CSE343/CSE543/ECE363/ECE563: Machine Learning Sec A (Monsoon 2024)
End-sem

Date of Examination: 12.12.2024 Duration: 2 hour Total Marks: 30 marks

---

Instructions:

- Attempt all questions. State any assumptions you have made clearly.
- MCQs may have multiple correct options. No evaluation without suitable justification.
- Standard institute plagiarism policy holds.

Section A: MCQ

1. **[1 mark]** You clustered n data points using k-means and Gaussian Mixtures, obtaining 5 clusters with identical cluster centres in both methods. Three points assigned to different clusters by k-means are observed to belong to the same cluster in the Gaussian mixture solution. Which of the following statements is/are most accurate regarding this observation?

   a. K-means assigns each point deterministically to the closest cluster centre, whereas Gaussian Mixtures assign probabilities to points, allowing overlaps in cluster boundaries.

   b. Identical cluster centres ensure identical assignments in both K-means and Gaussian Mixtures, so the observation is likely due to a computational error.

   c. K-means and Gaussian Mixtures optimise different objectives, so their assignments differ only when initialisation values for cluster centres are not identical.

   d. Gaussian Mixtures can have cluster components with high variances, potentially assigning points near the edges of clusters to the same cluster.

   Solution: Both A and D are correct, as Gaussian Mixtures allow probabilistic overlaps (A) and high variances (D) can lead to shared assignments near boundaries. Option B is incorrect because cluster centres alone do not guarantee identical assignments. Option C is incorrect since the methods' objectives differ regardless of initialization.
   1 mark for correct option and correct reason

2. **[1 mark]** In the Expectation-Maximization (EM) algorithm for Gaussian Mixture Models (GMMs), which statement about convergence is most accurate?

   a. EM for GMMs always converges to the global maximum of the log-likelihood.

   b. EM for GMMs can converge to a saddle point or a local maximum depending on initialization.

   c. EM for GMMs does not guarantee any improvement in the log-likelihood at each iteration.

   d. EM for GMMs is guaranteed to converge only if the covariance matrices are fixed and not learned.

   Solution: B. EM for GMMs can converge to a saddle point or a local maximum, depending on initialization.
   1 mark for correct option and correct reason

3. **[1 mark]** Suppose S_1 and S_2 are the set of support vectors and w_1 and w_2 are the learned weight vectors for a linearly separable problem using hard and soft margin linear SVMs, respectively. Which of the following is correct?

   a. $S_1 \subseteq S_2$

   b. $w_1 = w_2$

   c. S_1 may not be a subset of S_2

   d. w_1 may not be equal to w_2

   Solution: C, D
   (c) S1 may not be a subset of S2: This is correct. By allowing margin violations, the set of support vectors can change. Some points that were critical for the hard margin solution may no longer be support vectors, and new points may enter the set. Hence, it's possible that S1 is not contained within S2.
   (d) w1 may not be equal to w2: This is correct. The objective functions differ due to the slack variables and their associated penalties. This can yield a different optimal solution for the weight vector when comparing the hard-margin and soft-margin SVMs.
   1 mark for correct option and correct reason

4. **[1 mark]** If AI models are generated by biased, compromised, or skewed datasets, then some stakeholders will not adequately be protected from discriminatory harm. Which of the following steps could help in ensuring fairness in such situations:

   a. Ensuring equal representativeness of all groups.

   b. Having sufficient samples for training, even if the attributes/measurements are irrelevant.

   c. Ensuring the timeliness and recency of the data.

   d. Collecting data with domain knowledge to ensure relevance and appropriateness.

   e. All of the above.

   Solution: e
   1 mark for correct option and correct reason

5. **[1 mark]** Which of the following propositions is true about a CONV layer?

   a. The number of weights depends on the depth of the input volume.

   b. The number of biases is equal to the number of filters.

   c. The total number of parameters depends on the stride.

   d. The total number of parameters depends on the padding.

   Solution: A, B
   1 mark for correct option and correct reason

Section B: Short Answers

6. **[2 marks]** Let p be the probability of keeping neurons in a dropout layer. We have seen that in forward passes, we often scale activations by dividing them by p during training time. You accidentally train a model with dropout layers without dividing the activations by p. How would you resolve this issue at test time? Please justify your answer mathematically.
   Solution:

   - Dropout switched off at test time - expected activation thus larger than train by factor $1/p$; fix by multiplying outputs of each layer by p.
   - Some mathematical expressions involve an expectation over activations and how the scaling factor helps rescale the expected value.

