluster variance/separation in numerator and within cluster variance in denominator. Which of these is correct statement
   (a) the metric is wrong. between variance/separation should be in denominator and within cluster variance should be numerator
   (b) The metric value is between -1 and +1
   (c) higher the metric value, the better it is
   (d) lower the metric value, the better it is



3. If N is the size of the dataset, then selecting K = N in KNN causes
   (a) Curse of dimensionality
   (b) underfitting
   (c) overfitting
   (d) decision about underfitting/overfitting cannot be made without cross validation


Q1 (3) (b) underfitting
When K=N, as all the no. of clusters are more it results in underfitting as proximity of data pts are not considered

4. When elbow method is used to cross validate K in K-Means clustering, should WCSS be low or high?
   (a) Low
   (b) High
   (c) Only average WCSS should be high
   (d) Only WCSS averaged per cluster should be high


Q1 (4) (a) low. ✓
WCSS is the compactness or cohesion of data points
so WCSS should be low.

5. Which of the following imputation mechanisms will you choose given when you know the data is missing at random (MAR)

   A. Mean imputation

   B. Mode imputation

   C. Grouped Mean Imputation

   D. Iterative imputation

   (a) C or D
   (b) A, C or D
   (c) A or D
   (d) B
   (e) D
   (f) C
   (g) A
   (h) A or C

   Ans: (e)

6. The following assertions about StandardScaler are given. Choose all that are correct

   A. Standard Scaler transformation is impacted by outliers

   B. Records that were farther before Standard Scaler feature transformation may get closer after applying it

   C. Standard Scaler creates a ellipsoid around the origin that matches the Gaussian distribution

(a) A, B & C are correct
(b) A & C are correct. B is wrong
(c) B & C are correct, A is wrong
(d) A & B are correct, C is wrong
(e) B is correct, A & C are wrong

Ans: (d)

7. Curve fitting is the appropriate way to view ML classifiers.

(a) True only for binary classification.
(b) True only for multi class classification
(c) True for all classifiers
(d) False

Ans: (d)

8. Decision boundary with Gaussian mixture model clustering is

(a) always quadratic
(b) always quadratic or cubic
(c) always linear
(d) always linear or quadratic

Ans: (d)

9. Silhouette score is always a good metric for
A. DBSCAN
B. K-Means
C. Gaussian Mixture model clustering

(a) All three - A, B & C
(b) A & B, but not C
(c) A & C but not B
(d) B & C but not A

Ans: (d)

10. Given the silhouette plot, which cluster has outliers

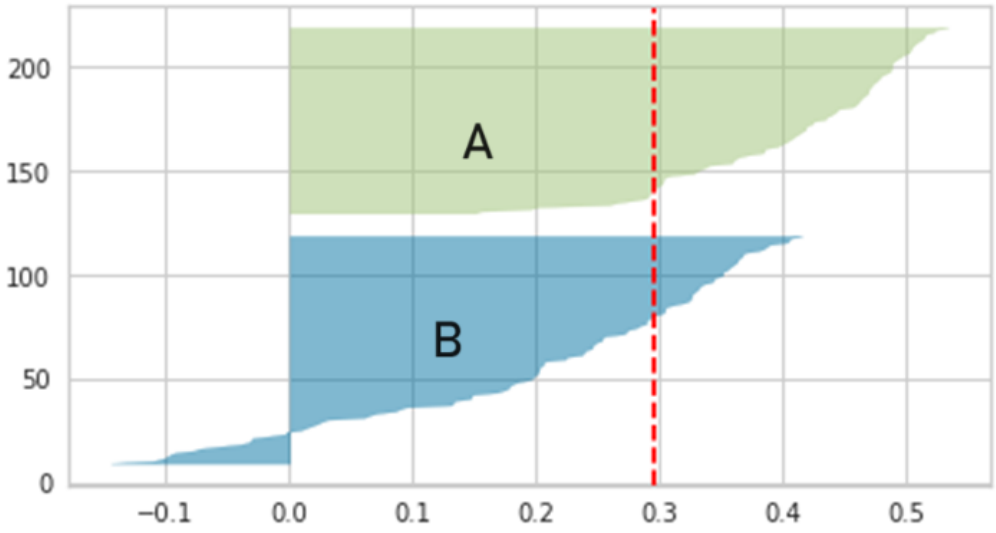


Figure 1: Silhouette analysis plot for K=2

(a) Both A and B

(b) Only A

(c) Only B

(d) None

Ans: (c)

Q2: [CO 1, BT 5] 10 marks Objective Type 5 questions. 2 mark each.

Provide your reasons along with your choice to get any marks

1. Two bivariate Gaussian distributions were fit for a dataset with two classes 1 and 2 for the target variable. The contour plots of the distributions and the decision boundary are as shown in the diagram below.

