

# **Graded**

---

This document has 11 questions.

# Note to Learners

---

## Statement

The projection is treated as a vector in all the questions. If we wish to talk about the length of the projection, then that would be mentioned explicitly. Likewise, the residue after the projection is also treated as a vector. If we wish to talk about the length of the residue, that would be mentioned explicitly.

## Question-1 [0.5 point]

---

### Statement

Consider a point  $\mathbf{x} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$  and a line passing through the origin which is represented by the vector  $\mathbf{w} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$ . What can you say about the following quantities? (MSQ)

- (1) the projection of  $\mathbf{x}$  onto the line
- (2) the residue

### Options

**(a)**

The residue is equal to the zero vector.

**(b)**

The residue is equal to the vector  $\mathbf{x}$ .

**(c)**

The projection is the zero vector.

**(d)**

The projection is equal to the vector  $\mathbf{x}$ .

### Answer

(b), (c)

## Solution

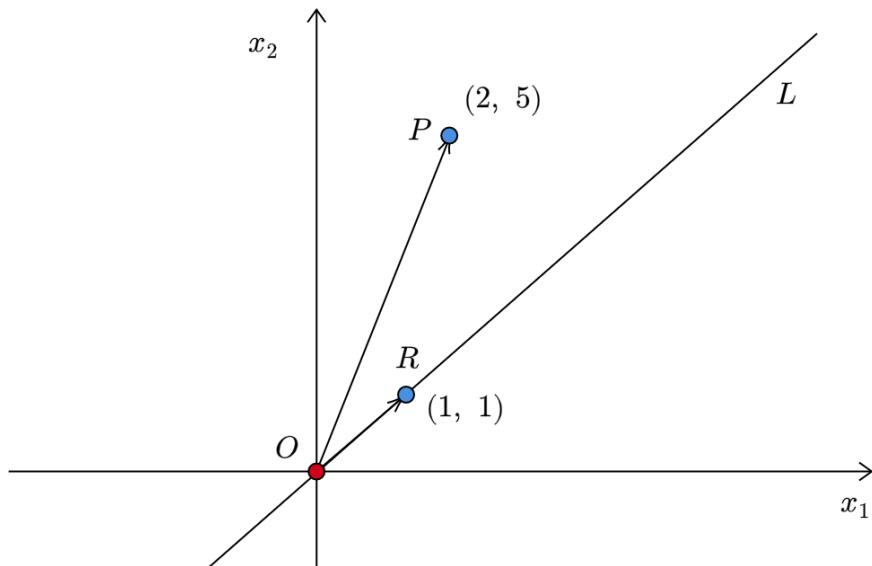
We have  $\mathbf{x}^T \mathbf{w} = 0$ . So, the projection is the zero vector. The residue is given by:

$$\mathbf{x} - (\mathbf{x}^T \mathbf{w})\mathbf{w} = \mathbf{x}$$

## Common data for questions (2) to (5)

### Statement

Consider a point  $P$  and a line  $L$  that passes through the origin  $O$ . The point  $R$  lies on the line.



We use the following notation:

$$\mathbf{w} = \overrightarrow{OR}$$

$$\mathbf{x} = \overrightarrow{OP}$$

## Question-2 [1 point]

---

### Statement

Consider the following statements:

Statement-1: The projection of  $\mathbf{x}$  on the line  $L$  is given by  $(\mathbf{x}^T \mathbf{w})\mathbf{w}$

Statement-2: The projection of  $\mathbf{x}$  on the line  $L$  is given by  $(\mathbf{x}^T \mathbf{w})\mathbf{x}$

Statement-3: The projection of  $\mathbf{x}$  on the line  $L$  is given by  $(\mathbf{x}^T \mathbf{x})\mathbf{w}$

Statement-4: The projection of  $\mathbf{x}$  on the line  $L$  is given by  $\mathbf{w}^T \mathbf{x}$

Which of the above statements is true?

### Options

**(a)**

Statement-1

**(b)**

Statement-2

**(c)**

Statement-3

**(d)**

Statement-4

**(e)**

None of these statements are true.

### Answer

(e)

### Solution

The projection of a point  $\mathbf{x}$  on a line  $\mathbf{w}$  is given by:

$$\frac{\mathbf{x}^T \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \mathbf{w}$$

This is the expression when  $\mathbf{w}$  does not have unit length. In this problem,  $\mathbf{w}$  does not have unit length. If  $\|\mathbf{w}\| = 1$ , then the expression becomes:

$$(\mathbf{x}^T \mathbf{w})\mathbf{w}$$

## **Question-3 [1 point]**

---

### **Statement**

Find the length of the projection of  $\mathbf{x}$  on the line  $L$ . Enter your answer correct to two decimal places.

### **Answer**

4.95

Range: [4.9, 5.0]

### **Solution**

The length of the projection is given by:

$$\frac{|\mathbf{x}^T \mathbf{w}|}{\|\mathbf{w}\|} = \frac{2 + 5}{\sqrt{2}} \approx 4.95$$

## Question-4 [1 point]

---

### Statement

Find the residue after projecting  $\mathbf{x}$  on the line  $L$ .

### Options

(a)

$$\begin{bmatrix} 3.5 \\ 3.5 \end{bmatrix}$$

(b)

$$\begin{bmatrix} -1.5 \\ 1.5 \end{bmatrix}$$

(c)

$$\begin{bmatrix} 2 \\ 5 \end{bmatrix}$$

(d)

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

### Answer

(b)

### Solution

The residue is given by:

$$\mathbf{x} - \frac{\mathbf{x}^T \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \mathbf{w} = \begin{bmatrix} 2 \\ 5 \end{bmatrix} - \frac{7}{2} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1.5 \\ 1.5 \end{bmatrix}$$

## Question-5 [1 point]

### Statement

Find the reconstruction error for this point. Enter your answer correct to two decimal places.

### Answer

4.5

Range: [4.4, 4.6]

### Solution

The reconstruction error is given by the square of the length of the residue. If the residue is  $\mathbf{x}'$ , then:

$$(\mathbf{x}')^T \mathbf{x}' = (-1.5)^2 + 1.5^2 = 4.5$$

Programming based solution. This is to be used only to verify the correctness of the calculations. The added benefit is that you get used to NumPy.

```
1 import numpy as np
2
3 x = np.array([2, 5])
4 w = np.array([1, 1])
5 w = w / np.linalg.norm(w)
6
7 # Projection
8 proj = (x @ w) * w
9 print(f'Projection = {np.linalg.norm(proj)}')
10 # Residue
11 res = x - proj
12 print(f'Residue = {res}')
13 # Reconstruction error
14 recon = res @ res
15 print(f'Reconstruction error = {recon}')
```

## Question-6 [0.5 point]

### Statement

Consider the following images of points in 2D space. The red line segments in one of the images represent the lengths of the residues after projecting the points on the line  $L$ . Which image is it?

Image-1

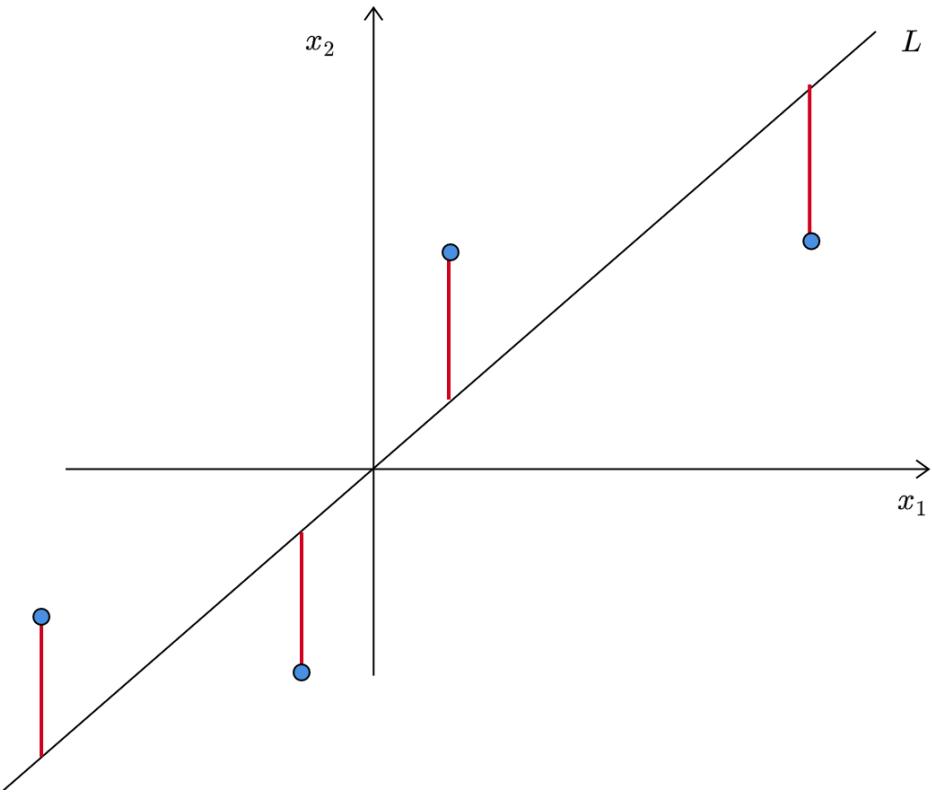
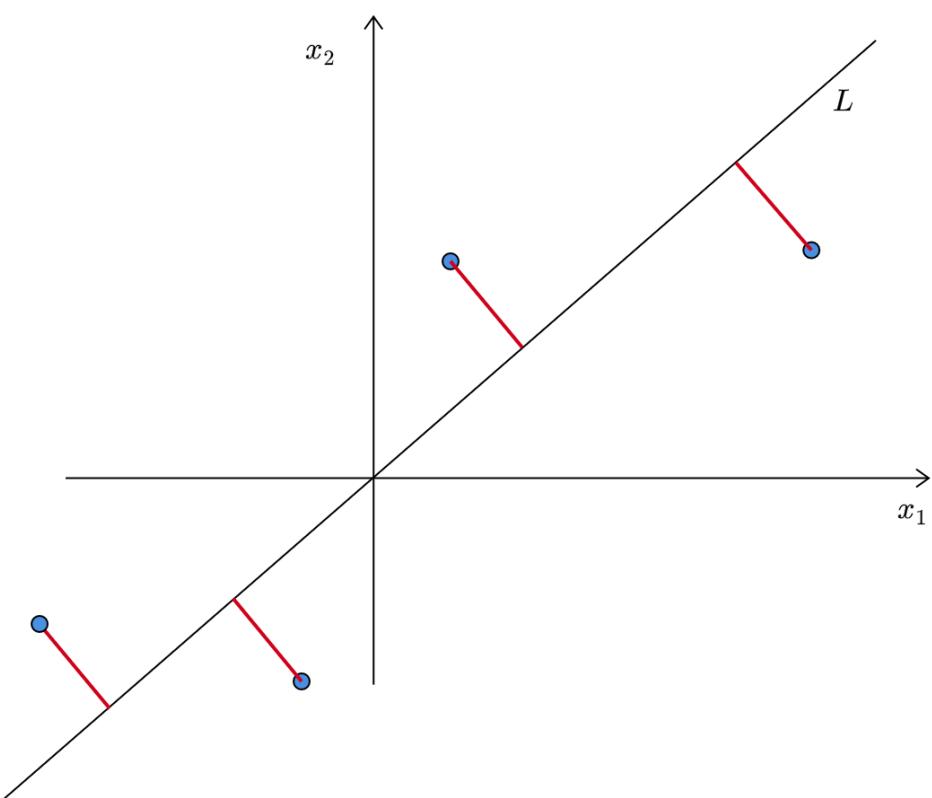


Image-2



## **Options**

**(a)**

Image-1

**(b)**

Image-2

## **Answer**

(b)

## **Solution**

The residue after the projection should be perpendicular to the line. Note that by projection we mean the orthogonal projection of a point on a line. The projection of a point on a line is one of the proxies for that point on the line, in fact it is the "best" possible proxy. But every proxy does not become a projection. The projection of a point on a line is unique.

## **Question-7 [1 point]**

---

### **Statement**

Consider a dataset that has 1000 samples, where each sample belongs to  $\mathbb{R}^{30}$ . PCA is run on this dataset and the top 4 principal components are retained, the rest being discarded. If it takes one unit of memory to store a real number, find the percentage decrease in storage space of the dataset by moving to its compressed representation. Enter your answer correct to two decimal places; it should lie in the range [0, 100].

### **Answer**

86.27

Range: [86.2, 86.3]

### **Solution**

$$\text{Original space} = 1000 \times 30 = 30000$$

$$\text{Compressed space} = 1000 \times 4 + 4 \times 30 = 4120$$

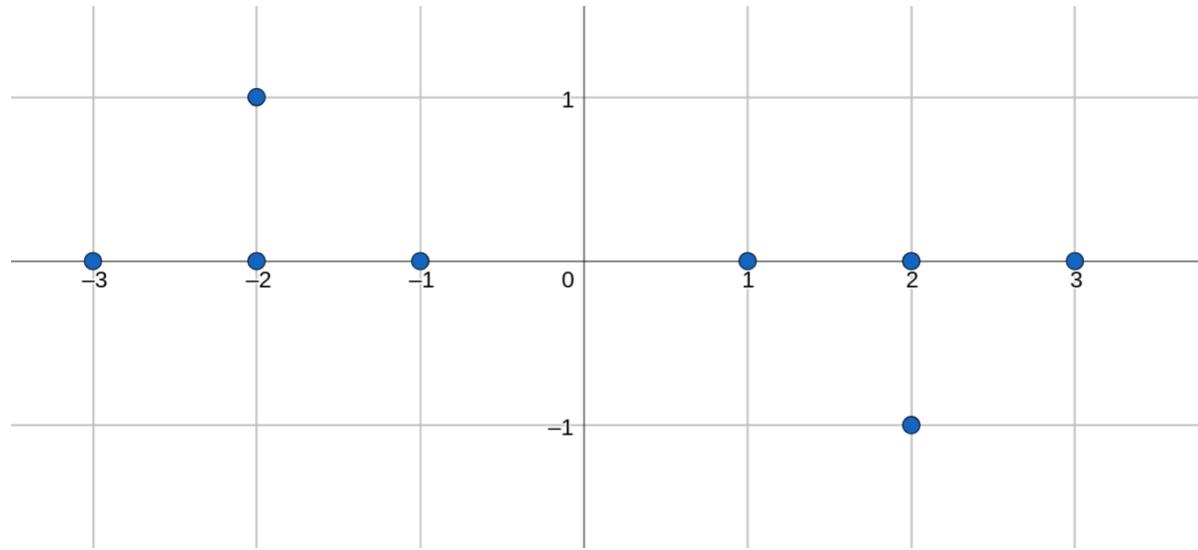
$$\text{Percentage decrease in space} = \frac{30000 - 4120}{30000} \times 100 \approx 86.27$$

## Common Data for questions (8) to (9)

---

### Statement

Consider a dataset that has 8 points all of which belong to  $\mathbb{R}^2$ :



## Question-8 [1 point]

---

### Statement

Find the covariance matrix of this dataset.

### Options

(a)

$$\begin{bmatrix} 4.5 & -0.5 \\ -0.5 & 0.25 \end{bmatrix}$$

(b)

$$\begin{bmatrix} 36 & -4 \\ -4 & 2 \end{bmatrix}$$

(c)

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

(d)

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

### Answer

(a)

### Solution

Let us first arrange the data in the form of a  $n \times d$  matrix. Here,  $n = 8$  and  $d = 2$ :

$$\mathbf{X} = \begin{bmatrix} -3 & 0 \\ -2 & 0 \\ -2 & 1 \\ -1 & 0 \\ 1 & 0 \\ 2 & 0 \\ 2 & -1 \\ 3 & 0 \end{bmatrix}$$

The covariance matrix is therefore:

$$\frac{1}{n} \cdot \mathbf{X}^T \mathbf{X} = \begin{bmatrix} 4.5 & -0.5 \\ -0.5 & 0.25 \end{bmatrix}$$

## Question-9 [1 point]

### Statement

If PCA is run on this dataset, find the variance of the dataset along the first principal component. The eigenvectors of the covariance matrix are given below:

$$\begin{bmatrix} -0.993 \\ 0.115 \end{bmatrix}, \quad \begin{bmatrix} -0.115 \\ -0.993 \end{bmatrix}$$

Recall that the first principal component is the most important. Enter your answer correct to two decimal places.