   1 mark for correct reason and 1 mark for correct justification

7. **[2 marks]** Consider a dataset with two points in 1D: $(x_1 = 0, y_1 = -1)$ and $(x_2 = \sqrt{2}, y_2 = 1)$ (You can consider these points as support vectors). We will map each point in 3D using the feature vector:

   $\phi(x) = [1, \sqrt{2} * x, x^2]^T$
   This mapping is equivalent to using a second-order polynomial kernel. The max-margin classifier has the form:
   minimize $\|w\|^2$
   subject to:
   $y1 * (w^T * \phi(x_1) + w_0) \geq 1$
   $y2 * (w^T * \phi(x_2)(+ w_0) \geq 1$

   (a) Write down a vector that is parallel to the optimal vector w.

   (b) What is the value of the margin achieved by this w?

   (c) Solve for w, using the fact that the margin is equal to $\frac{1}{\|w\|}$. Also, Solve for w_0 using your value for w and the inequalities above. Write down the form of the discriminant function $f(x) = w_0 + w^T * \phi(x)$ as an explicit function of x.

   Rubric- 0.5 for part a, 0.5 for part b and 1 for part c

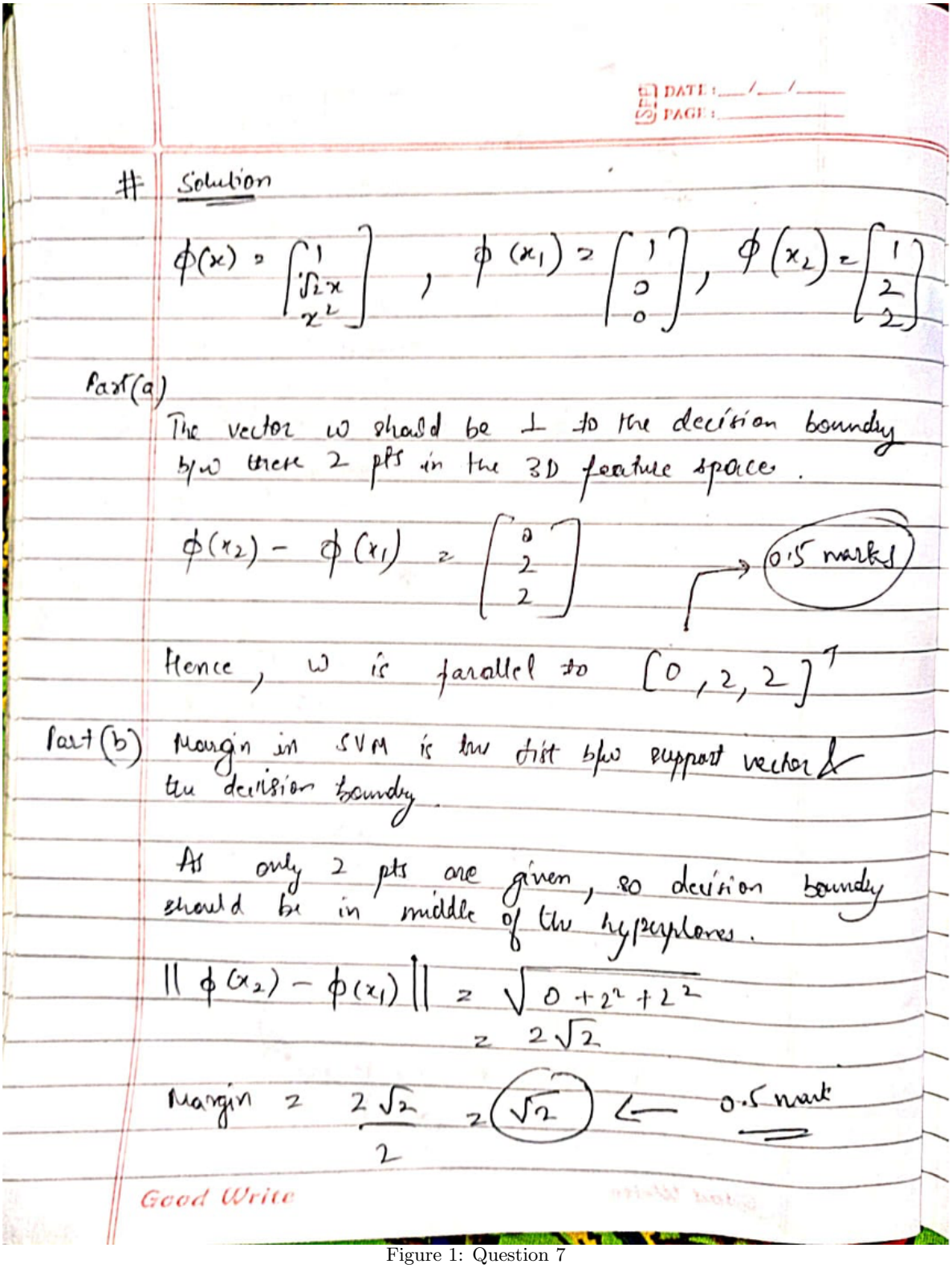


Figure 1: Question 7

8. **[2 marks]** In a Gaussian Mixture Model (GMM), the Expectation-MAximization (EM) algorithm alternates between:
   1. E-step: Compute responsibilities r_{ik}, the probability that datapoint x_i belongs to Gaussian component k.
   2. M-step: Update parameters μ_k and $\sum_k$ to maximize the likelihood of the data.

   The M-step update for the covariance matrix $\sum_k$ of the k-th Gaussian is given by:
   $\sum_k = \frac{\sum_i r_{ik}(x_i - \mu_k)(x_i - \mu_k)^T}{r_k}$ where $r_k = \sum_i r_{ik}$ or equivalently: $\sum_k = \frac{\sum_i r_{ik} x_i x_i^T}{r_k} - \mu_k \mu_k^T$.

   Variable:
   x_i : i-th datapoint.