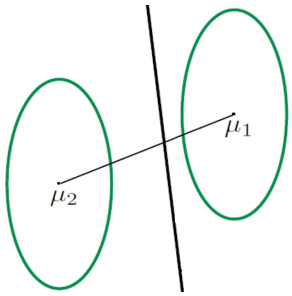


Figure 2: Two bivariate Gaussian distributions with decision boundary

Which of the conclusions are correct in addition to both distributions having same covariance matrices? Justify your reason along with your choice

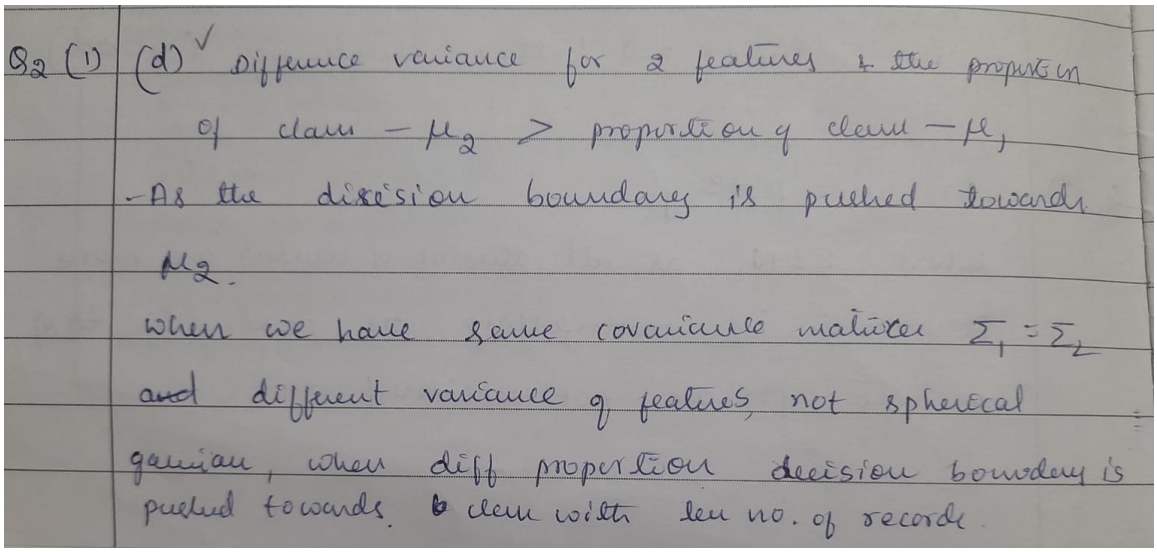
(a) Same variance for 2 features & both classes are present in equal proportion in the dataset

(b) Same variance for 2 features but the proportion of records from class corresponding to μ_1 is more in the dataset

(c) Same variance for 2 features and the proportion of records from class corresponding to μ_2 is more in the dataset

(d) Different variance for 2 features and the proportion of records from class corresponding to μ_2 is more in the dataset

(e) Different variance for 2 features and the proportion of records from class corresponding to μ_1 is more in the dataset


Q2 (1) (d) ✓ Diffrence variance for 2 features & the proportion of class $-\mu_2$ > proportion of class $-\mu_1$

- As the decision boundary is pushed towards μ_2.

When we have same covariance matrix $\Sigma_1 = \Sigma_2$ and different variance of features, not spherical gaussian, when diff proportion decision boundary is pushed towards class with less no. of records.

Figure 3: Q2-1 Solution

2. The following are given

I. $y = w^T x + b$

II. $w^T x + b = 0$

$x \in R^d$

The following assertions are made

A. I is a decision boundary for regression. II is decision boundary for classification

B. I and II can both be plotted in R^d

C. I is plotted in R^d, II is plotted in R^{d+1}

D. I is plotted in R^{d+1}, II is plotted in R^d

E. I is a equation for regression. II is decision boundary for classification

(a) A & C are correct

(b) A & D are correct

(c) A & B are correct

(d) B & E are correct

(e) C & E are correct

(f) D & E are correct

Ans: f

3. Anjali and Aishwarya were doing EDA for their mini project at the last minute. Anjali first applied StandardScaler transformation to two features x1 and x2 and obtained z1 and z2. Then Anjali realized she has to check the correlation between the two features, failing which their ML prediction will not be robust and they will lose marks. Since both were in hurry, Aishwarya created a new feature x3 as the product of z1 and z2 and calculated its expected value. She got 0.8 as the expected value of x3 and Aishwarya concluded the features x1 and x2 have high positive correlation.

   But Anjali being meticulous, was not happy with Aishwarya's method. But instead of calculating correlation between x1 and x2, she calculated the covariance between z1 and z2. To her utter surprise, she got the same value 0.8 and hence concluded the features x1 and x2 have high positive correlation.

   Whose method is flawed - Aishwarya's method or Anjali's method?. If the provided reason feels incomplete, you are encouraged to provided additional reasoning in 1-2 sentences to prove your point.