### Answer

4.55

Range: (4.5, 4.6)

### Solution

If  $(\lambda_k, \mathbf{w}_k)$  is the  $k^{th}$  eigenpair for  $\mathbf{C}$ , we have:

$$\lambda_k = \mathbf{w}_k^T \mathbf{C} \mathbf{w}_k$$

Of the two eigenvalues, the larger one is the answer.

Verification for these two problems:

```
1 import numpy as np
2
3 X = np.array([[-3,  0],
4               [-2,  0],
5               [-2,  1],
6               [-1,  0],
7               [1,  0],
8               [2,  0],
9               [2, -1],
10              [3,  0]])
11
12 C = X.T @ X / X.shape[0]
13 print(f'Covariance matrix = {C}')
14 eigval, eigvec = np.linalg.eigh(C)
15 print(f'Variance = {eigval[-1]}')
```

A more detailed version. The variance of the dataset along the  $j^{th}$  principal component is  $\sigma_j^2$  and is given by:

$$\sigma_j^2 = \frac{1}{n} \cdot \sum_{i=1}^n (\mathbf{x}_i^T \mathbf{w}_j)^2$$

$$= \mathbf{w}_j^T \left( \frac{1}{n} \cdot \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{w}_j$$

$$= \mathbf{w}_j^T \mathbf{C} \mathbf{w}_j$$

$$= \lambda_j$$

So, the variance along the  $j^{th}$  principal component is the  $j^{th}$  largest eigenvalue of the covariance matrix.

## Question-10 [1 point]

### Statement

Consider a dataset of 100 points all of which lie in  $\mathbb{R}^5$ . The eigenvalues of the covariance matrix are given below:

3.4, 2.8, 0.5, 0.4, 0.01

If we run the PCA algorithm on this dataset and retain the top- $k$  principal components, what is a good choice of  $k$ ? Use the heuristic that was discussed in the lectures.

### Answer

4

### Solution

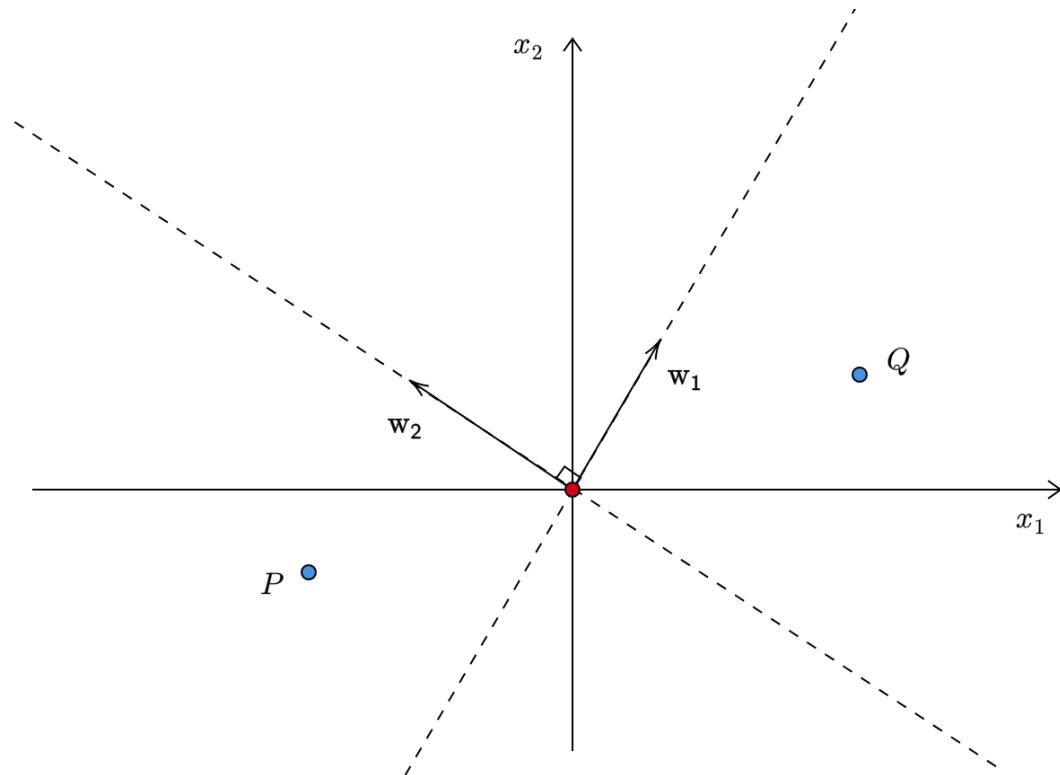
The top- $k$  principal components should capture 95% of the variance. Here is a code snippet to answer this question:

```
1 L = [3.4, 2.8, 0.5, 0.4, 0.01]
2
3 den = sum(L)
4 for k in range(1, len(L) + 1):
5     num = sum(L[: k])
6     if num / den >= 0.95:
7         break
8 print(k)
```

## Question-11 [1 point]

### Statement

PCA is run on a dataset that has 2 features. The resulting principal components are  $\mathbf{w}_1$  and  $\mathbf{w}_2$ . We represent the points in 2D space in terms of this new coordinate system made up of the principal components. The first coordinate corresponds to  $\mathbf{w}_1$  and the second to  $\mathbf{w}_2$ . In such a scenario, what would be the sign of the coordinates for the points  $P$  and  $Q$ ?



### Options

(a)

$$P : (-ve, -ve)$$

(b)

$$P : (-ve, +ve)$$

(c)

$$Q : (+ve, +ve)$$

(d)

$$Q : (+ve, -ve)$$

### Answer

(b), (d)

## **Solution**

Each vector  $\mathbf{w}$  is associated with a line perpendicular to it. This line divides the space into two halves. The basic idea is to identify the sign of the half-planes into which the line perpendicular to the vector  $\mathbf{w}$  divides the space.

# Week 2 Graded assignment

---

## Common data for Questions 1 and 2

---

A function  $k$  is defined as follows.

$$k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$$
$$k(x_1, x_2) = x_1^T x_2$$

### Question 1

---

#### Statement

Is  $k$  a valid kernel?

#### Options

(a)

Yes

(b)

No

#### Answer

(a)

#### Solution

Let  $\phi$  be an identity transformation that is

$$\phi : \mathbb{R}^d \rightarrow \mathbb{R}^d$$
$$\phi(x) = x$$

It is clear by the definition of given kernel that

$$k(x_1, x_2) = \phi(x_1)^T \phi(x_2)$$

It implies that  $k$  is a valid kernel.

### Question 2

---

If  $k$  is the valid kernel, we apply it to the three-dimensional dataset to run the kernel PCA. Select the correct options.

## Options

(a)

We cannot run the PCA as  $k$  is not a valid kernel.

(b)

It will be the same as PCA with no kernel.

(c)

It will be the same as the polynomial transformation of degree 2 and then run the PCA.

(d)

It will be the same as PCA with a third-degree polynomial kernel.

## Answer

(b)

## Solution

We have seen (in question 1) that  $k$  corresponds to the identity transformation. It implies that applying kernel and running PCA is same as standard PCA on the given dataset.

## Question 3

---

### Statement

Consider ten data points lying on a curve of degree two in a two-dimensional space. We run a kernel PCA with a polynomial kernel of degree two on the same data points. Choose the correct options.

## Options

(a)

The transformed data points will lie on a 5-dimensional subspace of  $\mathbb{R}^6$ .

(b)

The transformed data points will lie on a 6-dimensional subspace of  $\mathbb{R}^{10}$

(c)

There will be some  $w \in \mathbb{R}^6$  that all of the data points are orthogonal to.

(d)

There will be some  $w \in \mathbb{R}^{10}$  that all of the data points are orthogonal to.

## Answer

(a), (c)

## Solution

Since we are applying the polynomial kernel of degree two on the 2D dataset, the dataset will be transformed into a 6D feature space. (verify)

And the dataset is given to lying on a curve of degree two, the transformed dataset will live in the linear subspace of  $\mathbb{R}^6$ . and therefore, there will be some  $w \in \mathbb{R}^6$  that all of the data points are orthogonal to.

## Question 4

---

### Statement

Which of the following matrices can not be appropriate matrix  $K = X^T X$  for some data matrix  $X$ ?

### Options

(a)

$$\begin{bmatrix} 1 & 8 \\ 8 & -1 \end{bmatrix}$$

(b)

$$\begin{bmatrix} 1 & 8 \\ 8 & 1 \end{bmatrix}$$

(c)

$$\begin{bmatrix} 1 & 8 \\ -8 & 1 \end{bmatrix}$$

(d)

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

## Answer

(a), (b) and (c)

## Solution

We know that  $K$  matrix must be symmetric and positive semi-definite.

All the given matrices are symmetric.

We need to check whether they are positive semi-definite or not.

For that, we will check the eigenvalues of the matrices and if all the eigenvalues are non-negative, then the matrix is positive semi-definite.

### Option (a)

$$A = \begin{bmatrix} 1 & 8 \\ 8 & -1 \end{bmatrix}$$

Let  $\lambda$  be the eigenvalue of  $A$ , then

$$\begin{aligned} |A - \lambda I| &= 0 \\ \begin{vmatrix} 1 - \lambda & 8 \\ 8 & -1 - \lambda \end{vmatrix} &= 0 \\ (1 - \lambda)(-1 - \lambda) &= 64 \\ \lambda^2 - 1 &= 64 \\ \lambda &= \pm\sqrt{65} \end{aligned}$$

Since  $A$  has a non-negative eigenvalue,  $A$  is not a positive semi-definite matrix.

Similarly, check for all the options.

## Question 5

---

### Statement

A function  $k$  is defined as

$$\begin{aligned} k : \mathbb{R}^2 \times \mathbb{R}^2 &\rightarrow \mathbb{R} \\ k(x_1, x_2) &= (x_1^T x_2)^2 \end{aligned}$$

Is  $k$  a valid kernel?

### Options

(a)

Yes

(b)

No

### Answer

(a)

### Solution

The given function is

$$\begin{aligned} k : \mathbb{R}^2 \times \mathbb{R}^2 &\rightarrow \mathbb{R} \\ k(x_1, x_2) &= (x_1^T x_2)^2 \end{aligned}$$

Let  $x_1 = [a_1, a_2]^T$  and  $x_2 = [b_1, b_2]^T$  then

$$\begin{aligned}
k(x_1, x_2) &= ([a_1, a_2][b_1, b_2]^T)^2 \\
&= (a_1 b_1 + a_2 b_2)^2 \\
&= a_1^2 b_1^2 + 2a_1 b_1 a_2 b_2 + a_2^2 b_2^2 \\
&= [a_1^2, \sqrt{2}a_1 a_2, a_2^2][b_1^2, \sqrt{2}b_1 b_2, b_2^2]^T \\
&= \phi(x_1)^T \phi(x_2)
\end{aligned}$$

where  $\phi : \mathbb{R}^2 \rightarrow \mathbb{R}$  such that  $\phi([a_1, a_2]^T) = [a_1^2, \sqrt{2}a_1 a_2, a_2^2]^T$

It means there exists a transformation mapping  $\phi$  such that  $k(x_1, x_2) = \phi(x_1)^T \phi(x_2)$ . Therefore,  $k$  is a valid kernel.

## Question 6

---

### Statement

Kernel PCA was run on the four data points  $[1, 2]^T$ ,  $[2, 3]^T$ ,  $[2, -3]^T$ , and  $[4, 4]^T$  with the polynomial kernel of degree 2. What will be the shape of the matrix  $K$ ? Notations are used as per lectures.

### Options

(a)

$2 \times 2$

(b)

$4 \times 4$

(c)

$6 \times 6$

(d)

None of the above

### Answer

(b)

### Solution

The  $K$  matrix is defined as  $X^T X$  where  $X$  is a data matrix of shape  $(d, n)$ . That is  $K$  matrix is of shape  $(n, n)$  where  $n$  is a number of examples.

It is given that  $n = 4$ . Therefore, shape of  $K$  matrix is  $(4, 4)$

## Question 7

---

## Statement

Find the element at the index  $(2, 3)$  of the matrix  $K$  defined in Question 6. Take the points in the same order.

## Options

(a)

-4

(b)

16

(c)

13

(d)

196

## Answer

(b)

## Solution

The polynomial kernel of degree 2 is given by

$$k(x_1, x_2) = (x_1^T x_2 + 1)^2$$

The  $(2, 3)$ th element of  $K$  matrix will be  $k(x_2, x_3)$ .

$$k(x_2, x_3) = ([2, 3][2, -3]^T + 1)^2 = (-5 + 1)^2 = 16$$

## Question 8

---

## Statement

A dataset containing 200 examples in four-dimensional space has been transformed into higher dimensional space using the polynomial kernel of degree two. What will be the dimension of transformed feature space?

## Answer

15 (No range required)

## Solution

Let the features be  $x_1, x_2, x_3$ , and  $x_4$ . After the transformation of degree two, features will be  $1, x_1, x_2, x_3, x_4, x_1x_2, x_1x_3, x_1x_4, x_2x_3, x_2x_4, x_3x_4, x_1^2, x_2^2, x_3^2$ , and  $x_4^2$ . So, the dimension of transformed feature space will be 15.

## Question 9

---

### Statement

Let  $x_1, x_2, \dots, x_n$  be  $d$ -dimensional data points ( $d > n$ ) and  $X$  be the matrix of shape  $d \times n$  containing the data points. The  $k^{th}$  largest eigenvalue and corresponding unit eigenvector of  $X^T X$  is  $\lambda$  and  $\alpha_k$ , respectively. What will be the projection of  $x_i$  on the  $k^{th}$  principal component?

### Options

(a)

$$x_i^T \alpha_k$$

(b)

$$\frac{x_i^T \alpha_k}{\lambda}$$

(c)

$$\frac{x_i^T X \alpha_k}{\sqrt{\lambda}}$$

(d)

$$\frac{x_i^T X \alpha_k}{\sqrt{n\lambda}}$$

### Answer

(c)

### Solution

If the  $k^{th}$  largest eigenvalue and corresponding unit eigenvector of  $X^T X$  is  $\lambda$  and  $\alpha_k$ , respectively, the  $k^{th}$  principal component will be  $\frac{X \alpha_k}{\sqrt{\lambda}}$ .

Therefore, the projection of the point  $x_i$  on the  $k^{th}$  principal component will be  $\frac{x_i^T X \alpha_k}{\sqrt{\lambda}}$ .

## Question 10

---

### Statement

Let  $k_1$  and  $k_2$  be two valid kernels. Is  $3k_1 + 5k_2$  a valid kernel?

### Options

**(a)**

Yes

**(b)**

No

## Answer

(a)

## Solution

Let  $k_1$  and  $k_2$  be two valid kernels defined on  $\mathbb{R}^d \times \mathbb{R}^d$ , it implies that they satisfies the following two properties

(1)  $k_1$  and  $k_2$  are symmetric functions that is

$$\begin{aligned} k_1(x_1, x_2) &= k_1(x_2, x_1) \quad \forall x_1, x_2 \in \mathbb{R}^d \quad \text{and} \\ k_2(x_1, x_2) &= k_2(x_2, x_1) \quad \forall x_1, x_2 \in \mathbb{R}^d \end{aligned}$$

(2) For any  $x \in \mathbb{R}^n$ , we have

$$\begin{aligned} x^T K_1 x &\geq 0 \quad \text{and} \\ x^T K_2 x &\geq 0 \end{aligned}$$

where,

$K_1 = [k_1(x_i, x_j)]_{n \times n}$  and  $K_2 = [k_2(x_i, x_j)]_{n \times n}$  are  $K$  matrices corresponding to kernels  $k_1$  and  $k_2$ , respectively.

Assume that  $k = 3k_1 + 5k_2$

To show:  $k(x_1, x_2) = k(x_2, x_1) \quad \forall x_1, x_2 \in \mathbb{R}^d$

and

$x^T K x \geq 0 \quad \forall x \in \mathbb{R}^n$ , where  $K = [k(x_i, x_j)]_{n \times n}$  is  $K$  matrices corresponding to kernel  $k$

Now,

$$\begin{aligned} k(x_1, x_2) &= 3k_1(x_1, x_2) + 5k_2(x_1, x_2) \\ &= 3k_1(x_2, x_1) + 5k_2(x_2, x_1) \\ &= k(x_2, x_1) \end{aligned}$$

and

$$\begin{aligned} x^T K x &= x^T (3K_1 + 5K_2)x \\ &= 3x^T K_1 x + 5x^T K_2 x \geq 0 \end{aligned}$$

It implies that  $3k_1 + 5k_2$  is a valid kernel.