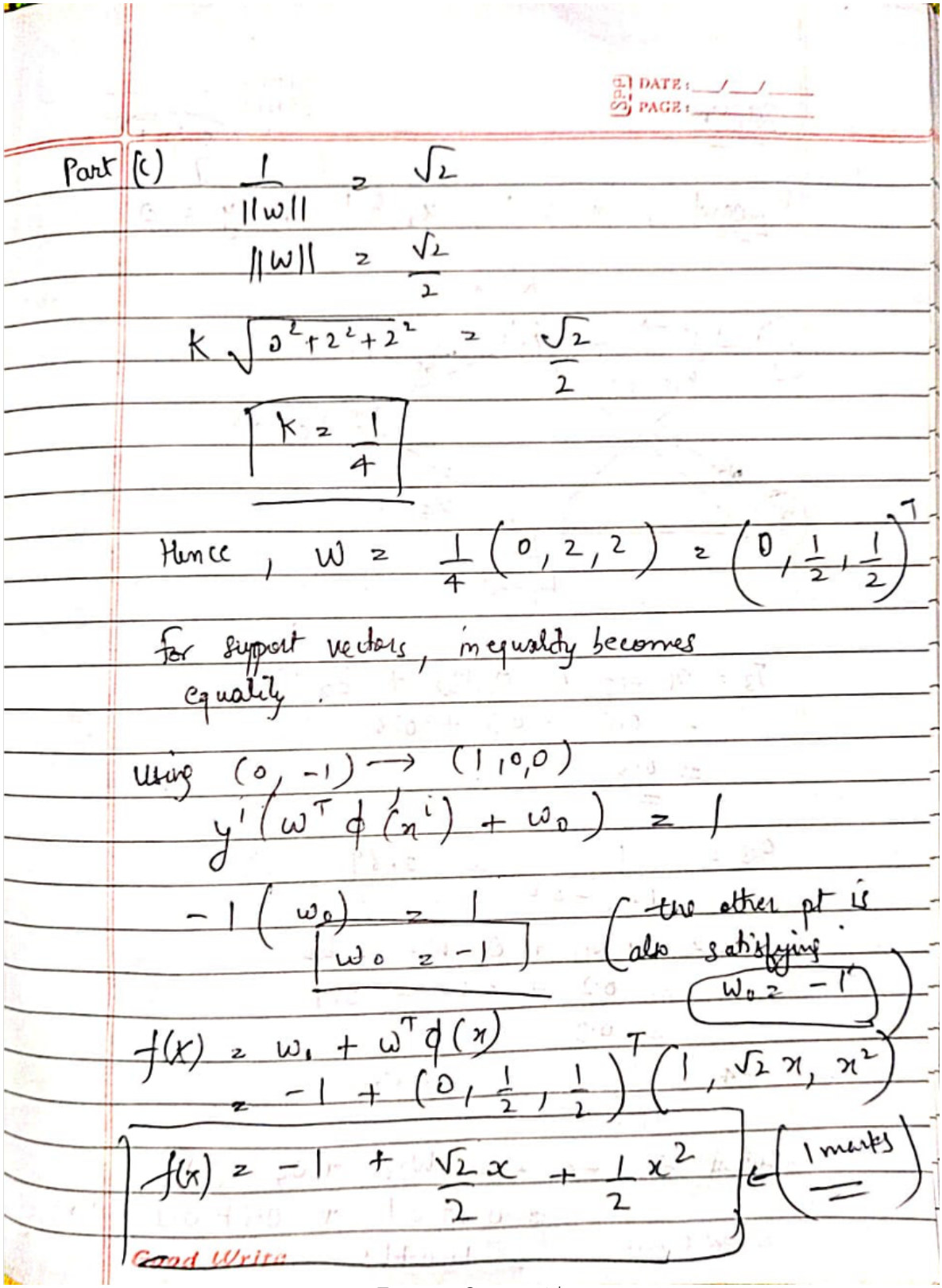


Figure 2: Question 7*

r_{ik}: Responsibility of component k for x_i.
$r_kc = \sum_i r_{ik}$: Total responsibility for component k.
μ_k: Mean of component k.
$\sum_k$: Covariance matrix for component k.
Task: Derive the formula for $\sum_k$ in both forms.
Rubric- 1 marks for correctly starting the definition of $\sum_k$ and expanding and simplifying $(x_i - \mu_k)(x_i - \mu_k)^T$
1 marks for deriving both form of $\sum_k$; the weighted covariance form and decomposition

1. **Definition of Covariance Matrix:** The covariance matrix Σ_k of the k-th Gaussian component is defined as the weighted expectation (using responsibilities) of the outer product of $(\mathbf{x}_i - \mu_k)$:

$$\Sigma_k = \mathbb{E}_{\text{responsibilities}}\left[(\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T\right].$$

Using the responsibilities r_{ik}, this becomes:

$$\Sigma_k = \frac{\sum_i r_{ik}(\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T}{r_k}, \quad \text{where } r_k = \sum_i r_{ik}.$$

2. **Expanding $(\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T$:**
The term $(\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T$ can be expanded as:

$$(\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T = \mathbf{x}_i\mathbf{x}_i^T - \mathbf{x}_i\mu_k^T - \mu_k\mathbf{x}_i^T + \mu_k\mu_k^T.$$

3. **Substitute Into the Summation:**
Substitute the expanded term into the weighted summation:

$$\Sigma_k = \frac{\sum_i r_{ik}\left(\mathbf{x}_i\mathbf{x}_i^T - \mathbf{x}_i\mu_k^T - \mu_k\mathbf{x}_i^T + \mu_k\mu_k^T\right)}{r_k}.$$