   (a) Aishwarya's method is flawed. Feature engineering a new feature as the product of two z-features and averaging them as if it is correlation coefficient has no mathematical basis

   (b) Anjali's method is flawed and not the right way to calculate correlation. Anjali only calculated the covariance and not correlation. She should have mean centered the features x1 & x2 or z1 & z2 and calculated correlation instead of covariance

   (c) Both methods are flawed. Getting 0.8 in both case was just a fluke

   (d) Both methods are right

   Ans: d. Using the standard formula of Covariance for a zero centered feature, we see it is product of expected values of features

4. A diagonal matrix with w1, w2... wn on the principal diagonal is used as weights in nearest centroid model training for weighting the features during euclidean distance calculation. Which of the following is equivalent to weighting with a diagonal matrix?

   (a) Covariance matrix with all variances = 1 and covariances = 0

   (b) Covariance matrix with variance of all features = 1

   (c) Covariance matrix with zero covariance between all features

   (d) Insufficient data to conclude in any of the three ways listed above

   Ans c

5. Milan is doing a high profile Machine Learning project for Food Safety and Standards Authority of India (FSSAI). He has collected data from students with many features in it. Among them is one feature about eating at Boda sheera and found that 75 % of students eating at Boda sheera ended up with upset stomach within next 24 hours. Another feature is whether students ate at Suresh mess in the last 48 hours. He found that in his dataset, 85 % of students eating at Suresh Mess caught viral fever within 48 hours. Using the data, he creates a machine learning model that correctly predicts the target variable of student sick Yes/No with an accuracy of 98 %. You as the best resident data scientist at MSIS are helping Milan derive right conclusions from the data and ML model predictions. Which of the following conclusion will you draw from the machine learning model

   (a) Eating at Boda sheera is causing students to fall sick

   (b) Eating at Suresh mess is causing students to fall sick

   (c) Eating at either Boda sheera or Suresh mess is causing students to fall sick

   (d) It cannot be said that eating at either Boda sheera or Suresh mess is causing students to fall sick

   Ans d. Reason: Correlation is not causation

Q3: [CO 2, BT 3] 10 marks Numerical problems 2 questions. 5 mark each.

Apply your steps clearly. Inaccuracies in actual numerical calculations is not an issue and will not result in losing any marks

The following dataset of points in 2D is given. {(2,5), (-1,3), (-2,3), (3,1), (1,4), (4,1), (4,2)}

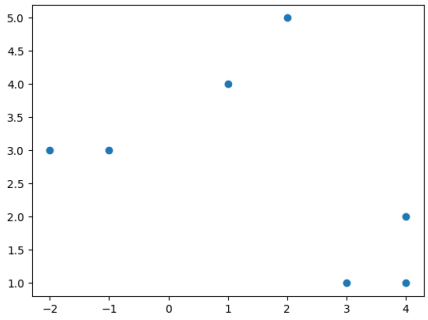


Figure 4: Plot of the data points

The task is to fit 3 clusters

1. Use k-means++ algorithms and extract robust centroids for initialization. Show all the steps and calculations, then sample the next centroids manually considering the individual probabilities as weights for each data point and arrive at final three centroids

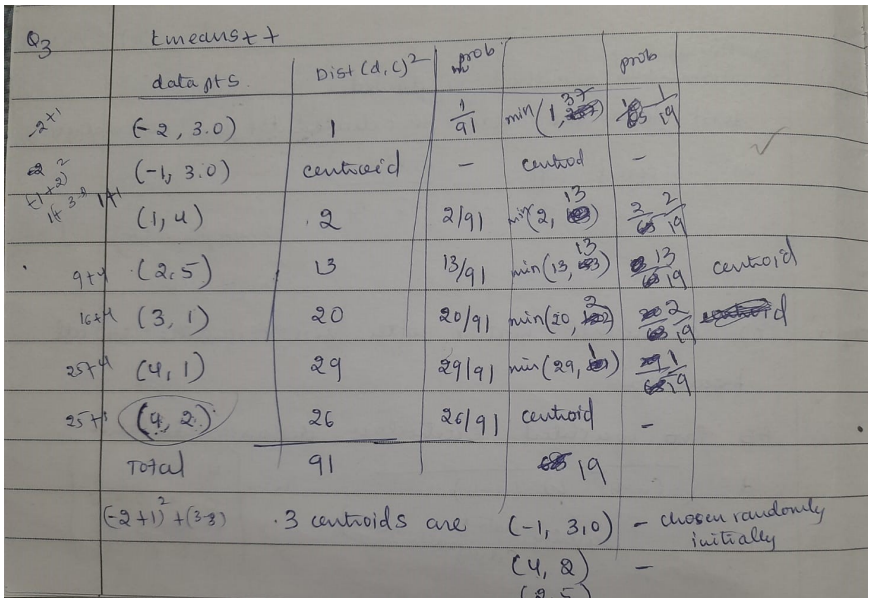


Figure 5: Q3-1 Solution

2. Using the centroids that you arrived at in previous step, perform two iterations of K-Means clustering algorithm. Show all the calculations.

   Ans

   Step 1: Let (-1,3), (2,5) and (4,1) be initial centroids μ_1, μ_2, μ_3.