# **Graded**

---

This document has 10 questions.

# Question-1

---

## Statement

What would be the correct relationship among the following three quantities?:

- (1)  $\sum_{i=1}^n \|x_i - \mu_{z_i^t}\|^2$ ,
- (2)  $\sum_{i=1}^n \|x_i - \mu_{z_i^{t+1}}^t\|^2$  and
- (3)  $\sum_{i=1}^n \|x_i - \mu_{z_i^{t+1}}^{t+1}\|^2$

where  $\mu_{z_i^t}$  and  $\mu_{z_i^{t+1}}^{t+1}$  refer to means of cluster  $z_i$  in iterations  $t$  and  $t + 1$  respectively. And  $\mu_{z_i^{t+1}}^t$  is the mean of the cluster  $z_i$  where  $x_i$  is going to move in the next (i. e.,  $(t + 1)^{th}$ ) iteration.

## Options

**(a)**

(1) > (2) < (3)

**(b)**

(1) < (2) < (3)

**(c)**

(1) > (2) > (3)

**(d)**

(1) < (2) > (3)

## Answer

(c)

## Solution

The first quantity represents the value of objective function in iteration  $t$ . the third quantity represents the value of objective function in iteration  $t + 1$ . The second quality represents an intermediate quantity which captures the distance of each data point from the mean that they will be moving towards, in the  $t+1$  iteration. Since in every iteration, the reassignment happens only if a data point has found a closer mean, (3) will be lesser than (1). Further, since every point will want to move towards a closer cluster center in the subsequent iteration, the value of (2) will be between (1) and (3).

## Question-2

---

### Statement

Consider that in an iteration  $t$  of Lloyd's algorithm, the partition configuration  $(P^t)$  is  $z_1^t, z_2^t, \dots, z_n^t$  where each  $z_i^t \in \{1, 2, \dots, k\}$ . Assume that the algorithm does not converge in iteration  $t$ , and hence some re-assignment happens, thus updating the partition configuration in the next iteration  $(P^{t+1})$  to  $z_1^{t+1}, z_2^{t+1}, \dots, z_n^{t+1}$ . How can we say that partition configuration  $P^{t+1}$  is better than  $P^t$ ?

### Options

(a)

The value of the objective function for  $P^{t+1}$  should be more than that for  $P^t$

(b)

The value of the objective function for  $P^{t+1}$  should be lesser than that for  $P^t$

(c)

The value of the objective function for  $P^{t+1}$  and  $P^t$  should be same.

### Answer

(b)

### Solution

Since in every iteration, the reassignment happens only if a data point has found a closer mean,  $P^{t+1}$  will be lesser than  $P^t$ .

## **Question-3**

---

### **Statement**

With respect to Lloyd's algorithm, choose the correct statements:

### **Options**

**(a)**

At the end of k-means, the objective function settles in a local minima and reaching global minima may not be guaranteed.

**(b)**

At the end of k-means, the objective function always settles in the global minima.

**(c)**

The clusters produced by K-means are optimal.

**(d)**

If the resources are limited and the data set is huge, it will be good to prefer K-means over K-means ++.

**(e)**

In practice, k should be as large as possible.

### **Answer**

(a), (d)

### **Solution**

(a), (b) K-means may not always settle in a global minima.

(c) Finding optimal clusters is an NP-hard problem. K-means provides approximate clusters.

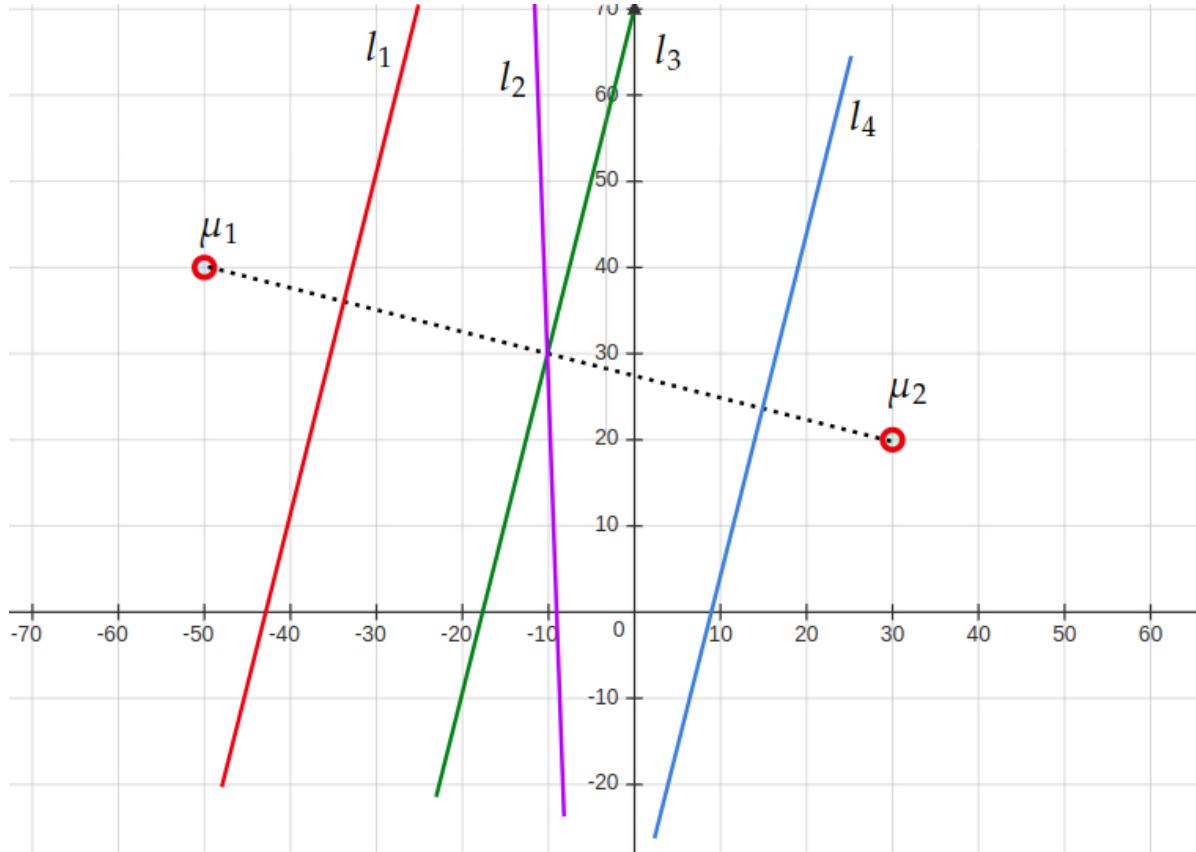
(d) If the dataset is huge, the elaborate initialization step in K-means++ will take a lot of time.

(e) In practice, k should neither be very small nor very large, because in both these cases, we may not be able to uncover groupings present in the data.

## Question-4

### Statement

Consider two cluster centres  $\mu_1$  and  $\mu_2$  corresponding to two clusters  $C_1$  and  $C_2$  as shown in the below image. Consider four half spaces represented by lines  $l_1, l_2, l_3$  and  $l_4$ . Where would the data points falling in cluster  $C_1$  lie?



### Options

(a)

To the left of  $l_1$

(b)

Between  $l_1$  and  $l_2$

(c)

Between  $l_3$  and  $l_4$

(d)

To the left of  $l_3$

**(e)**

To the left of  $l_2$

## **Answer**

(d)

## **Solution**

Half-spaces are perpendicular bisectors of the line joining the cluster centers.

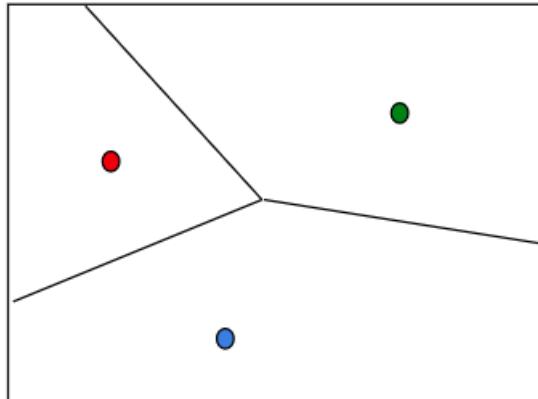
## Question-5

### Statement

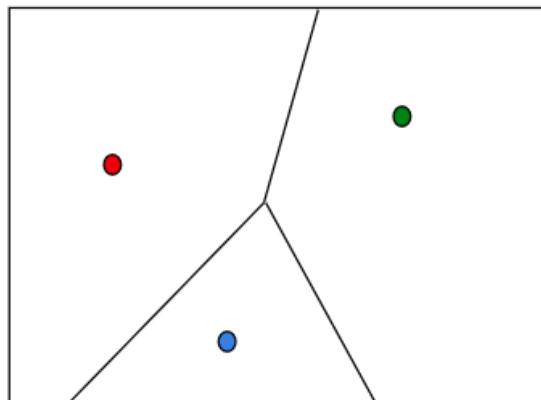
Which of the following best represents a valid voronoi diagram for K-means algorithm with K = 3?  
(The dots represent the cluster centres of respective clusters.)

### Options

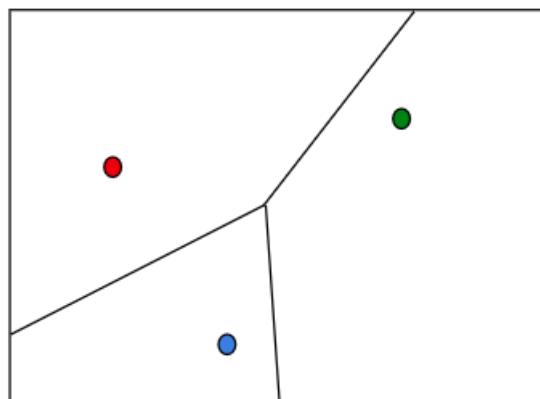
(a)



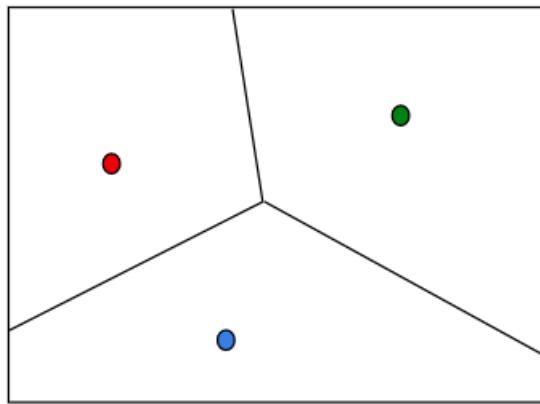
(b)



(c)



(d)



## Answer

(d)

## Solution

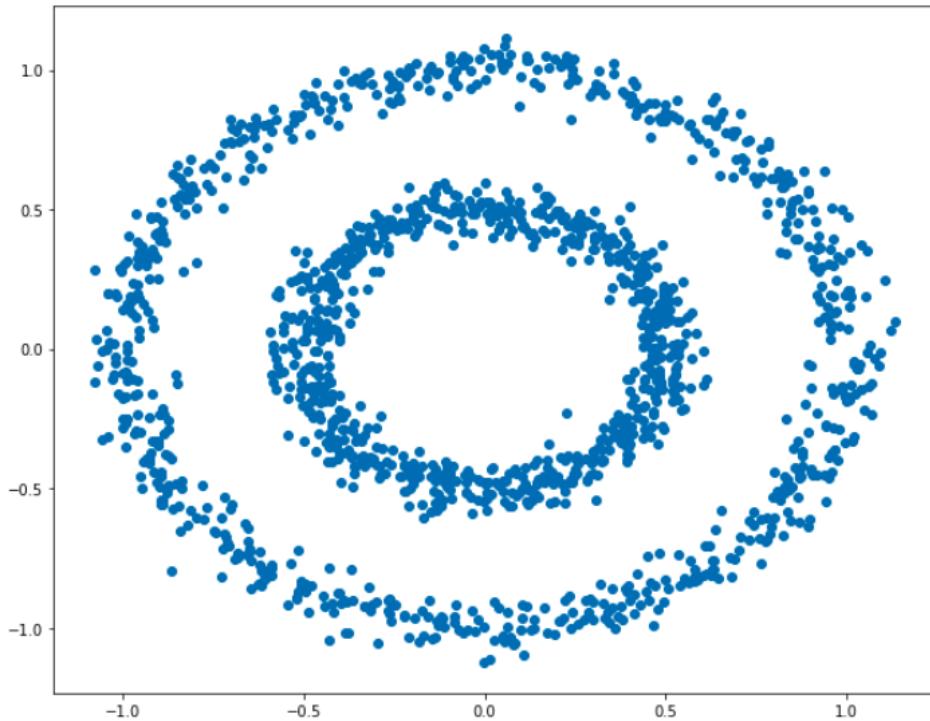
Half-spaces are perpendicular bisectors of the line joining the cluster centers.

## Question-6

---

### Statement

Consider the following data points:

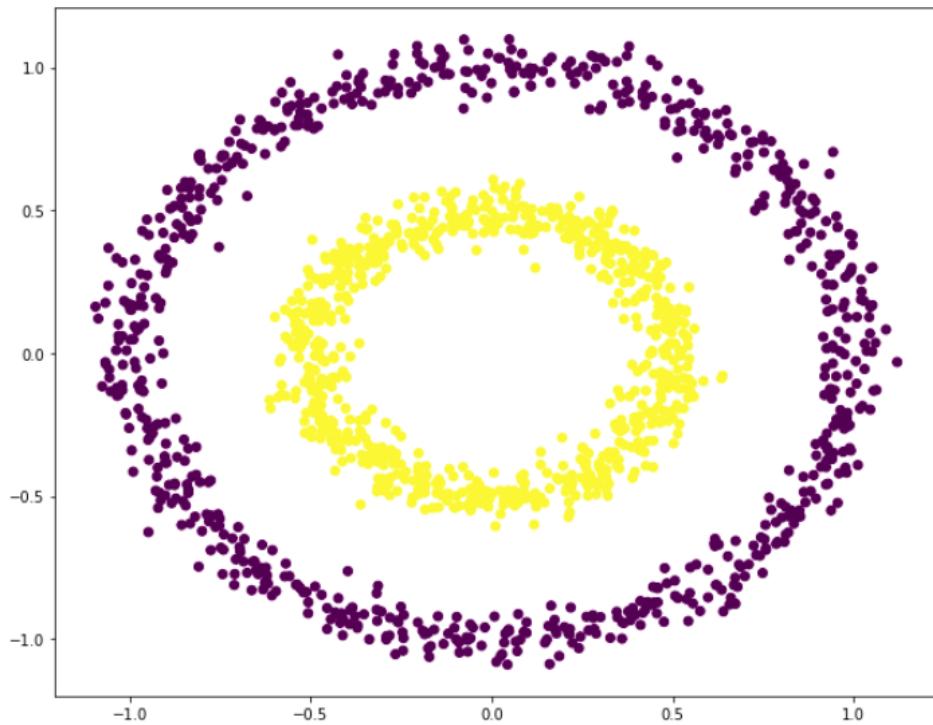


Assume that K-means is applied on this data with  $k = 2$ . Which of the following are expected to be the clusters produced?