4. **Simplify Each Term:**

- The first term:

$$\frac{\sum_i r_{ik}\mathbf{x}_i\mathbf{x}_i^T}{r_k}.$$

- The second and third terms combine since they are symmetric:

$$\frac{\sum_i r_{ik}\mathbf{x}_i}{r_k}\mu_k^T + \mu_k\frac{\sum_i r_{ik}\mathbf{x}_i^T}{r_k}.$$

Using the definition of the mean:

Figure 3: Question 8

$$\mu_k = \frac{\sum_i r_{ik}\mathbf{x}_i}{r_k},$$

these terms simplify to:

$$-\mu_k\mu_k^T.$$

- The last term:

$$\frac{\sum_i r_{ik}\mu_k\mu_k^T}{r_k} = \mu_k\mu_k^T.$$

Substituting these into the equation:

$$\Sigma_k = \frac{\sum_i r_{ik}\mathbf{x}_i\mathbf{x}_i^T}{r_k} - \mu_k\mu_k^T.$$

5. **Final Covariance Matrix Form:**
Therefore, the covariance matrix can be written in two forms:

- Explicit weighted covariance form:

$$\Sigma_k = \frac{\sum_i r_{ik}(\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T}{r_k}.$$

- Alternative decomposition:

$$\Sigma_k = \frac{\sum_i r_{ik}\mathbf{x}_i\mathbf{x}_i^T}{r_k} - \mu_k\mu_k^T.$$

Figure 4: Question 8*

9. **[2 marks]** While doing the recruitment for a job role, company PQR considers the following criteria: Education level, Skills and experience, Job performance during the interview, Relevant certifications or qualifications, Race, Gender, Age, Religion, Sexual orientation, Physical disability, Marital status, and Nationality. Given the above features, partition those into two classes:
Should Use: For fairness, which features should PQR take into account.
Must Avoid: For fairness, the decision must not discriminate based on these features, as much as possible.
Give suitable justification.