   Step 2 (Expectation): Calculate which points are close to the centroids

Point	$\|p-\mu_1\|^2$	$\|p-\mu_2\|^2$	$\|p-\mu_3\|^2$	$\underset{\mu_i}{\text{argmin}} \ \|p-\mu_i\|^2$
(-2,3)	1	20	40	μ_1
(-1,3) (μ_1)	0	gt 0	gt 0	μ_1
(1,4)	5	2	18	μ_2
(2,5) (μ_2)	gt 0	0	gt 0	μ_2
(3,1)	20	17	1	μ_3
(4,1) (μ_3)	gt 0	gt 0	0	μ_3
(4,2)	26	13	1	μ_3

Step 3 (Maximization): Calculate new centroids using the current clustering

$$\mu_{1_{new}} = \frac{1}{2}\left(\begin{bmatrix} -2 \\ \\ 3 \end{bmatrix} + \begin{bmatrix} -1 \\ \\ 3 \end{bmatrix} \right) = \begin{bmatrix} -3/2 \\ \\ 3 \end{bmatrix}$$

$$\mu_{2_{new}} = \frac{1}{2}\left(\begin{bmatrix} 1 \\ \\ 4 \end{bmatrix} + \begin{bmatrix} 2 \\ \\ 5 \end{bmatrix} \right) = \begin{bmatrix} 3/2 \\ \\ 9/2 \end{bmatrix}$$

$$\mu_{3_{new}} = \frac{1}{3}\left(\begin{bmatrix} 3 \\ \\ 1 \end{bmatrix} + \begin{bmatrix} 4 \\ \\ 1 \end{bmatrix} + \begin{bmatrix} 4 \\ \\ 2 \end{bmatrix} \right) = \begin{bmatrix} 11/3 \\ \\ 4/3 \end{bmatrix}$$

Repeat Step 2 and 3 one more time.

Q4: [CO 2, BT 4] 10 marks 2 questions

1. (3 marks)

   (a) (1 mark) Classify the three algorihms, Nearest Centroid, kNN and Kmeans clustering into categories Supervised and unsupervised

   (b) (2 marks) Fill the table below by categorizing the 3 algorithms into appropriate categories. Categories are specified as horizontal rows

Algorithm	Nearest Centroid	kNN	Gaussian Model
Eager vs Lazy	?	?	?
Batch vs Instance	?	?	?
Parametric vs Non-parametric	?	?	?
Discriminative vs Generative	?	?	?

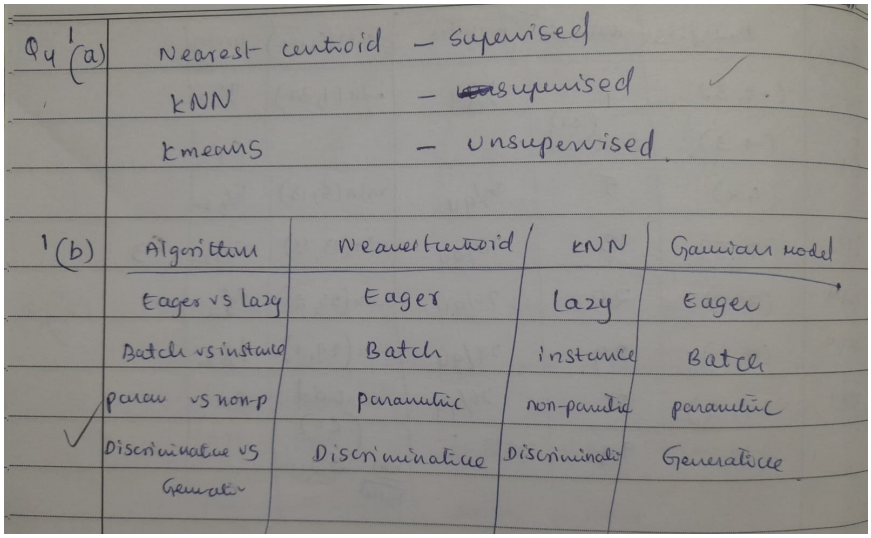


Figure 6: Q4-1 Solution

2. (3 marks) A dataset consists of m records and each record has n features. All features are floating point. The records belong to 3 classes - 0, 1 and 2 (much like an Iris dataset). The classes can be considered to be integers.

   It takes 4 bytes to store a floating point. Each integer can be stored in 1 byte.

   Two models are developed using Nearest Centroid and kNN algorithms respectively. kNN algorithm has two integer hyperparameters viz. k and weighting-mechanism. Answer the following questions to contrast the two models.

   (a) (1 mark) What is the storage space (in bytes) taken up by nearest centroid model?

   (b) (2 marks) What is the storage space (in bytes) taken up by kNN model?

   Show your calculations in each case.