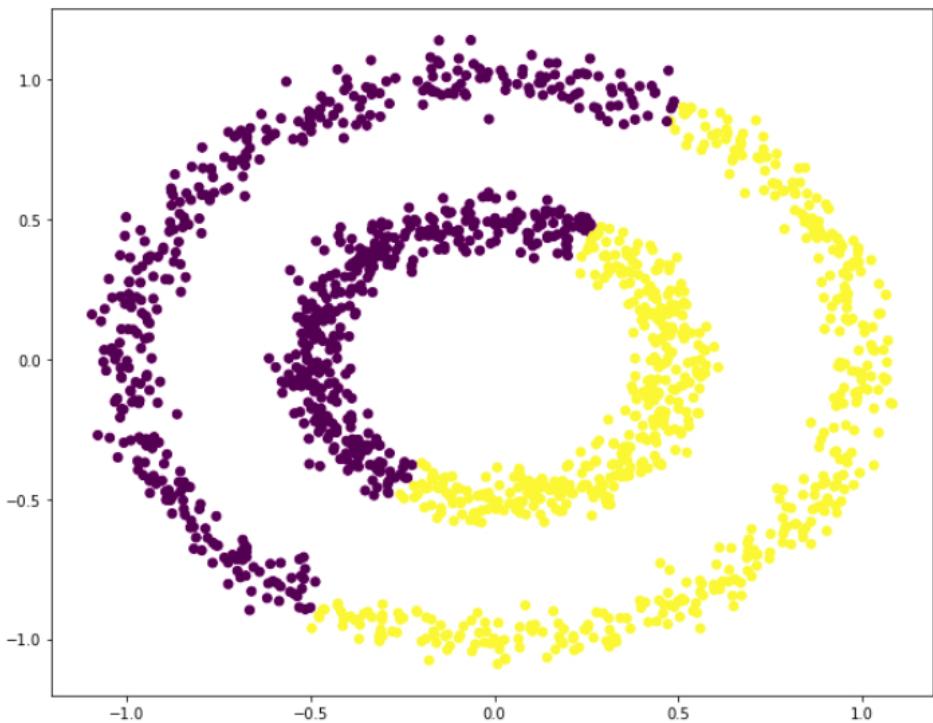
Note: Different colors represent different clusters.

### Options

(a)



(b)



## Answer

(b)

## Solution

Half-spaces are perpendicular bisectors of the line joining the cluster centers.

In the given data, in case of option (a), the cluster centers will coincide, which is something does not happen as a result of applying k-means.

## Question-7

---

### Statement

Assume that in the initialization step of k-means++, the squared distances from the closest mean for 10 points  $x_1, x_2, \dots, x_{10}$  are: 25, 67, 89, 24, 56, 78, 90, 85, 35, 95. Which point has the highest probability of getting chosen as the next mean and how much will that probability be?

### Options

(a)

$x_4, 0.24$

(b)

$x_4, 0.037$

(c)

$x_{10}, 0.95$

(d)

$x_{10}, 0.1475$

### Answer

(d)

### Solution

$$25 + 67 + 89 + 24 + 56 + 78 + 90 + 85 + 35 + 95 = 644$$

$$\text{Probability for } x_{10} = 95/644 = 0.1475$$

$$\text{Probability for } x_4 = 24/644 = 0.037$$

## Question-8

---

### Statement

Consider 7 data points  $x_1, x_2, \dots, x_7$ :  $\{(0, 4), (4, 0), (2, 2), (4, 4), (6, 6), (5, 5), (9, 9)\}$ . Assume that we want to form 3 clusters from these points using K-Means algorithm. Assume that after first iteration, clusters  $C1, C2, C3$  have the following data points:

$$C1: \{(2,2), (4,4), (6,6)\}$$

$$C2: \{(0,4), (4,0)\}$$

$$C3: \{(5,5), (9,9)\}$$

After second iteration, which of the clusters is the data point (2, 2) expected to move to?

### Options

(a)

$C_1$

(b)

$C_2$

(c)

$C_3$

(d)

Can't say, it is not deterministic.

### Answer

(b)

### Solution

$$C1: (4,4), C2: (2,2), C3: (7,7)$$

$C2$  mean is the closest to (2,2) with distance 0.

## **Question-9**

---

### **Statement**

Which of the following statements are True?

1. K-means is extremely sensitive to cluster center initializations.
2. Bad initialization can lead to poor convergence speed.
3. Bad initialization can lead to bad overall clustering.

### **Options**

**(a)**

1 and 3

**(b)**

1 and 2

**(c)**

2 and 3

**(d)**

1, 2, and 3

### **Answer**

(d)

### **Solution**

1. Different cluster center initializations may result in different clusters produced by k-means.
2. Some initializations may take more time to converge.
3. Some initializations may converge either in a local minima rather than global minima.

## **Question-10**

---

### **Statement**

If the data set has two features  $x_1$  and  $x_2$ , which of the following are true for K means clustering with k = 3?

1. If  $x_1$  and  $x_2$  have a correlation of 1, the cluster centres will be in a straight line.
2. If  $x_1$  and  $x_2$  have a correlation of 0, the cluster centres will be in straight line.

### **Options**

**(a)**

1

**(b)**

2

**(c)**

None of these. Correlation does not affect cluster centres' position.

### **Answer**

(a)

### **Solution**

If  $x_1$  and  $x_2$  have a correlation of 1, all data points will lie along a line.

Hence the cluster centers will also lie along the same line.

# **Graded**

---

This document has 11 questions.

# Note to Learners

---

## Statement

For all questions involving the Bernoulli distribution, the parameter  $p$  is  $P(x = 1)$ .

## Question-1

---

### Statement

Consider a dataset that has 10 zeros and 5 ones. What is the likelihood function if we assume a Bernoulli distribution with parameter  $p$  as the probabilistic model?

### Options

(a)

$$p^{15}$$

(b)

$$(1 - p)^{15}$$

(c)

$$p^{10} \cdot (1 - p)^5$$

(d)

$$p^5 \cdot (1 - p)^{10}$$

### Answer

(d)

### Solution

We shall use the i.i.d. assumption. If  $x_i$  is the random variable corresponding to the  $i^{th}$  data-point, we have:

$$P(x_i = 1) = p$$

$x_i$  and  $x_j$  are independent for  $i \neq j$  and they are identically distributed. The likelihood is therefore the product of 15 terms, five of which correspond to ones and the rest to zeros:

$$L(p; D) = p^5 \cdot (1 - p)^{10}$$

## Question-2

---

### Statement

In the previous question, what is the estimate of  $\hat{p}_{ML}$ ? Enter your answer correct to two decimal places.

### Answer

0.33

Range: [0.32, 0.34]

### Solution

The estimate is the fraction of ones:

$$\hat{p} = \frac{5}{15} = \frac{1}{3} \approx 0.33$$

## Question-3

---

### Statement

Consider a dataset that has a single feature ( $x$ ). The first column in the table below represents the value of the feature, the second column represents the number of times it occurs in the dataset.

$x$	Frequency
-1	1
0	1
2	4
4	2
5	2

If we use a Gaussian distribution to model this data, find the maximum likelihood estimate of the mean.

### Options

(a)

2

(b)

0

(c)

2.5

(d)

The mean cannot be computed as the variance of the Gaussian is not explicitly specified.

### Answer

(c)

### Solution

$$\hat{\mu}_{ML} = \frac{-1 + 0 + 4 \times 2 + 2 \times 4 + 2 \times 5}{10} = 2.5$$

## Question-4

---

### Statement

Consider a beta prior for the parameter  $p$  of a Bernoulli distribution:

$$p \sim \text{Beta}(3, 2)$$

The dataset has 15 ones and 10 zeros. What is the posterior?

### Options

(a)

$$\text{Beta}(3, 2)$$

(b)

$$\text{Beta}(30, 17)$$

(c)

$$\text{Beta}(18, 12)$$

(d)

$$\text{Beta}(17, 11)$$

### Answer

(c)

### Solution

Since the beta distribution is a conjugate prior of the Bernoulli distribution, the posterior is also a beta distribution. Specifically, if the prior is  $\text{B}(\alpha, \beta)$  and the dataset has  $n_1$  ones and  $n_0$  zeros, then the posterior:

$$\text{Beta}(\alpha + n_1, \beta + n_0)$$

In this problem,  $\alpha = 3, \beta = 2, n_1 = 15, n_0 = 10$ .

## Question-5

---

### Statement

In the previous question, we use the expected value of the posterior as a point-estimate for the parameter of the Bernoulli distribution. What is  $\hat{p}$ ? Enter your answer correct to two decimal places.

### Answer

0.6

Range: [0.59, 0.61]

### Solution

The expected value of a beta distribution with parameters  $\alpha$  and  $\beta$  is:

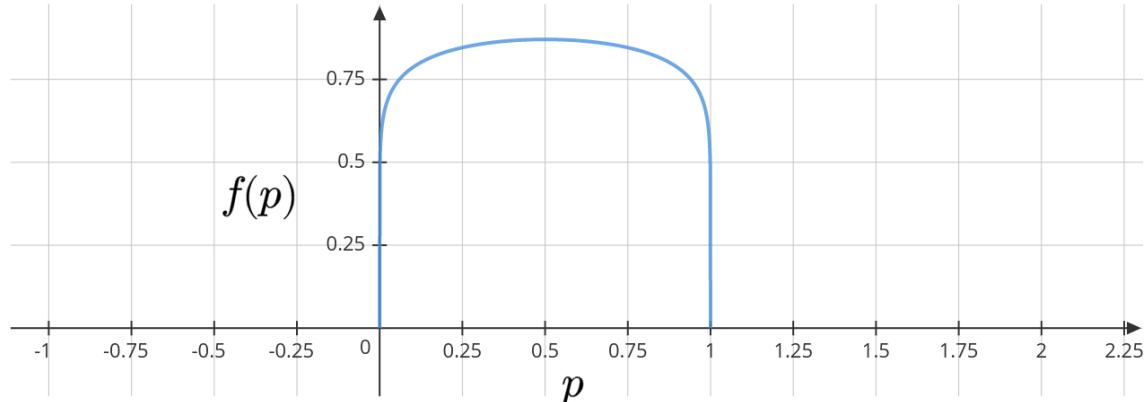
$$\frac{\alpha}{\alpha + \beta}$$

For the posterior, we have  $\alpha = 18, \beta = 12$ .

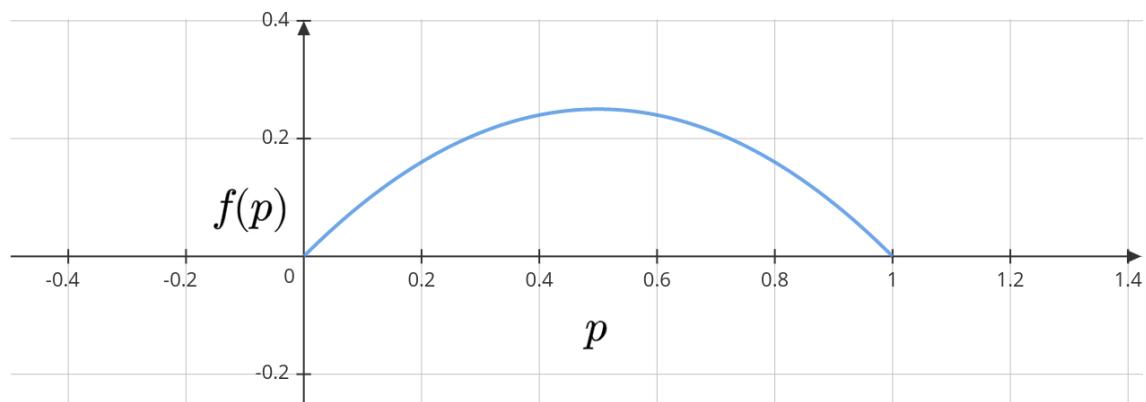
## Question-6

### Statement

Consider the following prior distribution (Beta) of the parameter  $p$  of a Bernoulli distribution:



After observing 10 data-points, the following is the posterior distribution:



Ignore the values on the Y-axis and just focus on the shapes of the distributions. Which of the following could correspond to the observed data?

### Options

(a)

$$\{1, 1, 1, 0, 1, 1, 0, 1, 1, 1\}$$

(b)

$$\{0, 1, 0, 0, 0, 1, 0, 0, 0, 0\}$$

**(c)**

$$\{1, 1, 0, 1, 0, 0, 0, 1, 1, 0\}$$

## **Answer**

(c)

## **Solution**

The prior encodes the belief that coin is somewhat unbiased. The posterior seems to have made that belief stronger. So, the data should have been something that strengthens the belief in the prior, meaning, an equal number of ones and zeros.

## Common Data for questions (7) to (9)

---

### Statement

We wish to fit a GMM with  $K = 2$  for a dataset having 4 points. At the beginning of the  $t^{th}$  time step of the EM algorithm, we have  $\theta^{(t)}$  as follows:

$$\begin{aligned}\pi_1 &= 0.3, & \pi_2 &= 0.7 \\ \mu_1 &= 2, & \sigma_1^2 &= 1 \\ \mu_2 &= 3, & \sigma_2^2 &= 1\end{aligned}$$

The density of the points given a particular mixture is given to you for all four points.  $f$  is the density of a Gaussian.

$x_i$	$f(x_i   z_i = 1)$	$f(x_i   z_i = 2)$
1	0.242	0.054
2	0.399	0.242
3	0.242	0.399
4	0.054	0.242

Use three decimal places for all quantities throughout the questions.

## Question-7

---

### Statement

What is the value of  $\lambda_k^i$  for  $i = 1$  and  $k = 2$  after the E-step? Enter your answer correct to three decimal places.

### Answer

0.342

Range: [0.33, 35]

### Solution

From Bayes rule, we have:

$$\lambda_k^i = P(z_i = k \mid x_i) = \frac{P(z_i = k) \cdot f(x_i \mid z_i = k)}{f(x_i)}$$

So:

$$\lambda_2^1 = \frac{0.7 \times 0.054}{0.7 \times 0.054 + 0.3 \times 0.242} \approx 0.342$$

## Question-8

---

### Statement

If we pause the algorithm at this stage (after the E-step) and use the  $\lambda_k^i$  values to do a hard-clustering, what would be the cluster assignment? We use the following rule to come up with cluster assignments:

$$z_i = \operatorname{argmax}_k \lambda_k^i$$

The answer is in the form of a vector:  $\mathbf{z} = [z_1 \ z_2 \ z_3 \ z_4]^T$ .

### Options

(a)

$$[1 \ 1 \ 1 \ 1]^T$$

(b)

$$[2 \ 2 \ 2 \ 2]^T$$

(c)

$$[1 \ 1 \ 2 \ 2]^T$$

(d)

$$[1 \ 2 \ 2 \ 2]^T$$

### Answer

(d)

### Solution

We need to compute the table of  $\lambda_k^i$  values from which we can read off the cluster assignments.

$x_i$	$\lambda_1^i$	$\lambda_2^i$	$z_i$
1	0.658	0.342	1
2	0.414	0.586	2
3	0.206	0.794	2
4	0.087	0.912	2

## Question-9

---

### Statement

What is the value of  $\mu_1$  after the M-step? Enter your answer correct to three decimal places.

### Answer

1.796

Range: [1.7, 1.9]

### Solution

Now for  $\mu_1$ :

$$\mu_1 = \frac{\sum_{i=1}^4 \lambda_1^i x_i}{\sum_{i=1}^4 \lambda_1^i}$$

Which is:

$$\mu_1 = \frac{0.658 \times 1 + 0.414 \times 2 + 0.206 \times 3 + 0.087 \times 4}{0.658 + 0.414 + 0.206 + 0.087} \approx 1.796$$

## Question-10

---

### Statement

A GMM is fit for a dataset with 5 points. At some time-step in the EM algorithm, the following are the values of  $\lambda_k^i$  for all points in the dataset for the  $k^{th}$  mixture after the E-step:

$$\begin{aligned}\lambda_k^1 &= 0.3 \\ \lambda_k^2 &= 0.1 \\ \lambda_k^3 &= 0.4 \\ \lambda_k^4 &= 0.8 \\ \lambda_k^5 &= 0.2\end{aligned}$$

What is the estimate of  $\pi_k$  after the M-step? Enter your answer correct to two decimal places.

### Answer

0.36

Range: [0.35, 0.37]

### Solution

$$\pi_k = \frac{1}{5} \cdot \sum_{i=1}^5 \lambda_k^i = \frac{0.3 + 0.1 + 0.4 + 0.8 + 0.2}{5} = \frac{1.8}{5} = 0.36$$

## Question-11

---

### Statement

What is the value of the following expression after the E-step at time-step  $t$  in the EM algorithm? There are 100 data-points and 3 mixtures.

$$\sum_{i=1}^{100} \sum_{k=1}^3 \lambda_k^i$$

### Options

(a)

3

(b)

100

(c)

103

(d)

300

(e)

1

(f)

The answer depends on the time-step  $t$  we are at

### Answer

(b)

### Solution

We know that the  $\lambda_k^i$  values should sum to 1 for each data-point. Since there are 100 data-points, the expression should sum to 100.

# Graded Assignment

---

## Note:

1. In the following assignment,  $X$  denotes the data matrix of shape  $(d, n)$  where  $d$  and  $n$  are the number of features and samples, respectively.
2.  $x_i$  denotes the  $i^{th}$  sample and  $y_i$  denotes the corresponding label.
3.  $w$  denotes the weight vector (parameter) in the linear regression model.