Solution: Should Use: Education level, Skills and experience, Job performance during the interview, Relevant certifications or qualifications
Must Avoid: Race, Gender, Age, Religion, Sexual orientation, Physical disability, Marital status, Nationality

0.25 marks for each correct classification for any 8 right answers. No marks without justification.

10. **[2 marks]** You are implementing the K-means algorithm. You are provided with a dataset as shown in the figure.

    a. What is the predicted value of y when $x = 1$?
    b. What is the predicted value of y when $x = 3$?
    c. What is the predicted value of y when $x = 5$?

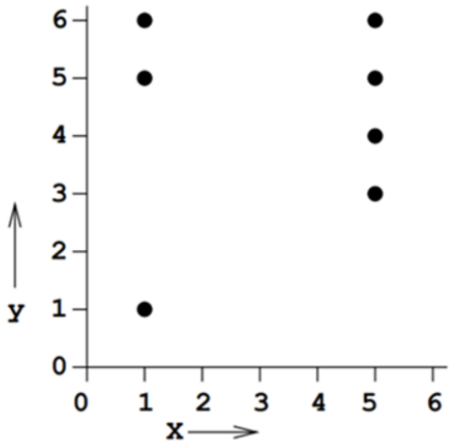


Figure 5

(a) mean(6,5,1) = 4
(b) X is equidistant from all, so (6+5+1+6+5+4+3)/7 = 4.3
(c) mean(6+5+4+3)/4 = 4.5
0.5 marks for a, 1 marks for b, 0.5 marks for c

Section C: Descriptive Answers

11. **[4 marks]** We learned a two-class linear SVM for linearly separable input data. Let W and b be the parameters we obtained for the primal SVM formulation. In the standard SVM formulation (SVM1), we use the following constraints:
For all x in class 1:

$$W^T x + b \geq 1$$

For all x in class 0:

$$W^T x + b \leq -1$$

Assume that we learned a new SVM model (SVM2) using the following constraints instead:
For all x in class 1:

$$W^T x + b \geq 0$$

For all x in class 0:

$$W^T x + b < 0$$

(i) If we compare the margin of SVM2 to that of SVM1, we can say that:

a. The margin increased
b. The margin decreased
c. The margin stayed the same
d. Impossible to tell

Solution: Decrease. If we set the threshold at 0 then there will be no margin and since this is a linearly separable dataset the margin will decrease.
1 mark for correct option and correct reason

(ii) Assume that we are using a new SVM, SVM3, which uses $\frac{W}{2}$ and $\frac{b}{2}$ where W and b are the parameters learned for SVM1. With these new parameters:
Are we guaranteed that SVM3 would not make any mistakes on the training data? (Recall that an SVM classifier determines the class based on the sign of $W^T x + b$, where x is the input.)

a. Yes
b. No

Solution: Yes. This is a linearly separable problem and everything that was higher than 0 before remains higher now and similarly for lower than 0.
1 mark for correct option and correct reason

(iii) How would the margin for SVM3 compare to the margin of SVM1?

a. The margin would increase
b. The margin would decrease
c. The margin would stay the same
d. Impossible to tell

Solution: The margin would increase. The margin is given by $\frac{2}{\mathbf{W}^T\mathbf{W}}$, and since $\mathbf{W}$ is divided by 2, the margin would increase.
1 mark for correct option and correct reason

(iv) The number of support vectors for SVM3 compared to SVM1 would (recall that support vectors are those inputs that are either exactly on the $+1$ or -1 planes or those points that are between these planes and the decision boundaries):

a. The number of support vectors would likely increase
b. The number of support vectors would likely decrease
c. The number of support vectors would likely stay the same
d. Impossible to tell

Solution: The number of support vectors would likely increase. All previous support vectors now lie between the margin and the decision line and there could only be new support vectors added.
1 mark for correct option and correct reason

12. **[3 marks]** Consider a CNN architecture with an input size of 24×24. The architecture consists of the following layers:

1. A Convolutional Layer with 64 filters, each of size 3×3, with a stride of 1 and no padding.
2. A Max Pooling Layer with a pool size of 3×3 and a stride of 2, with no padding.

For each layer, determine the following:

- The size of the feature map after the layer.
- The number of trainable parameters in the layer.