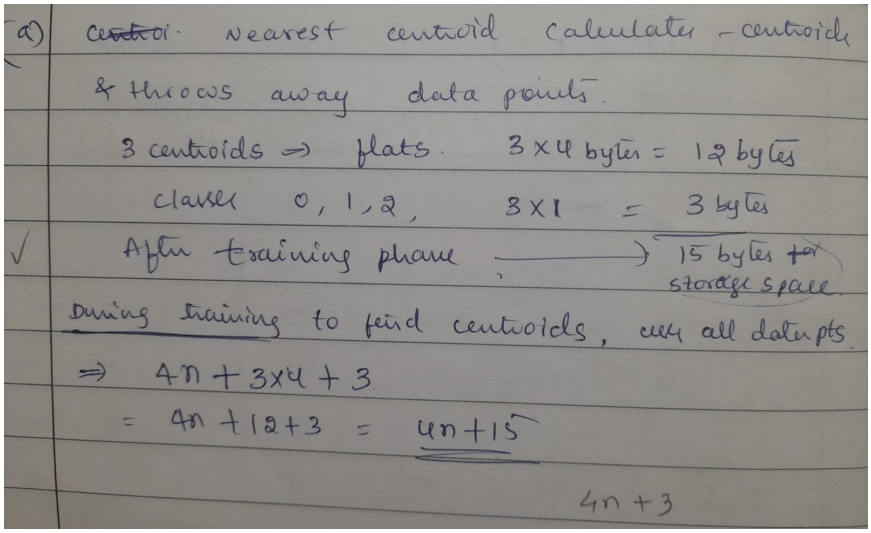


Figure 7: Q4-2a Solution

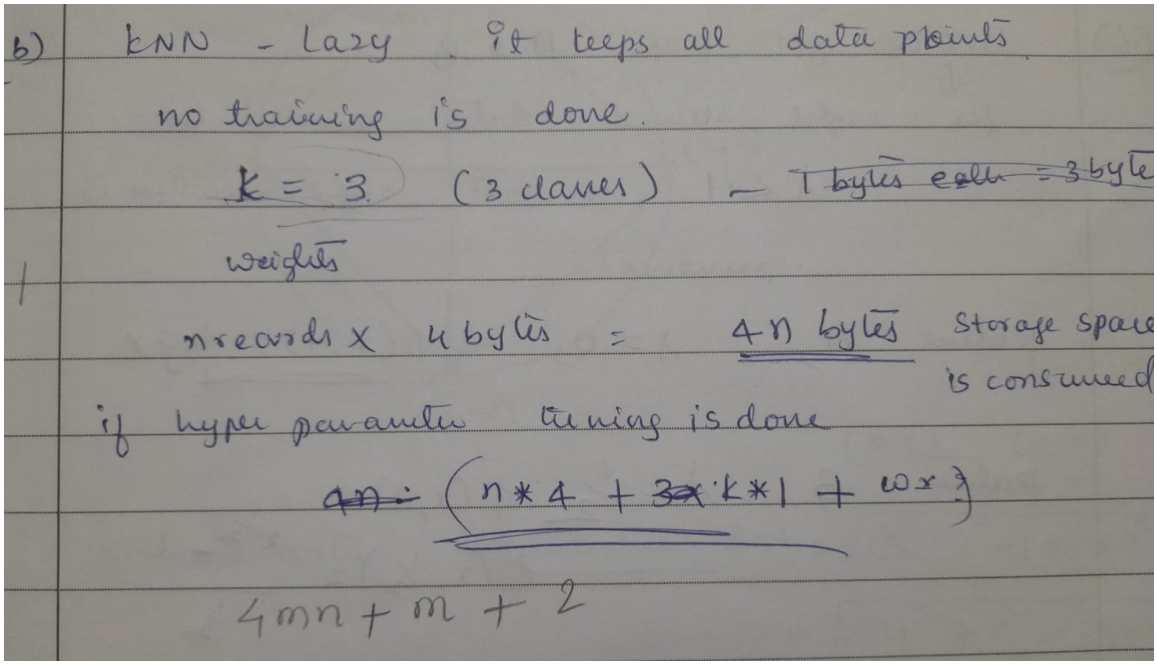


Figure 8: Q4 2b Solution

3. (4 marks) Feature Transformation question

   A dataset consists of two features A and B as shown in the diagrams (left and right figures respectively). Feature A follows lognormal distribution. Feature B has moderately right skewed distribution and both feature values are greater than 0.

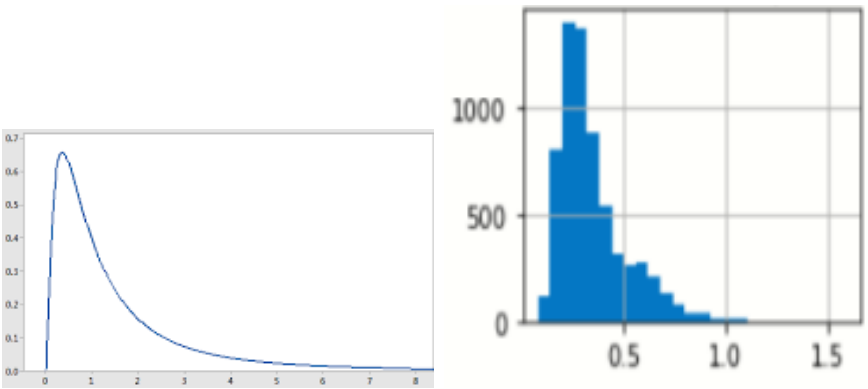


Figure 9: Lognormal distributed Feature A, moderately right skewed distribution of Feature B