## Question 1

---

### Statement

An ML engineer comes up with two different models for the same dataset. The performances of these two models on the training dataset and test dataset are as follows:

- Model 1: Training error = 0.9; Test error = 0.1
- Model 2: Training error = 0.1; Test error = 10

Which model you would select?

### Options

(a)

Model 1

(b)

Model 2

### Answer

(a)

### Solution

In model 1, the test error is very low compared to model 2 even though the training error is high in model 1. We choose model 1 as it worked well on unseen data.

## Question 2

---

### Statement

Consider a model  $h$  for a given  $d$ -dimensional training data points  $\{x_1, x_2, \dots, x_n\}$  and corresponding labels  $\{y_1, y_2, \dots, y_n\}$  as follows:

$$h : \mathbb{R}^d \rightarrow \mathbb{R}$$
$$h(x_i) = \bar{y}$$

where  $\bar{y}$  is the average of all the labels. Which of the following error function will always give the zero training error for the above model?

## Options

(a)

$$\sum_{i=1}^n (h(x_i) - y_i)^2$$

(b)

$$\sum_{i=1}^n |(h(x_i) - y_i)|$$

(c)

$$\sum_{i=1}^n (h(x_i) - y_i)$$

(d)

$$\sum_{i=1}^n (h(x_i) - y_i)^3$$

## Answer

(c)

## Solution

The sum of squared error and absolute error will give zero error only if predicted values are the same as actual values for all the examples.

But for option (3), we have

$$\begin{aligned}\sum_{i=1}^n (h(x_i) - y_i) &= \sum_{i=1}^n (\bar{y} - y_i) \\ &= n\bar{y} - \sum_{i=1}^n y_i \\ &= n\bar{y} - n\bar{y} = 0\end{aligned}$$

This error function will give zero error for the above model.

---

## Common data for questions 3 and 4

---

Consider the following dataset with one feature  $x_1$ :

$x_1$	label ( $y$ )
-1	5
0	7
1	6

## Question 3

---

## Statement

We want to fit a linear regression model of the form  $y_i = w^T x_i$ . Assume that the initial weight vector is  $w = [2]$ . What will be the weight after one iteration using the gradient descent algorithm assuming the squared loss function? Assume the learning rate is  $\eta = 1$ .

## Answer

-4 No range is required

## Solution

At iteration  $t = 0$ , we have  $w^0 = [2]$

At  $t = 1$ , we have

$$w^1 = w^0 - \eta[2XX^T w^0 - 2Xy]$$

Here  $w^0 = [2]$

$X = [-1, 0, -1]$

$y = [5, 7, 6]^T$

Put the values and we get

$$w^1 = [-4]$$

## Question 4

### Statement

If we stop the algorithm at the weight calculated in question 1, what will be the prediction for the data point  $x_1 = -2$ ?

## Answer

8 No range is required

## Solution

The model is given as

$$y_i = -4x_i$$

at  $x = -2$ ,

$$y = (-4)(-2) = 8$$

## Question 5

## Statement

Assume that  $w^t$  denotes the updated weight after the  $t^{th}$  iteration in the stochastic gradient descent. At each step, a random sample of the data points is considered for weight update. What will be the final weight  $w$  after  $T$  iterations?

## Options

(a)

$$w^T$$

(b)

$$w^1 + w^2 + \dots + w^T$$

(c)

$$\frac{1}{T} \sum_{i=1}^T w^i$$

(d)

any of the  $w^t$

## Answer

(c)

## Solution

The final weight is given by the average of all the weights updated in all the iterations. That is why option (c) is correct.

## Common data for questions 6 and 7

---

Kernel regression with a polynomial kernel of degree two is applied on a data set  $\{X, y\}$ . Let the weight vector be

$$w = X[0.3, 1.6, 4.2, -0.5, 0.9]$$

## Question 6

---

### Statement

Which data point plays the most important role in predicting the outcome for an unseen data point? Write the data point index as per matrix  $X$  assuming indices start from 1.

### Answer

3, No range is required

## Solution

Since  $w$  is written as  $X[0.3, 1.6, 4.2, -0.5, 0.9]$ , the data point which is associated with the highest weight (coefficient) will have the most importance. The third data point is associated with the highest coefficient (4.2) therefore, the third data point has the highest importance.

## Question 7

---

What will be the prediction for the data point  $[0, 0, 0, 0, 0]^T$ ?

### Answer

6.5 No range is required

## Solution

The polynomial kernel of degree 2 is given by

$$k(x_i, x_j) = (x_i^T x_j + 1)^2$$

The coefficient  $\alpha$  vector is given as  $[0.3, 1.6, 4.2, -0.5, 0.9]$ .

The prediction for a point  $x_{\text{test}}$  is given by

$$\alpha_1 k(x_{\text{test}}, x_1) + \alpha_2 k(x_{\text{test}}, x_2) + \dots + \alpha_n k(x_{\text{test}}, x_n)$$

Since,  $x_{\text{test}} = [0, 0, 0, 0, 0]^T$

$$x_{\text{test}}^T x_j = [0, 0, 0, 0, 0]^T x_j = 0 \quad \forall j$$

That is

$$k(x_{\text{test}}, x_j) = 1 \quad \forall j$$

Therefore the prediction will be

$$\alpha_1 + \alpha_2 + \dots + \alpha_n = 0.3 + 1.6 + 4.2 - 0.5 + 0.9 = 6.5$$

## Question 8

---

### Statement

If  $w^*$  be the solution to the optimization problem of the linear regression model, which of the following expression is always correct?

### Options

(a)

$$y^T X^T w^* = 0$$

**(b)**

$$(y - X^T w^*)^T (X^T w^*) = 0$$

**(c)**

$$(X^T w^*)(X^T w^*) = 0$$

**(d)**

$$y - X^T w^* = 0$$

## Answer

(b)

## Solution

We know that  $X^T w^*$  is the projection of labels  $y$  on the subspace spanned by the features that is  $(y - X^T w^*)$  will be orthogonal to  $X^T w^*$ . For details, check the lecture 5.4.

## Question 9

### Statement

The gradient descent with a constant learning rate of  $\eta = 1$  for a convex function starts oscillating around the local minima. What should be the ideal response in this case?

### Options

**(a)**

Increase the value of  $\eta$

**(b)**

Decrease the value of  $\eta$

## Answer

(b)

## Solution

One possible reason of oscillation is that the weight vector jumps the local minima due to greater step size ( $\eta$ ). That is if we decrease the value of  $\eta$ , the weight vector may not jump the local minima and the GD will converge to that local minima.

## Question 10

---

## **Statement**

Is the following statement true or false?

Error in the linear regression model is assumed to have constant variance.

## **Options**

**(a)**

True

**(b)**

False

## **Answer**

(a)

## **Solution**

We make the assumption in the regression model that the error follows gaussian distribution with zero mean and a constant variance.

# **Graded**

---

This document has 10 questions.

# Question-1

---

## Statement

Assume that for a certain linear regression problem involving 4 features, the following weight vectors produce an equal amount of mean square error:

$$w_1 = [2, 2, 3, 1]$$

$$w_2 = [1, 1, 3, 1]$$

$$w_3 = [3, 2, 4, 1]$$

$$w_4 = [1, 2, 1, 1]$$

Which of the weight vector is likely to be chosen by ridge regression?

## Options

(a)

$$w_1$$

(b)

$$w_2$$

(c)

$$w_3$$

(d)

$$w_4$$

## Answer

(d)

## Solution

$$\text{Total error} = \text{MSE} + \lambda ||w||^2$$

If MSE for all the given weights is same, the weight vector whose squared length is the least will be chosen by Ridge Regression.

## Question-2

---

### Statement

Assuming that in the constrained version of ridge regression optimization problem, following are the weight vectors to be considered, along with the mean squared error (MSE) produced by each:

$$w_1 = [2, 2, 3, 1], \text{ MSE} = 3$$

$$w_2 = [1, 1, 3, 1], \text{ MSE} = 5$$

$$w_3 = [3, 2, 4, 1], \text{ MSE} = 8$$

$$w_4 = [1, 2, 1, 1], \text{ MSE} = 9$$

If the value of  $\theta$  is 13, which of the following weight vectors will be selected as the final weight vector by ridge regression?

Note:  $\theta$  is as per lectures. That is,  $\|w\|^2 \leq \theta$

### Options

(a)

$w_1$

(b)

$w_2$

(c)

$w_3$

(d)

$w_4$

### Answer

(b)

### Solution

We need to minimize MSE such that  $\|w\|^2 \leq \theta$

$$\|w_1\|^2 = 18, \|w_2\|^2 = 12, \|w_3\|^2 = 30, \|w_4\|^2 = 7$$

$$\|w\|^2 \leq 13 \text{ for } w_2 \text{ and } w_4.$$

However, the MSE for  $w_2$  is lesser than  $w_4$ .

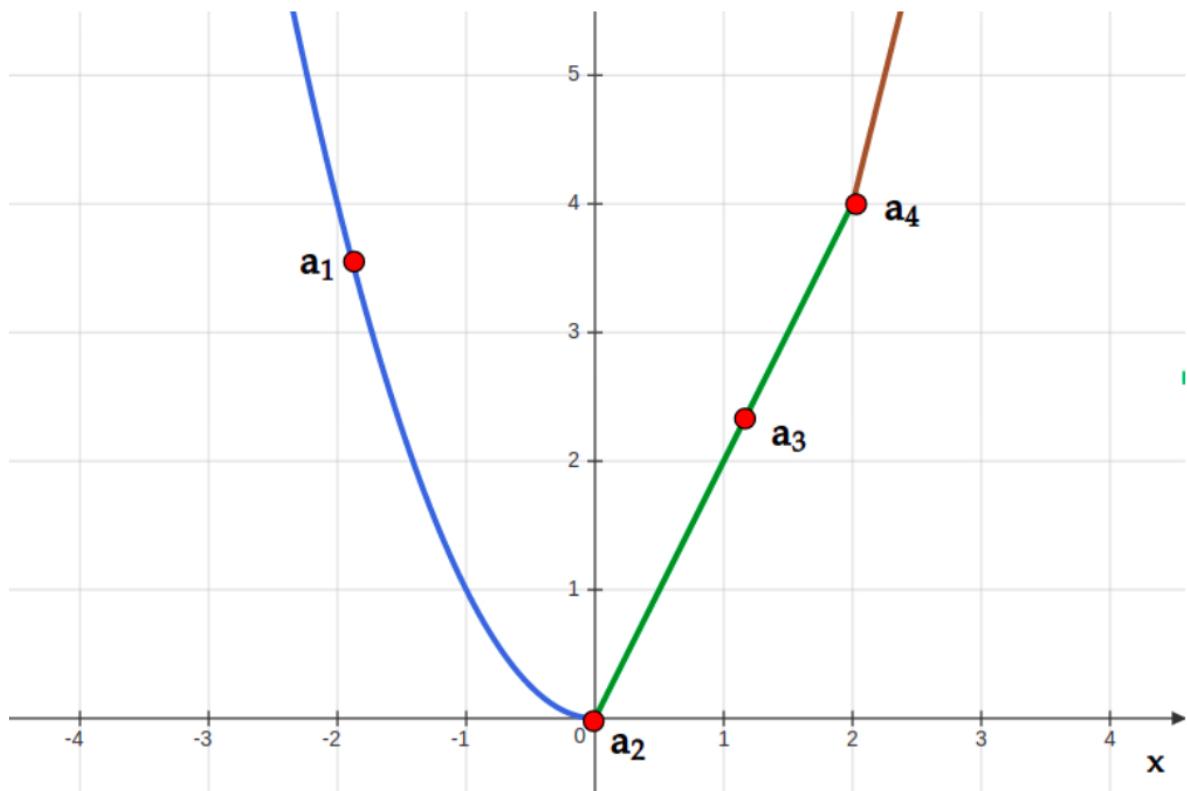
Hence,  $w_2$  will be chosen.

## Question-3

### Statement

Consider the following piece-wise function as shown in the image:

$$y(x) = \begin{cases} x^2 & x \leq 0 \\ 2x & 0 \leq x \leq 2 \\ 4x - 4 & 2 \leq x \end{cases}$$



How many sub-gradients are possible at points  $a_1, a_2, a_3$  and  $a_4$ ?

### Options

(a)

$a_1 : \text{Many}, a_2 : \text{One}, a_3 : \text{Many}, a_4 : \text{One}$

(b)

$a_1 : \text{One}, a_2 : \text{Many}, a_3 : \text{Many}, a_4 : \text{One}$

(c)

$a_1 : \text{One}, a_2 : \text{Many}, a_3 : \text{One}, a_4 : \text{Many}$

**(d)**

$a_1 : \text{Many}$ ,  $a_2 : \text{One}$ ,  $a_3 : \text{One}$ ,  $a_4 : \text{Many}$

## Answer

(c)

## Solution

$a_1$  lies on the part of the function which is differentiable. For a differentiable function (subpart), only one sub-gradient is possible which is the gradient itself.

$a_2$  lies at the intersection of two  $x^2$  and  $2x$ . The function is not differentiable at this point (as left slope is different from right slope). Hence there are multiple sub-gradients possible at  $a_2$ .

$a_3$  lies on the part of the function which is differentiable. For a differentiable function (subpart), only one sub-gradient is possible which is the gradient itself.

$a_4$  lies at the intersection of two  $2x$  and  $4x - 4$ . The function is not differentiable at this point (as left slope is different from right slope). Hence there are multiple sub-gradients possible at  $a_2$

## **Question-4**

---

### **Statement**

For a data set with 1000 data points and 50 features, 10-fold cross-validation will perform validation of how many models?

### **Options**

**(a)**

10

**(b)**

50

**(c)**

1000

**(d)**

500

### **Answer**

(a)

### **Solution**

In 10-fold cross validation, the data will be divided into 10 parts. In each of ten iterations, a model will be built using nine of these parts and the remaining part will be used for validation. Hence, in total, ten models will be validated.

## **Question-5**

---

### **Statement**

For a data set with 1000 data points and 50 features, assume that you keep 80% of the data for training and remaining 20% of the data for validation during k-fold cross-validation. How many models will be validated during cross-validation?

### **Options**

**(a)**

80

**(b)**

20

**(c)**

5

**(d)**

4

### **Answer**

(c)

### **Solution**

If 20% of the data is used for validation, that means, 1/5th part is used for validation, which means, 5-fold cross validation is being performed. In each iteration, one model will be validated. Hence, total 5 models will be validated.

## **Question-6**

---

### **Statement**

For a data set with 1000 data points and 50 features, how many models will be trained during Leave-One-Out cross-validation?

### **Options**

**(a)**

1000

**(b)**

50

**(c)**

5000

**(d)**

20

### **Answer**

(a)

### **Solution**

In leave one out cross-validation, only one data point is used for validation in each iteration, and the remaining  $n-1$  data points are used for training. Hence a total of  $n = 1000$  models will be trained.

## **Question-7**

---

### **Statement**

The mean squared error of  $\hat{w}_{ML}$  will be small if

### **Options**

**(a)**

The eigenvalues of  $XX^T$  are small.

**(b)**

The eigenvalues of  $(XX^T)^{-1}$  are large.

**(c)**

The eigenvalues of  $XX^T$  are large.

**(d)**

The eigenvalues of  $(XX^T)^{-1}$  are small.

### **Answer**

(c), (d)

### **Solution**

Mean Squared error of  $\hat{w}_{ML} = \sigma^2 \text{trace}(XX^T)^{-1}$ . Trace of a matrix = sum of eigenvalues.

If the eigenvalues of  $XX^T$  are large, the eigenvalues of  $(XX^T)^{-1}$  will be small. Hence, trace will be small and in turn MSE will be small.

## **Question-8**

---

### **Statement**

The eigenvalues of a  $3 \times 3$  matrix  $A$  are 2, 5 and 1. What will be the eigenvalues of the matrix  $A^{-1}$

### **Options**

**(a)**

4, 25, 1

**(b)**

2, 5, 1

**(c)**

0.5, 0.2, 1

**(d)**

0.6, 0.9, 0.1

### **Answer**

(c)

### **Solution**

If the eigenvalues of  $A$  are  $a, b$  and  $c$ , then the eigenvalues of  $A^{-1}$  will be  $1/a, 1/b$  and  $1/c$ .



# **Graded**

---

This document has 10 questions.