Provide your calculations and mention the total number of parameters for the entire architecture along with the final output size. Make any necessary assumptions clear.
Solution: Input Size: 24×24
Filter Size: 3×3
Number of Filters: 64
Stride: 1
Padding: 0 (no padding)
The size of the output feature map is calculated as:

$$\text{Output size} = \frac{\text{Input size} - \text{Filter size}}{\text{Stride}} + 1$$

For both height and width:

$$\text{Output size} = \frac{24 - 3}{1} + 1 = 22$$

Thus, the feature map size is 22×22 with a depth of 64. The trainable parameters include:

- **Weights:** Each filter has 3×3 weights, and there are 64 filters applied to the input depth of 1 (grayscale image).

$$\text{Weights per filter} = 3 \times 3 \times 1$$

$$\text{Total weights} = 3 \times 3 \times 1 \times 64 = 576$$

- **Biases:** Each filter has one bias term.

$$\text{Total biases} = 64$$

Total trainable parameters:

$$\text{Total parameters} = \text{Weights} + \text{Biases} = 576 + 64 = 640$$

Layer 2: Max Pooling Layer Input Size: $22 \times 22 \times 64$
Pool Size: 3×3
Stride: 2
Padding: 0 (no padding) **Feature Map Size** The size of the output feature map is calculated as:

$$\text{Output size} = \frac{\text{Input size} - \text{Pool size}}{\text{Stride}} + 1$$

For both height and width:

$$\text{Output size} = \frac{22 - 3}{2} + 1 = 10$$

Thus, the feature map size is 10×10 with a depth of 64.
Total Trainable Parameters Max pooling has no trainable parameters, as it simply performs a reduction operation.
Total parameters:

$$\text{Total parameters} = 0$$

Summary

Layer	Feature Map Size	Number of Trainable Parameters
Convolutional	$22 \times 22 \times 64$	640
Max Pooling	$10 \times 10 \times 64$	0

Total Number of Trainable Parameters=640+0
Layer 1 (Convolutional Layer): 1 mark (0.5 marks each), Layer 2 (Max Pooling Layer): 1 mark (0.5 marks each), Final Result: 1 mark (0.5 marks each)

13. **[4 marks]** Show that K-means is guaranteed to converge (to a local optimum).
solution:
2 mark for Assignment (I), 2 marks for Refitting step. Give step marking

To prove convergence of the K-means algorithm, we show that the loss function is guaranteed to decrease monotonically in each iteration until convergence for the *assignment step* and for the *refitting step*. Since the loss function is non-negative, the algorithm will eventually converge when the loss function reaches its (local) minimum.

Let $z = (z_1, \ldots, z_n)$ denote the cluster assignments for the n points.

(i) *Assignment step*
We can write down the original loss function $L(\mu)$ as follows:

$$L(\mu, z) = \sum_{i=1}^{n} \|x_i - \mu_{z_i}\|_2^2$$

Let us consider a data point x_i, and let z_i be the assignment from the previous iteration and z_i^* be the new assignment obtained as:

$$z_i^* \in \arg \min_{j \in \{1,\ldots,k\}} \|x_i - \mu_j\|_2^2$$

Let z^* denote the new cluster assignments for all the n points. The change in loss function after this assignment step is then given by:

$$L(\mu, z^*) - L(\mu, z) = \sum_{i=1}^{n} \left(\|x_i - \mu_{z_i^*}\|_2^2 - \|x_i - \mu_{z_i}\|_2^2 \right) \leq 0$$

The inequality holds by the rule z_i^* is determined, i.e. to assign x_i to the nearest cluster.

(ii) *Refitting step*
We can write down the original loss function $L(\mu)$ as follows:

$$L(\mu, z) = \sum_{j=1}^{k} \Big(\sum_{i:z_i=j} \|x_i - \mu_j\|_2^2 \Big)$$

Figure 6: Question 13

Let us consider the j^{th} cluster, and let μ_j be the cluster center from the previous iteration and μ_j^* be the new cluster center obtained as:

$$\mu_j^* = \frac{1}{|\{i : z_i = j\}|} \sum_{i:z_i=j} x_i$$

Let μ^* denote the new cluster centers for all the k clusters. The change in loss function after this refitting step is then given by:

$$L(\mu^*, z) - L(\mu, z) = \sum_{j=1}^{k} \Big(\big(\sum_{i:z_i=j} \|x_i - \mu_j^*\|_2^2 \big) - \big(\sum_{i:z_i=j} \|x_i - \mu_j\|_2^2 \big) \Big) \leq 0$$

The inequality holds because the update rule of μ_j^* essentially minimizes this quantity.