(a) (1 mark) Which of the following is most appropriate to apply on Feature A? LogTransformer, Square Transformer, Square-root transformer? Provide reasoning/mathematical calculation

(b) (1 mark) Which of the following is most appropriate to apply for Feature B? LogTransformer, Square Transformer, Square-root transformer? Provide reasoning/mathematical calculation

(c) (2 marks) If Box-Cox transformation were to be applied to both features A and B, then suggest a value for lambda for transformation of each feature A and B. Provide reasoning/mathematical calculation

Box-Cox transformation is applied when features are greater than 0 and lambda is real value generally between -5 and +5. Formula is given below:

$$\phi(x, \lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ logx, & \text{if } \lambda = 0 \end{cases}$$

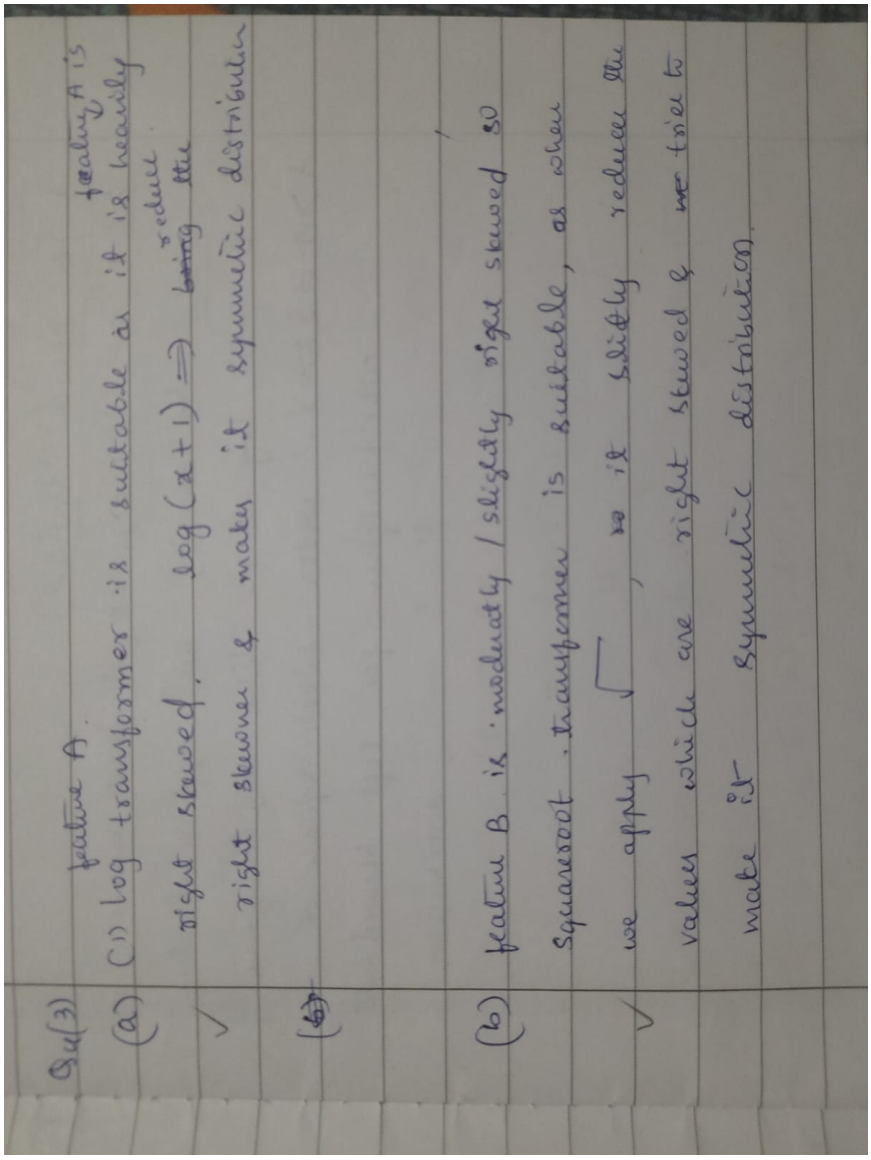


Figure 10: Q4-3ab Solution

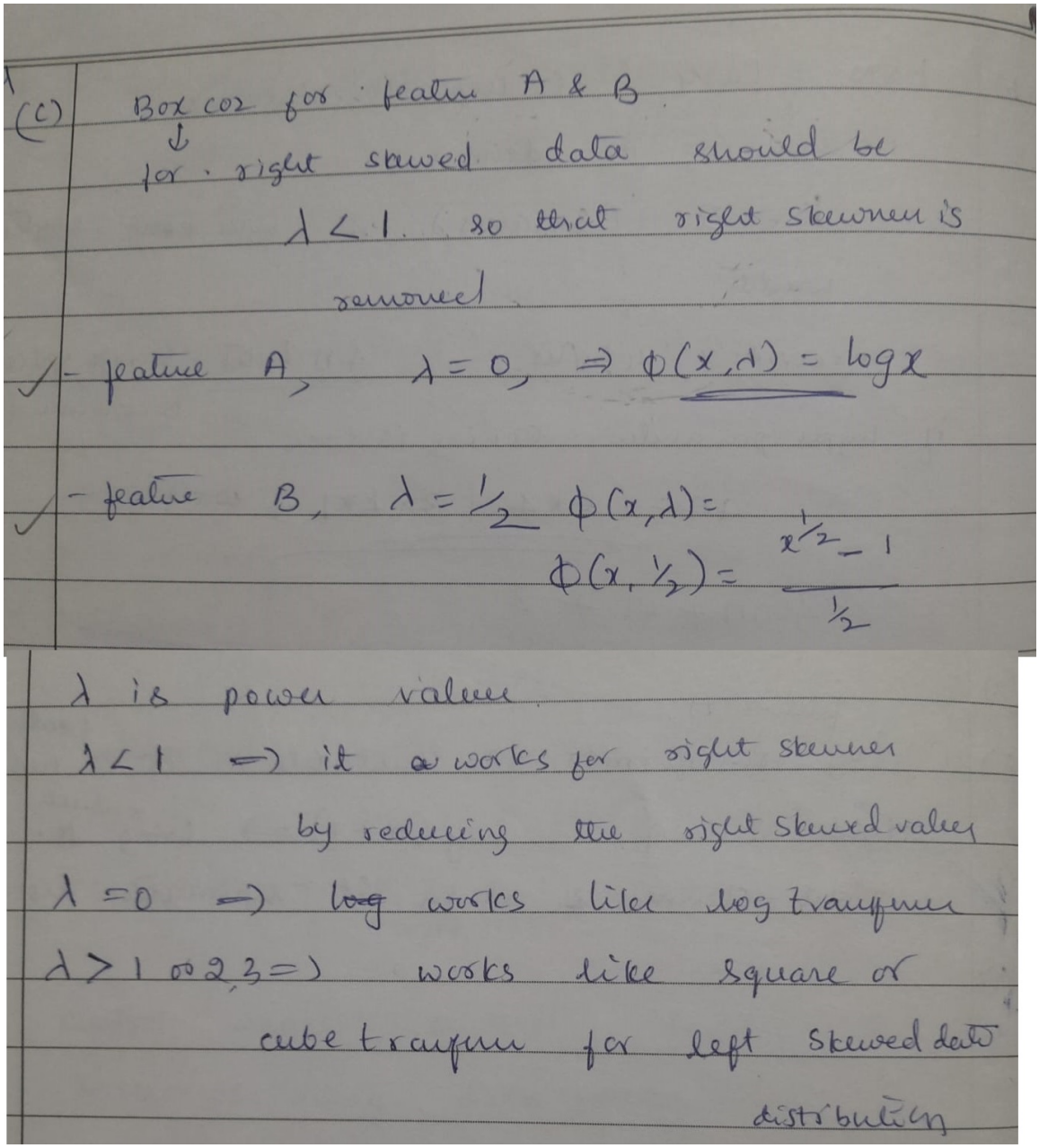


Figure 11: Q4 3c Solution

Q5: [CO 2, BT 3] 10 marks. 2 questions.

Please show your steps clearly. Inaccuracies in actual numerical calculations is not an issue and will not result in losing any marks