# Question-1

---

## Statement

We have a dataset of 1000 points for a classification problem using  $k$ -NN algorithm. Now consider the following statements:

S1: If  $k = 10$ , it is enough if we store any 10 points in the training dataset.

S2: If  $k = 10$ , we need to store the entire dataset.

S3: The number of data-points that we have to store increases as the size of  $k$  increases.

S4: The number of data-points that we have to store is independent of the value of  $k$ .

## Options

(a)

S1 and S3 are true statements

(b)

S2 and S4 are true statements

(c)

S1 alone is a true statement

(d)

S3 alone is a true statement

(e)

S4 alone is a true statement

## Answer

(b)

## Solution

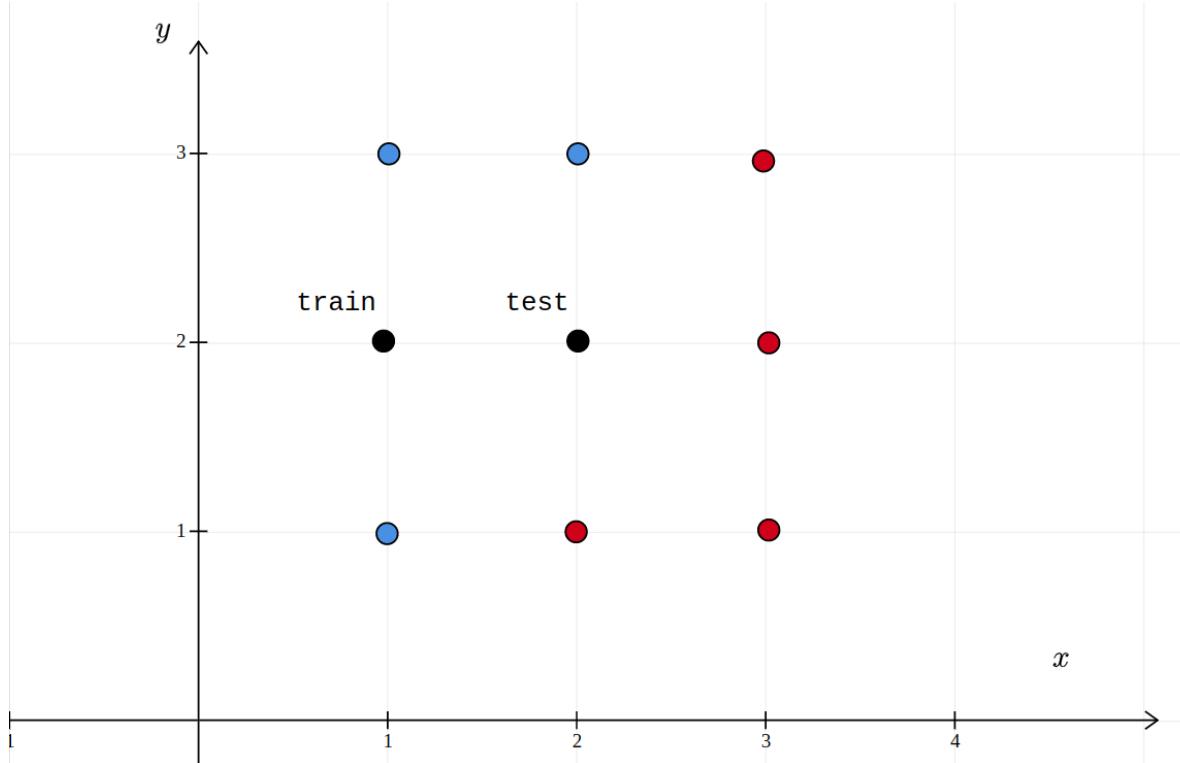
The entire training dataset has to be stored in memory. For predicting the label of a test-point, we have to perform the following steps:

- Compute distance of the test-point from each training point.
- Sort the training data points in ascending order of distance.
- Choose the first  $k$  points in this sequence.
- Return that label which garners the maximum vote among these  $k$  neighbors.

## Question-2

### Statement

The blue and the red points belong to two different classes. Both of them are a part of the training dataset. The black point at  $(1, 2)$  also belongs to the training dataset, but its true color is hidden from our view. The black point at  $(2, 2)$  is a test-point.



How should we recolor the black train point if the test point is classified as "red" without any uncertainty by a  $k$ -NN classifier, with  $k = 4$ ? Use the Euclidean distance metric for computing distances.

### Options

(a)

blue

(b)

red

(c)

Insufficient information

### Answer

(b)

## Solution

Since we are looking at the  $k$ -NN algorithm with  $k = 4$ , we need to look at the four nearest neighbors of the test data-point. The four points from the training dataset that are closest to the test data-point are the following:

- (1, 2): black
- (2, 3): blue
- (3, 2): red
- (2, 1): red

Each of them is at unit distance from the test data-point. From the problem statement, it is given that the test data-point is classified as "red" without any uncertainty. Let us now consider two scenarios that concern the black training data-point at (1, 2):

### Black training data-point is colored red

There are three red neighbors and one blue neighbor. Therefore, the test-data point will be classified as red. There is no uncertainty in the classification. This is what we want. However, for the sake of completeness, let us look at the alternative possibility.

### Black training data-point is colored blue

There will be exactly two neighbors that are blue and two that are red. In such a scenario, we can't classify the black test-point without any uncertainty. That is, we could call it either red or blue. This is one of the reasons why we choose an odd value of  $k$  for the  $k$ -NN algorithm. If  $k$  is odd, then this kind of a tie between the two classes can be avoided.

## Question-3

### Statement

Consider the following feature vectors:

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \\ -1 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 5 \\ -3 \\ -5 \\ 10 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 3 \\ 1 \\ 2 \\ 4 \end{bmatrix}, \mathbf{x}_4 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \mathbf{x}_5 = \begin{bmatrix} 10 \\ 7 \\ -3 \\ 2 \end{bmatrix}$$

The labels of these four points are:

$$y_1 = 1, y_2 = 0, y_3 = 1, y_4 = 0, y_5 = 0$$

If use a  $k$ -NN algorithm with  $k = 3$ , what would be the predicted label for the following test point:

$$\mathbf{x}_{\text{test}} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

### Answer

1

### Solution

The distances are:

- $d(\mathbf{x}_{\text{test}}, \mathbf{x}_1)^2 = 5$
- $d(\mathbf{x}_{\text{test}}, \mathbf{x}_2)^2 = 149$
- $d(\mathbf{x}_{\text{test}}, \mathbf{x}_3)^2 = 14$
- $d(\mathbf{x}_{\text{test}}, \mathbf{x}_4)^2 = 2$
- $d(\mathbf{x}_{\text{test}}, \mathbf{x}_5)^2 = 134$

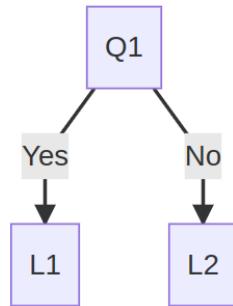
We see that among the three nearest neighbors, two have label 1 and one has label 0. Hence the predicted label is 1. For those interested in a code for the same:

```
1 import numpy as np
2
3 x_1 = np.array([1, 2, 1, -1])
4 x_2 = np.array([5, -3, -5, 10])
5 x_3 = np.array([3, 1, 2, 4])
6 x_4 = np.array([0, 1, 1, 0])
7 x_5 = np.array([10, 7, -3, 2])
8
9 x_test = np.array([1, 1, 1, 1])
10
11 for x in [x_1, x_2, x_3, x_4, x_5]:
12     print(round(np.linalg.norm(x_test - x) ** 2))
```

# Comprehension Type (4 to 6)

## Statement

Consider the following split at some node in a decision tree:



The following is the distribution of data-points and their labels:

Node	Num of points	Labels
Q1	100	0
Q1	100	1
L1	50	0
L1	30	1
L2	50	0
L2	70	1

For example, L1 has 80 points of which 50 belong to class 0 and 30 belong to class 1. Use  $\log_2$  for all calculations that involve logarithms.

## **Question-4**

---

### **Statement**

If the algorithm is terminated at this level, then what are the labels associated with L1 and L2?

### **Options**

**(a)**

L1 : 0

**(b)**

L1 : 1

**(c)**

L2 : 0

**(d)**

L2 : 1

### **Answer**

(a), (d)

### **Solution**

- $L_1$  has 80 data-points out of which 50 belong to class-0 and 30 belong to class-1. Since the majority of the points belong to class-0, this node will have 0 as the predicted label.
- $L_2$  has 120 data-points out of which 50 belong to class-0 and 70 belong to class-1. Since the majority of the points belong to class 1, this node will have 1 as the predicted label.

## Question-5

---

### Statement

What is the impurity in L1 if we use entropy as a measure of impurity? Report your answer correct to three decimal places.

### Answer

0.954

Range: [0.94, 0.96]

### Solution

If  $p$  represents the proportion of the samples that belong to class-1 in a node, then the impurity of this node using entropy as a measure is:

$$-p \log p - (1-p) \log(1-p)$$

For  $L_1$ ,  $p = \frac{30}{30+50} = \frac{3}{8}$ . So, the impurity for  $L_1$  turns out to be:

$$-\frac{3}{8} \log\left(\frac{3}{8}\right) - \frac{5}{8} \log\left(\frac{5}{8}\right) \approx 0.954$$

Code for reference:

```
1 | import math
2 | imp = lambda p: -p * math.log2(p) - (1 - p) * math.log2(1 - p)
3 | print(imp(3 / 8))
```

## Question-6

---

### Statement

What is the information gain for this split? Report your answer correct to three decimal places. Use at least three decimal places in all intermediate computations.

### Answer

0.030

Range: [0.025, 0.035]

### Solution

The information gain because of this split is equal to the decrease in impurity. Here,  $|L_1|$  and  $|L_2|$  denote the cardinality of the leaves.  $N$  is the total number of points before the split at node  $Q$ .

$$IG = E(Q) - \left( \frac{|L_1|}{N} E(L_1) + \frac{|L_2|}{N} E(L_2) \right)$$

For this problem, the variables take on the following values:

- $N = 200$ , there are 200 points in the node  $Q$ .
- $|L_1| = 80$ , there are 80 points in the node  $L_1$ .
- $|L_2| = 120$ , there are 120 points in the node  $L_2$ .

To calculate the entropy of the three nodes, we need the proportion of points that belong to class-1 in each of the three nodes. Let us call them  $p$  for node  $Q$ ,  $p_1$  for node  $L_1$  and  $p_2$  for node  $L_2$ :

- $p = \frac{100}{100+100} = \frac{1}{2}$
- $p_1 = \frac{30}{30+50} = \frac{3}{8}$
- $p_2 = \frac{70}{70+50} = \frac{7}{12}$

Now, we have all the data that we need to compute  $E(Q)$ ,  $E(L_1)$  and  $E(L_2)$ :

- $E(Q) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$
- $E(L_1) = -\frac{3}{8}\log\left(\frac{3}{8}\right) - \frac{5}{8}\log\left(\frac{5}{8}\right) \approx 0.954$
- $E(L_2) = -\frac{7}{12}\log\left(\frac{7}{12}\right) - \frac{5}{12}\log\left(\frac{5}{12}\right) \approx 0.980$

Now, we have all the values to compute the information gain:

$$IG = 1 - \left( \frac{80}{200} 0.954 + \frac{120}{200} 0.980 \right) \approx 0.030$$

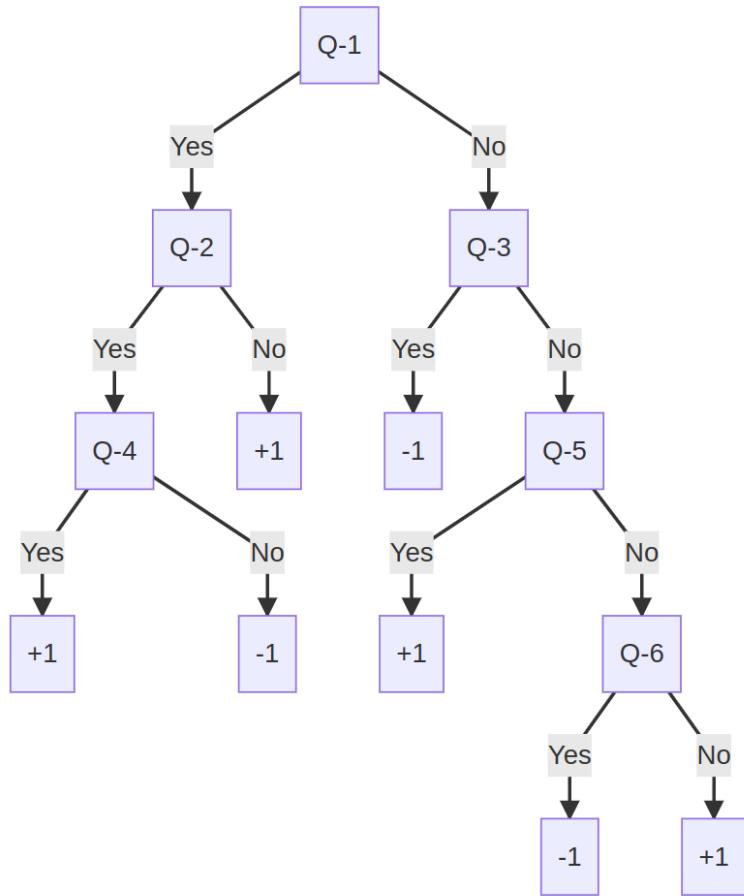
Code for reference:

```
1 import math
2 imp = lambda p: -p * math.log2(p) - (1 - p) * math.log2(1 - p)
3
4 p_0 = 1 / 2
5 p_1 = 3 / 8
6 p_2 = 7 / 12
7
8 n = 200
9 l_1 = 80
10 l_2 = 120
11
12 ig = imp(p_0) - ((l_1 / n) * imp(p_1) + (l_2 / n) * imp(p_2))
13 print(ig)
```

# Question-7

## Statement

Consider the following decision tree. Q-i corresponds to a question. The labels are  $+1$  and  $-1$ .



If a test-point comes up for prediction, what is the minimum and maximum number of questions that it would have to pass through before being assigned a label?

## Options

(a)

$\min = 1$

(b)

$\min = 2$

(c)

$\min = 3$

(d)

$\max = 3$

**(e)**

$\max = 4$

## **Answer**

(b), (e)

## **Solution**

Look at all paths from the root to the leaves. Find the shortest and longest path.

## **Question-8**

---

### **Statement**

$p$  is the proportion of points with label 1 in some node in a decision tree. Which of the following statements are true? [MSQ]

### **Options**

**(a)**

As the value of  $p$  increases from 0 to 1, the impurity of the node increases

**(b)**

As the value of  $p$  increases from 0 to 1, the impurity of the node decreases

**(c)**

The impurity of the node does not depend on  $p$

**(d)**

$p = 0.5$  correspond to the case of maximum impurity

### **Answer**

(d)

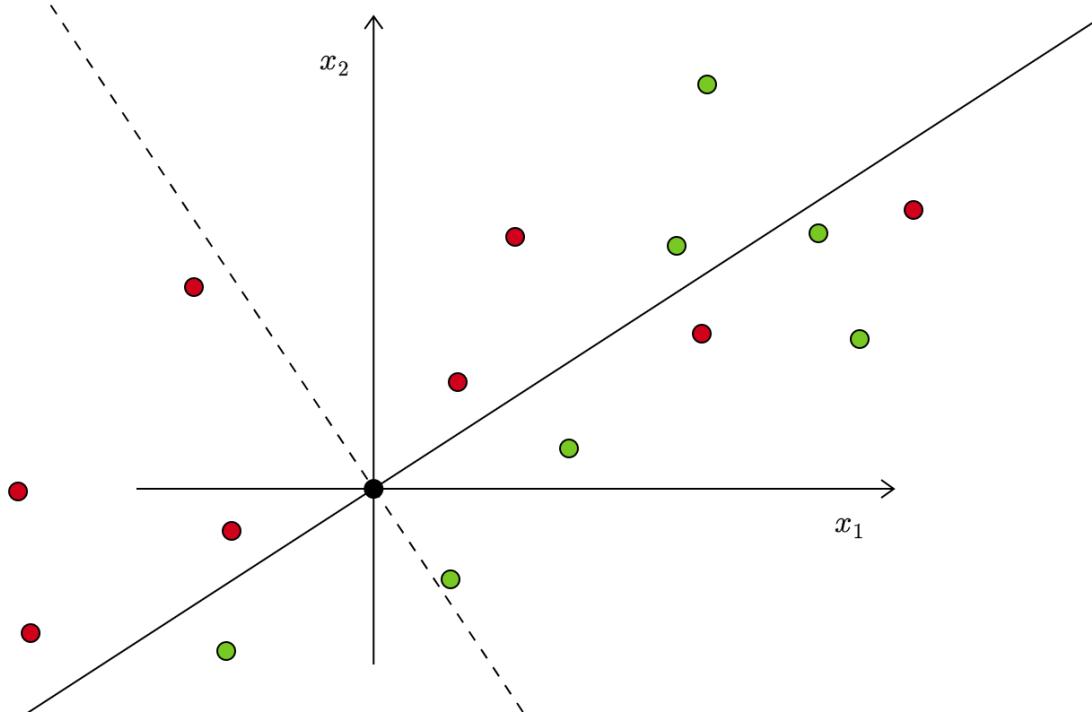
### **Solution**

Options (a) and (b) are incorrect as the impurity increases from  $p = 0$  to  $p = 0.5$  and then decreases. Option-(c) is incorrect for obvious reasons.