Figure 7: Question 13*

14. **[4 marks]** Consider a multilayer feed-forward neural network given in Figure. Let the learning rate be 0.5. Assume initial values of weights and biases as given in the table below. Train the network for the training tuples (1, 1, 0), where the last number is the target output. Show weight and bias updates by using a back-propagation algorithm (calculations are mandatory). Assume that the sigmoid activation function and mean square error are used in the network.

W13	W14	W23	W24	W35	W45	b3	b4	b5
0.5	0.2	-0.3	0.5	0.1	0.3	0.6	-0.4	0.8

1 mark for the output layer, 1 mark for the hidden layer and 2 marks for weight update. Students may have used the sigmoid activation function at different layers. Consider the approach, and if they are correct, give marks.

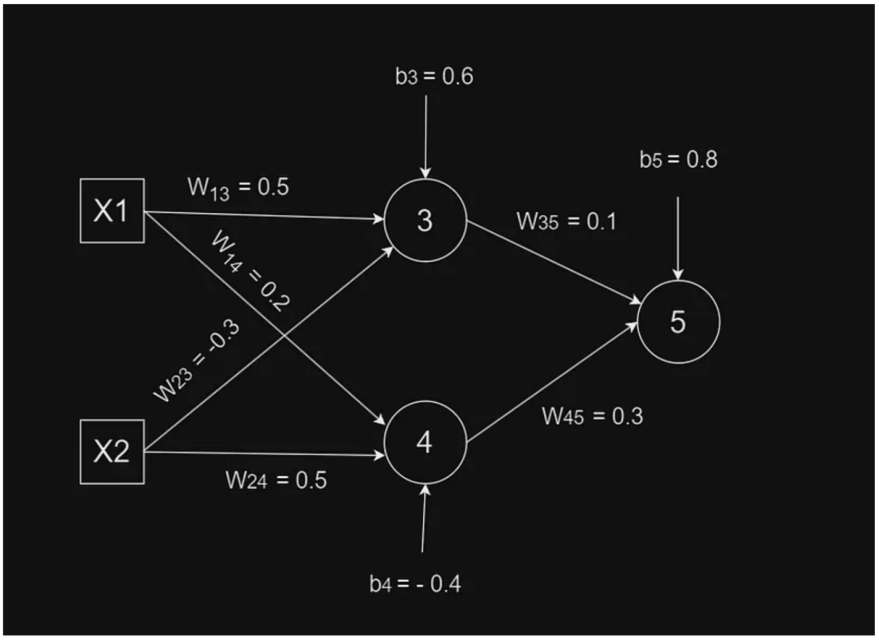


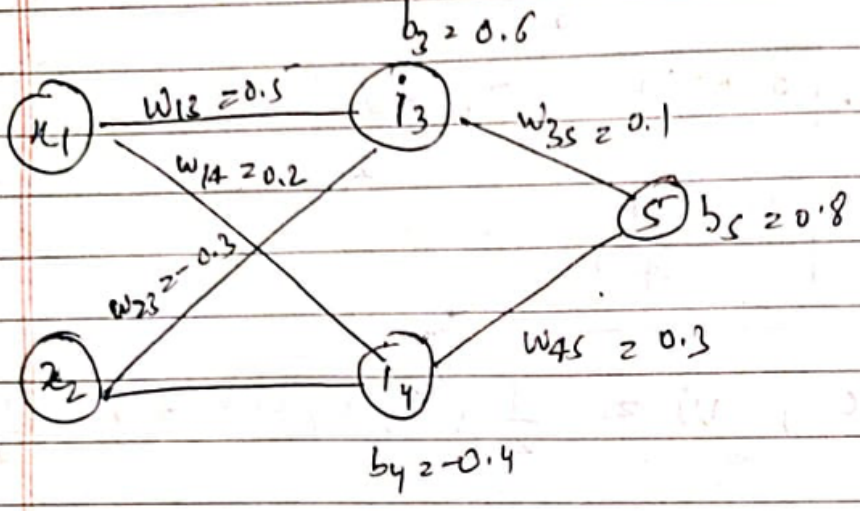
Figure 8: Question 14

DATE: __/__/____
PAGE: _____

Solution

output

Forward, $x_1 = 1$, $x_2 = 1$, $y = 0$



$I_3 = O_1 w_{13} + O_2 w_{23} + b_3$

$= 0.5 - 0.3 + 0.6$

$= 0.8$

$=$

$O_3 = \frac{1}{1 + e^{-0.8}} = 0.69$

$I_4 = O_1 w_{13} + O_2 w_{24} + b_4$

$= 0.2 + 0.5 - 0.4$

$= 0.3$

$O_4 = \frac{1}{1 + e^{-0.3}} = 0.57$

Output $\Rightarrow$ $I_5 = O_3 w_{35} + O_4 w_{45} + b_5$

$= 0.69\ 0.1 + 0.57\ 0.3 + 0.8$

$= 1.0913$

Good Write

Figure 9: Question 14

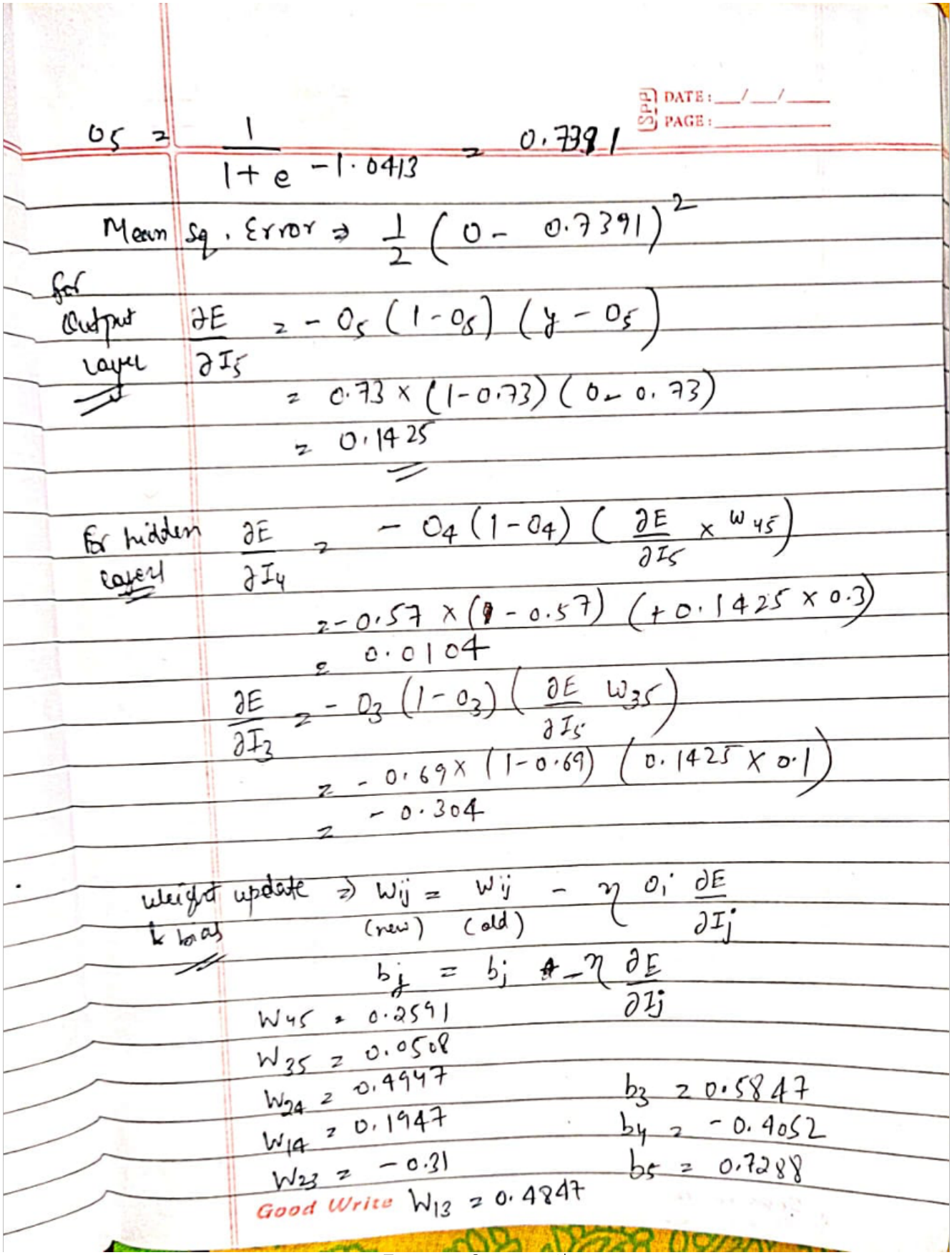


Figure 10: Question 14*