1. (4 marks) Below is a plot of a circle and rhombus that intersects x-axis at (4,0) and y-axis at (0,4) in the first quadrant. Similarly other points of intersection are shown.

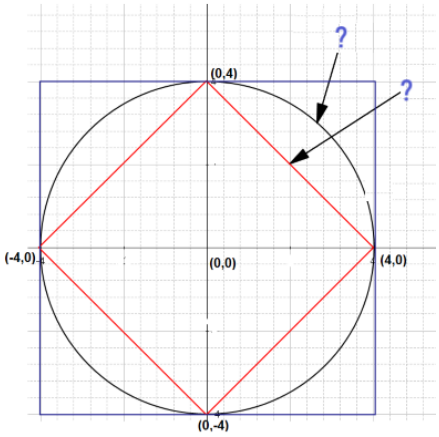


Figure 12: Circle and rhombus

   (a) (2 marks) All points on the circle are equidistant from the origin when a certain distance metric is used. Which is that distance metric? Provide reasons preferably using mathematical equation of the circle. (Hint: you can use Pythagoras theorem to identify any point on the circle)

   (b) (2 marks) All points on the rhombus are equidistant from the origin when a certain distance metric is used. Which is that distance metric? Provide reasons preferably using mathematical equation of the rhombus.(Hint: you can use any two points to get the equation of the straight line that forms one side of the rhombus in a given quadrant)

2. (6 marks) Titanic dataset captures the passenger details of the Titanic ship that sank after colliding with a iceberg. The survival rate was 30% [P(y=1) = 0.3].