## Question-9

### Statement

Consider a binary classification problem in which all data-points are in  $\mathbb{R}^2$ . The red points belong to class  $+1$  and the green points belong to class  $-1$ . A linear classifier has been trained on this data. The decision boundary is given by the solid line.



This classifier misclassifies four points. Which of the following could be a possible value for the weight vector?

### Options

(a)

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

(b)

$$\begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

(c)

$$\begin{bmatrix} -1 \\ -2 \end{bmatrix}$$

(d)

$$\begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

## **Answer**

(b)

## **Solution**

The weight vector is orthogonal to the decision boundary. So it will lie on the dotted line. This gives us two quadrants in which the vector can lie in: second or fourth. In other words, we only need to figure out its direction. If it is pointing in the second quadrant, then there will be four misclassifications. If it is pointing in the fourth quadrant then all but four points will be misclassified.

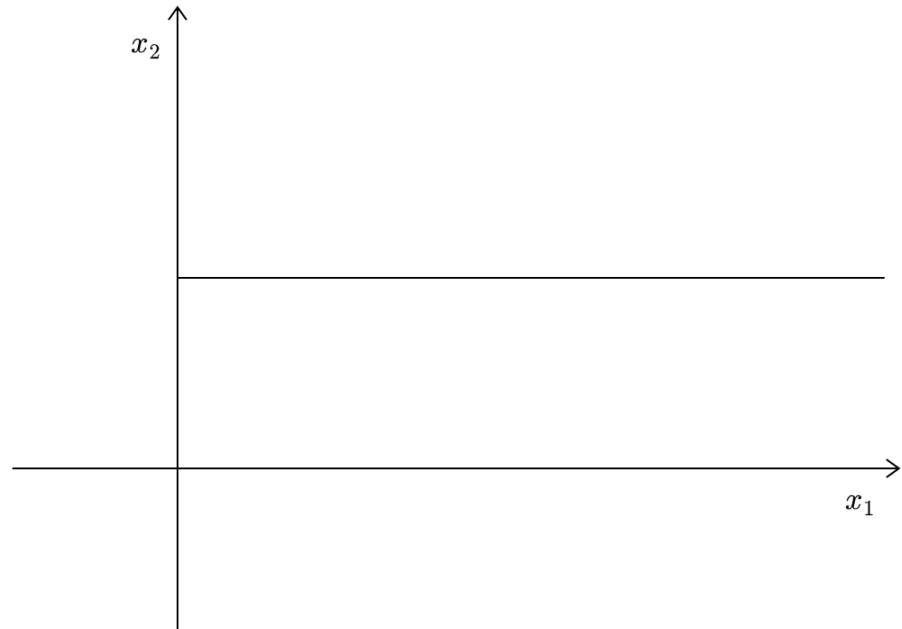
## Question-10

### Statement

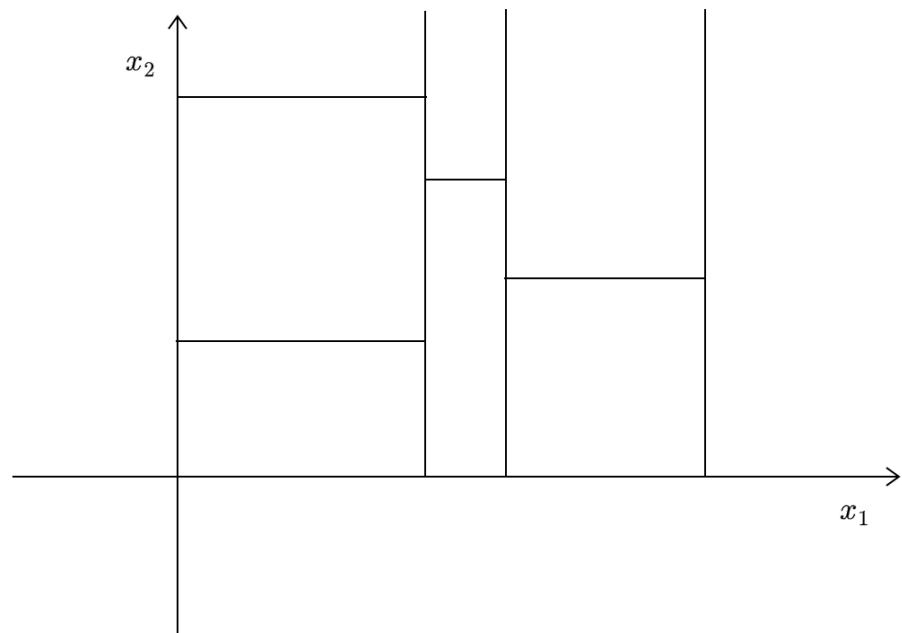
Which of the following are valid decision regions for a decision tree classifier for datapoints in  $\mathbb{R}^2$ ? The question in every internal node is of the form  $f_k \leq \theta$ . Both the features are positive real numbers.

### Options

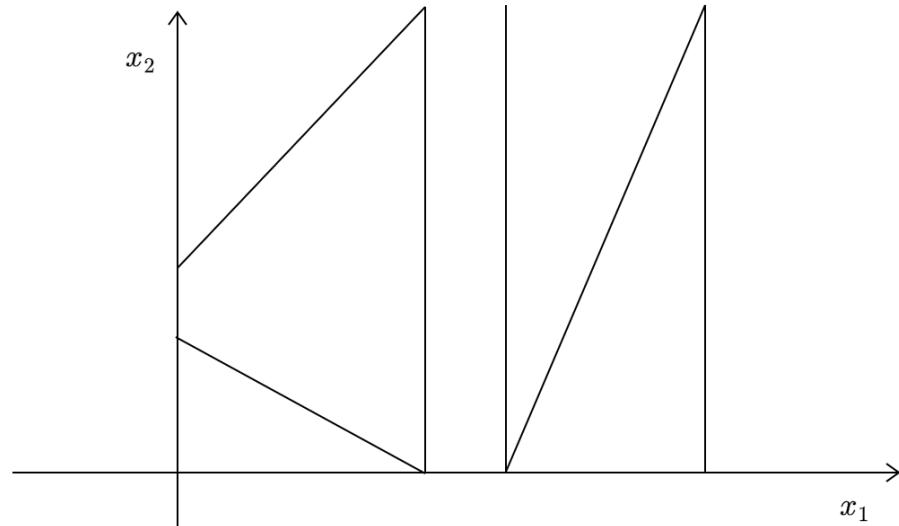
(a)



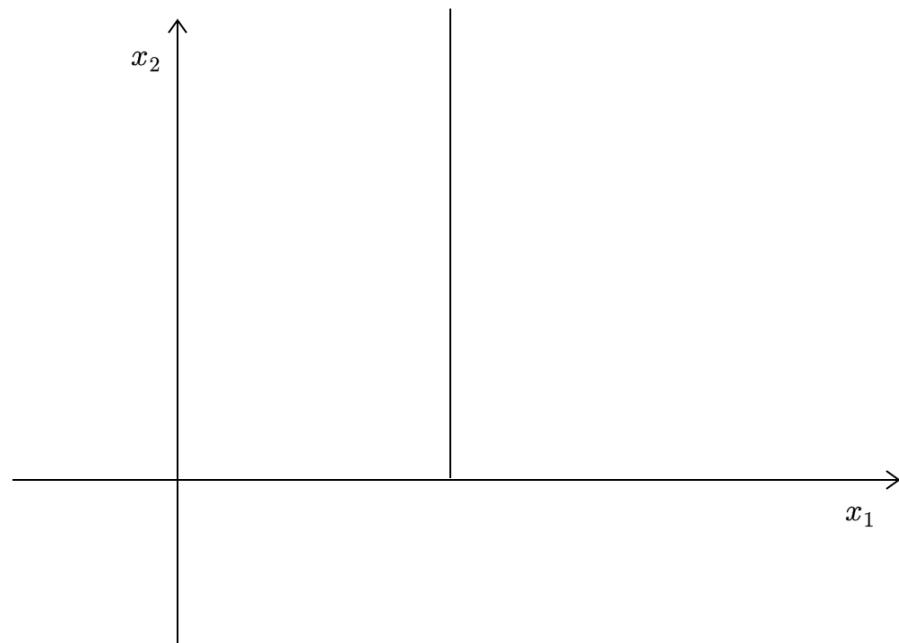
(b)



(c)



(d)



## Answer

(a), (b), (d)

## Solution

A question of the form  $f_k \leq \theta$  can only result in one of these two lines:

- a horizontal line
- a vertical line

It cannot produce a slanted line as shown in option-(c). Options (a) and (d) correspond to what are called decision stumps: a single node splitting into two child nodes.

# Graded assignment

---

## Question 1

---

### Statement

Consider the two different generative model-based algorithms.

1. Model 1: chances of occurring a feature are affected by the occurrence of other features and the model does not impose any additional condition on conditional independence of features.
2. Model 2: chances of occurring a feature are not affected by the occurrence of other features and therefore, the model assumes that features are conditionally independent of the label.

Which model has more independent parameters to estimate?

### Options

(a)

Model 1

(b)

Model 2

### Answer

(a)

### Solution:

In the first model, features are not independent, therefore, we need to find the probabilities (or density) for each and every possible example given the labels whereas in model 2, the features are independent, therefore we need to find the pmf (or pdf) of the features only.

That is the model 1 has more parameters to estimate.

---

## Question 2

---

### Statement

Which of the following statement is/are always correct in context to the naive Bayes classification algorithm for binary classification with all binary features? Here,  $\hat{p}_j^y$  denotes the estimate for the probability that the  $j^{th}$  feature value of a data point is 1 given that the point has the label  $y$ .

## Options

(a)

If  $\hat{p}_j^y = 0.2$  for  $y = 0$ , then  $\hat{p}_j^y = 0.8$  for  $y = 1$

(b)

$\sum_{j=1}^d \hat{p}_j^y = 1$  for any  $y$

(c)

If  $\hat{p}_j^y = 0$  for  $y = 0$ , then  $\hat{p}_j^y = 0$  for  $y = 1$

(d)

If  $\hat{p}_j^1 = 0$ , no labeled 1 example in the training dataset takes  $j^{th}$  feature values as 1.

(e)

None of the above

## Answer

(d)

## Solution

In general,  $\hat{p}_j^y$  = estimate for  $P(f_j = 1|y)$ .

It means that  $\hat{p}_j^y$  denotes the parameters of different distributions  $f_j|y$  for different  $y$  and for different  $j$ .

Therefore, If  $\hat{p}_j^y = 0.2$  for  $y = 0$ , then it is not mandatory that  $\hat{p}_j^y = 0.8$  for  $y = 1$  as they come from different distributions.

For different  $j$ , distributions  $f_j|y$  are different distributions. Therefore, it is not necessary that  $\sum_{j=1}^d \hat{p}_j^y = 1$ . What we can say is that  $\sum_{i \in R(f_j)} \hat{p}_{j=i}^y = 1$

If  $\hat{p}_j^y = 0$  for  $y = 0$ , It implies that there is no labeled zero examples such that  $j^{th}$  feature value is 1. It doesn't mean that all  $j^{th}$  feature value is 1 for all labeled one examples.

If  $\hat{p}_j^1 = 0$ , no labeled 1 example in the training dataset takes  $j^{th}$  feature values as 1.

---

## Question 3

## Statement

A naive Bayes model is trained on a dataset containing  $d$  features  $f_1, f_2, \dots, f_d$ . Labels are 0 and 1. If a test point was predicted to have the label 1, which of the following expression should be sufficient for this prediction?

## Options

(a)

$$P(y = 1) > P(y = 0)$$

(b)

$$\prod_{i=1}^d P(f_i|y = 1) > \prod_{i=1}^d P(f_i|y = 0)$$

(c)

$$\left( \prod_{j=1}^d (\hat{p}_j^1)^{f_j} (1 - \hat{p}_j^1)^{1-f_j} \right) P(y = 1) > \left( \prod_{j=1}^d (\hat{p}_j^0)^{f_j} (1 - \hat{p}_j^0)^{1-f_j} \right) P(y = 0)$$

(d)

None of the above

## Answer

(c)

## Solution

A test example is predicted label 1, it implies that

$$\begin{aligned} & P(y = 1|x) > P(y = 0|x) \\ \Rightarrow & \frac{P(x|y = 1) \cdot P(y = 1)}{P(x)} > \frac{P(x|y = 0) \cdot P(y = 0)}{P(x)} \\ \Rightarrow & P(x|y = 1) \cdot P(y = 1) > P(x|y = 0) \cdot P(y = 0) \\ \Rightarrow & \left( \prod_{j=1}^d (\hat{p}_j^1)^{f_j} (1 - \hat{p}_j^1)^{1-f_j} \right) P(y = 1) > \left( \prod_{j=1}^d (\hat{p}_j^0)^{f_j} (1 - \hat{p}_j^0)^{1-f_j} \right) P(y = 0) \end{aligned}$$

## Question 4

---

### Statement

Consider a binary classification dataset contains only one feature and the data points given the label follow the given distribution

$$\begin{aligned} x|(y = 0) &\sim N(0, 2) \\ x|(y = 1) &\sim N(2, \sigma^2) \end{aligned}$$

If the decision boundary learned using the gaussian naive Bayes algorithm is linear, what is the value of  $\sigma^2$ ?

## Answer

2

## Solution

Since the decision boundary is linear, both theiances will be the same. That is  $\sigma^2 = 2$

## Question 5

---

### Statement

Consider a binary classification dataset with two binary features  $f_1$  and  $f_2$ . The  $f_2$  feature values are 0 for all label '0' examples but the label '1' examples take both values 1 and 0 for the feature  $f_2$ . If we apply the naive Bayes algorithm on the same dataset, what will be the prediction for point  $[1, 1]^T$ ?

### Options

(a)

Label 0

(b)

Label 1

(c)

Insufficient information to predict.

## Answer

(c)

## Solution

Given that the  $f_2$  feature values are 0 for all label '0' examples it implies that  $\hat{p}_2^0 = 0$ .

Therefore,  $p(y = 0|x) = 0$

But the label '1' examples take both values 1 and 0 for the feature  $f_2$ . It implies that  $\hat{p}_2^1 > 0$ .

Still the value of  $p(y = 1|x)$  can be 0 if the value of  $\hat{p}_1^1$  is zero. So, we need the value of  $\hat{p}_1^1$  to make any conclusion.

---

## Common data for questions 6 and 7

---

## Statement

Consider the following binary classification dataset with two features  $f_1$  and  $f_2$ . The data points given the labels follow the Gaussian distribution. The dataset is given as

$f_1$	$f_2$	label $y$
0.5	1.3	1
0.7	1.1	1
1.3	2.0	0
2.3	2.4	0

## Question 6

---

### Statement

What will be the value of  $\hat{p}$ , the estimate for  $P(y = 1)$ ?

### Answer

0.5

### Solution

$$\begin{aligned}\hat{p} &= \frac{\sum_{i=1}^n y_i}{n} \\ &= 2/4 = 0.5\end{aligned}$$

## Question 7

---

### Statement

What will be the value of  $\hat{\mu}_0$ ?