   Survival is the target variable that takes two values 0 and 1. The dataset was analyzed and one feature "Age" is selected. Age is a numeric field obeying Gaussian distribution.

   For those that survived, Expected Value of age, $\mathbb{E}[Age] = 30$ with a variance of 8 For those that did not survive, Expected Value of age, $\mathbb{E}[Age] = 50$ with a variance of 10

   Apply the generative modeling approach of fitting one distribution per class using the data provided above and using Bayes theorem

$$P(y = 1|Age =?) = \frac{P(Age =?|y = 1)P(y = 1)}{P(Age =?)}$$

$$P(y = 0|Age =?) = \frac{P(Age =?|y = 0)P(y = 0)}{P(Age =?)}$$

   For $P(Age =?|y = 1)$ and $P(Age =?|y = 0)$, use univariate Gaussian.

   What does your model predict for survival of a person with age 35? Show the steps and calculations. Arriving at the actual numeric values is not important. But the steps should clearly indicate that they unambiguously lead to the final answer.

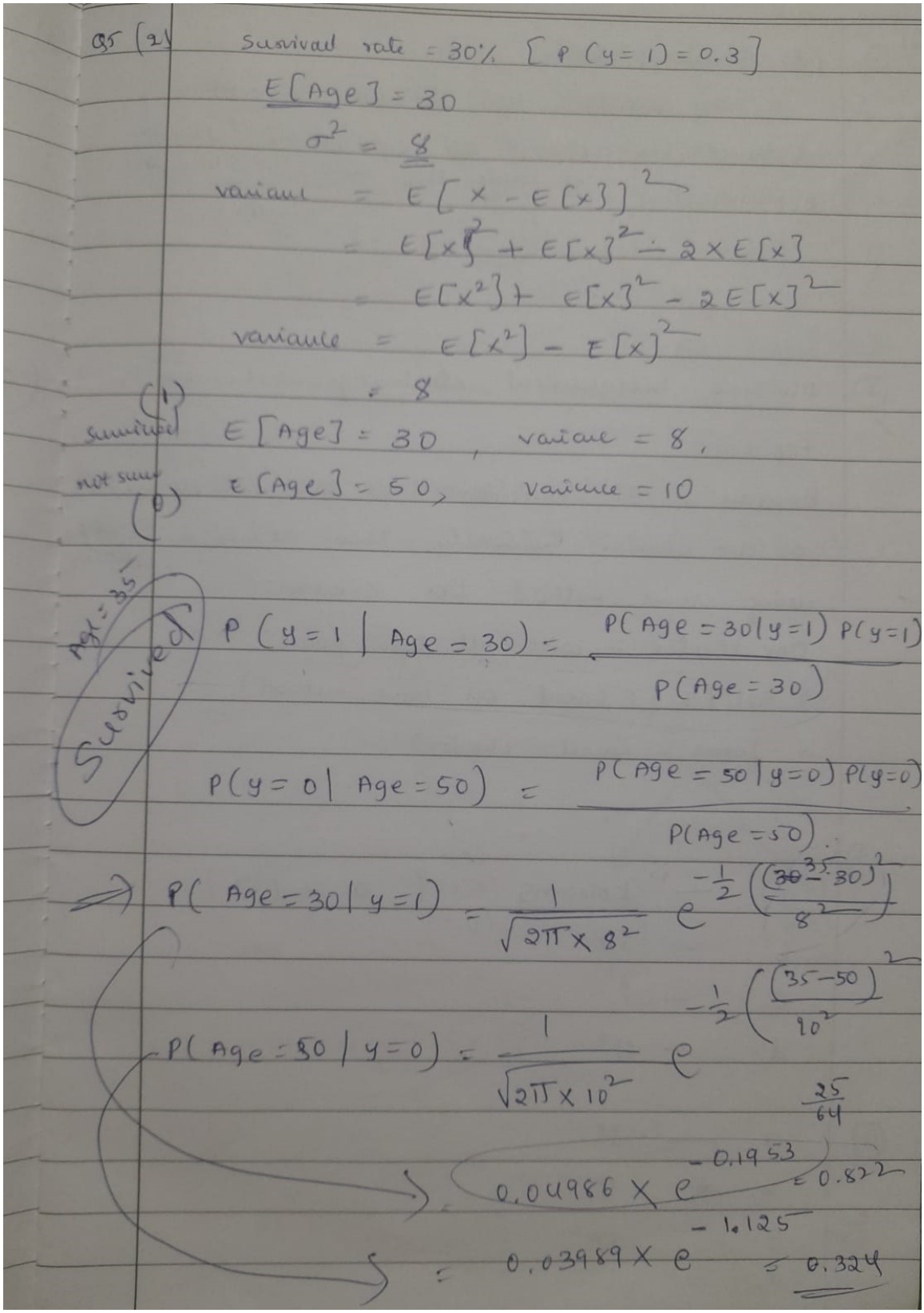


Figure 13: Q5 Solution