### Options

(a)

(1.8, 2.2)

(b)

(0.6, 1.2)

**(c)**

(2.0, 2.0)

**(d)**

(0.8, 1.2)

## Answer

(a)

## Solution

$$\begin{aligned}\hat{\mu}_0 &= \frac{\sum_{i=1}^n \mathbb{1}(y_i = 0)x_i}{\sum_{i=1}^n \mathbb{1}(y_i = 0)} \\ &= \frac{(1.3, 2.0) + (2.3, 2.4)}{2} \\ &= (1.8, 2.2)\end{aligned}$$

## Question 8

---

### Statement

Consider a binary classification dataset containing two features  $f_1$  and  $f_2$ . The feature  $f_1$  is categorical which can take three values and the feature  $f_2$  is numerical that follows the Gaussian distribution. How many independent parameters must be estimated if we apply the naive Bayes algorithm to the same dataset?

### Answer

9

## Solution

We need one parameter for  $P(y = 1)$  as  $y$  takes only two values.

For a given label (say  $y = 1$ )

feature  $f_1$  can take three values, therefore we need two estimates for  $P(f_1 = 0|y = 1)$  and  $P(f_1 = 1|y = 1)$ .

Similarly, two estimates if  $y = 0$

For feature  $f_2$ , we need  $\mu_0, \mu_1, \Sigma_0$  and  $\Sigma_1$ .

Therefore, total number of parameters =  $1 + 2 + 2 + 4 = 9$

---

## Common data for questions 9 and 10

---

## Statement

A binary classification dataset has 1000 data points belonging to  $\{0, 1\}^2$ . A naive Bayes algorithm was run on the same dataset that results in the following estimate:

$\hat{p}$ , estimate for $P(y = 1)$	<b>0.3</b>
$\hat{p}_1^0$ , estimate for $P(f_1 = 1 y = 0)$	0.2
$\hat{p}_2^0$ , estimate for $P(f_2 = 1 y = 0)$	0.3
$\hat{p}_1^1$ , estimate for $P(f_1 = 1 y = 1)$	0.1
$\hat{p}_2^1$ , estimate for $P(f_2 = 1 y = 1)$	0.02

## Question 9

---

### Statement

What is the estimated value of  $P(f_2 = 0|y = 1)$ ? Write your answer correct to two decimal places.

### Answer

0.98 Range: [0.97, 0.99]

### Solution

$$\begin{aligned}\text{estimate for } P(f_2 = 0|y = 1) &= 1 - (\text{estimate for } P(f_2 = 1|y = 1)) \\ &= 1 - 0.02 = 0.98\end{aligned}$$

---

## Question 10

### Statement

What will be the predicted label for the data point  $[0, 1]$ ?

### Answer

0 No range is required

### Solution

$$\begin{aligned}P(y = 0|x) &\propto P(x|y = 0). P(y = 0) \\ &\propto P(f_1 = 0|y = 0). P(f_2 = 1|y = 0). P(y = 0) \\ &\propto (1 - 0.2)(0.3)(1 - 0.3) \\ &= 0.168\end{aligned}$$

$$\begin{aligned}P(y = 1|x) &\propto P(x|y = 1). P(y = 1) \\ &\propto P(f_1 = 0|y = 1). P(f_2 = 1|y = 1). P(y = 1) \\ &\propto (1 - 0.1)(0.02)(0.3) \\ &= 0.054\end{aligned}$$

Since  $P(y = 0|x) > P(y = 1|x)$ , therefore the point will be predicted label 0.

# **Graded**

---

## Question-1

---

### Statement

Assume that Perceptron algorithm is applied to a data set in which the maximum of the lengths of the data points is 4 and the value of margin ( $\gamma$ ) of the optimal separator is 1. If the algorithm has made 10 mistakes at some point of the execution of the algorithm, which of the following can be valid squared length(s) of the weight vector obtained in the 11th iteration?

### Options

(a)

90

(b)

150

(c)

170

(d)

190

### Answer

(b), (c)

### Solution

Given,  $R = 4, \gamma = 1$

$l = 10$

Need to find  $w^{l+1}$ .

$$\|w^{l+1}\|^2 \leq (l + 1)R^2$$

and

$$\|w^{l+1}\|^2 \geq (l + 1)^2\gamma^2$$

Hence,

$$\|w^{l+1}\|^2 \leq (10 + 1)4^2$$

$$\|\mathbf{w}^{l+1}\|^2 \leq 176$$

and

$$\|w^{l+1}\|^2 \geq (10 + 1)^21^2$$

$$\|\mathbf{w}^{l+1}\|^2 \geq 121$$

Hence both (b) and (c) will be correct.

## Question-2

### Statement

Consider the following data set:

f_1	f_2	y
-1	-1	-1
0	1	+1
1	0	+1
1	1	+1

If Perceptron algorithm is applied on this data set with the weight vector initialized to  $[0, 0]$ , how many times the weight vector will be updated during the training process?

### Options

(a)

0

(b)

1

(c)

2

(d)

3

### Answer

(b)

### Solution

$$w_0 = [0 \ 0]$$

$I_1$ :

The predictions for each of the four data points  $x_1, x_2, x_3, x_4$  as per  $w^T x$  will be  $+1, +1, +1, +1$ .

The mistake occurs for  $x_1$

Hence,  $w_1 = w_0 + x_1 y_1$

Resulting into  $w_1 = [1 \ 1]$

**I<sub>2</sub>:**

The predictions for each of the four data points  $x_1, x_2, x_3, x_4$  as per  $w^T x$  will be  $-1, +1, +1, +1$ , which are correct.

Hence, the weight update happened once.

## Question-3

---

### Statement

Consider the following data set with three data points:

$$\left( \begin{bmatrix} 2 \\ 2 \end{bmatrix}, +1 \right), \left( \begin{bmatrix} 2 \\ -2 \end{bmatrix}, +1 \right), \left( \begin{bmatrix} -2 \\ 1 \end{bmatrix}, -1 \right)$$

If the Perceptron algorithm is applied to this data with the initial weight vector  $w^0$  to be a zero vector, what will be the outcome?

### Options

(a)

The algorithm will converge with  $w = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$

(b)

The algorithm will converge with  $w = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$

(c)

The algorithm will converge with  $w = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$

(d)

The algorithm will never converge.

### Answer

(c)

### Solution

$$w_0 = [0 \ 0]$$

**I<sub>1</sub>:**

The predictions for each of the three data points  $x_1, x_2, x_3$  as per  $w^T x$  will be  $+1, +1, +1$

The mistake occurs for  $x_3$

$$\text{Hence, } w_1 = w_0 + x_3 y_3$$

$$\text{Resulting into } w_1 = [2 \ -1]$$

**I<sub>2</sub>:**

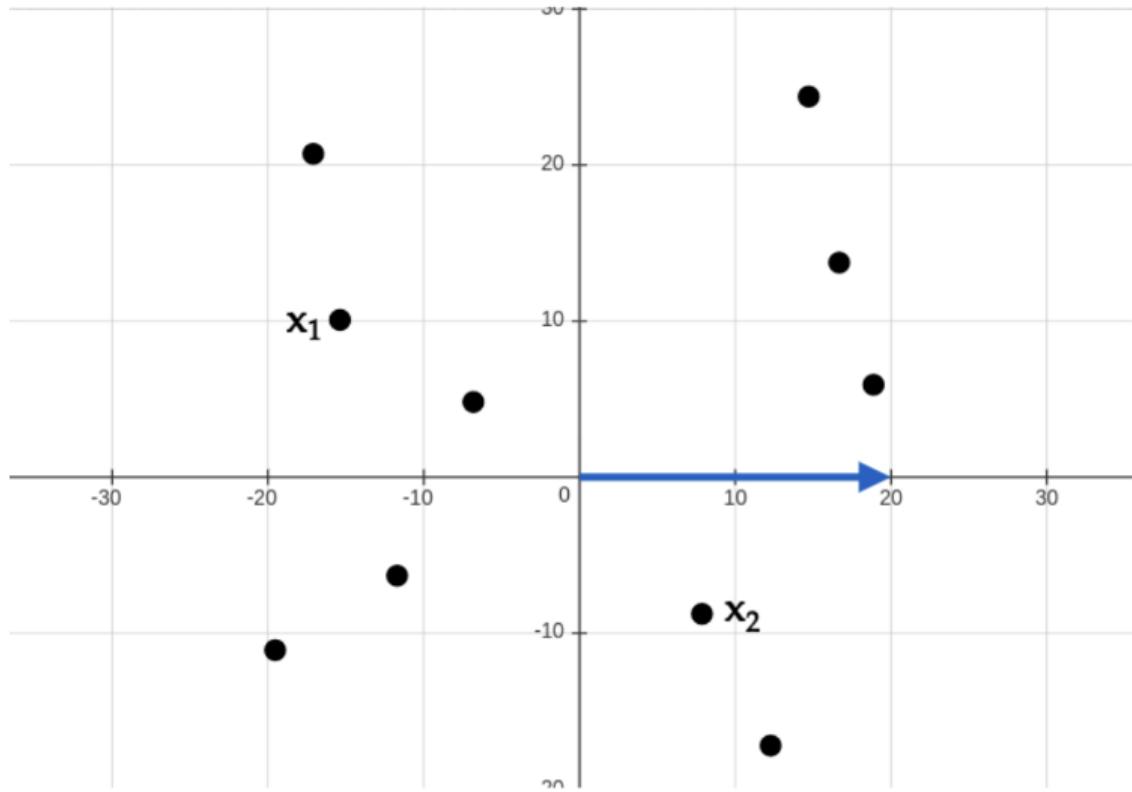
The predictions for each of the three data points  $x_1, x_2, x_3$  as per  $w^T x$  will be  $+1, +1, -1$ , which are correct.

Hence, the algorithm will converge with  $w = [2 \ -1]$

## Question-4

### Statement

Consider ten data points as shown in the following image:



The blue line represents the weight vector. As per this weight vector, the Perceptron algorithm will predict which classes for the data points  $x_1$  and  $x_2$ ?

### Options

(a)

$$x_1 : +1, x_2 : -1$$

(b)

$$x_1 : -1, x_2 : +1$$

(c)

$$x_1 : +1, x_2 : +1$$

(d)

$$x_1 : -1, x_2 : -1$$

## **Answer**

(b)

## **Solution**

The decision boundary will be perpendicular to  $w$ . For the data points on the right side of it,  $w^T x$  will be greater than equal to zero, on the LHS, it will be less than zero.

Accordingly, RHS data points (and hence  $x_2$ ) will be predicted as +1.

And, LHS data points (and hence  $x_1$ ) will be predicted as -1.

## Question-5

---

### Statement

In the previous question, if the weight vector is multiplied by -1, which classes will be predicted by the Perceptron for the data points  $x_1$  and  $x_2$ ?

### Options

(a)

$x_1 : +1, x_2 : -1$

(b)

$x_1 : -1, x_2 : +1$

(c)

$x_1 : +1, x_2 : +1$

(d)

$x_1 : -1, x_2 : -1$

### Answer

(a)

### Solution

If the weight vector is multiplied by -1, then for the data points on the RHS,  $w^T x$  will be less than 0 and on the LHS, it will be greater than equal to zero.

Accordingly, LHS data points (and hence  $x_1$ ) will be predicted as +1.

And, RHS data points (and hence  $x_2$ ) will be predicted as -1.

## **Question-6**

---

### **Statement**

Given a data set with  $R = 4$ ,  $\gamma = 2$ , what is the maximum number of mistakes that Perceptron algorithm can make on the data?

### **Options**

**(a)**

2

**(b)**

4

**(c)**

8

**(d)**

16

### **Answer**

(b)

### **Solution**

The maximum number of mistakes is given by  $R^2/\gamma^2$

$$\text{Which is } \frac{4^2}{2^2} = 4$$

## Question-7

---

### Statement

If the scores (i.e,  $w^T x$  values) for some data points are -4, 3, 1, 2, -6 respectively, what will be the probabilities returned for these points by Logistic Regression?

### Options

(a)

0.25, 0.1875, 0.0625, 0.125, 0.375

(b)

-1, 1, 1, 1, -1

(c)

0.017, 0.95, 0.73, 0.88, 0.002

(d)

0, 1, 1, 1, 0

### Answer

(c)

### Solution

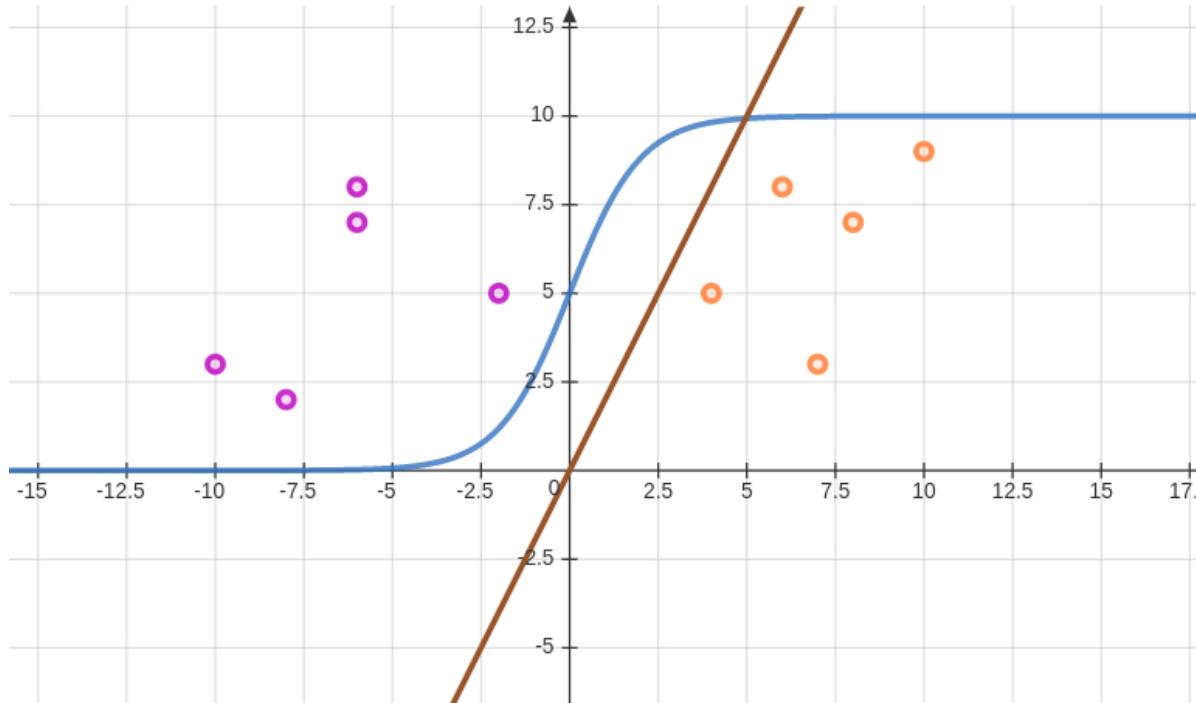
Put the values -4, 3, 1, 2, -6 in the formula  $g(z) = \frac{1}{1 + e^{-z}}$  one by one.

$$\text{Ex: } g(-4) = \frac{1}{1 + e^4} = \frac{1}{1 + 54.59} = \frac{1}{55.59} = 0.017$$

## Question-8

### Statement

Which of the lines (blue or brown) in the following image may represent the decision boundary of Logistic Regression?



### Options

(a)

Blue line

(b)

Brown line

(c)

Both

(d)

None of these

### Answer

(b)

### Solution

The decision boundary in logistic regression is always linear (Brown line).

It's just that the values obtained from a linear combination are reduced to values between 0 and 1 by the sigmoid function (blue line).

## Question-9

---

### Statement

Consider a data set  $x_1 = [-1, -1]$ ,  $x_2 = [-1, 0]$ ,  $x_3 = [0, 1]$ ,  $x_4 = [0, -1]$ . Let the corresponding class labels be  $y_1 = y_2 = y_4 = -1$  and  $y_3 = +1$ .

Assume you try to find the  $w$  using the Perceptron algorithm. You decide to cycle through points in the order  $\{x_4, x_3, x_2, x_1\}$  repeatedly until you find a linear separator. How many mistakes does your algorithm make and what is the linear separator your algorithm outputs?

### Options

(a)

3, [1, 0]

(b)

2, [1, 1]

(c)

3, [-1, 1]

(d)

4, [-1, 0]

### Answer

(b)

### Solution

When initial weight vector is not given, we will take zero vector as initial weight vector.

$$w^0 = [0 \ 0]$$

The order in which we have to traverse the data points is given.

We start with  $x_4$ .

$w^T x_4$  predicts +1 class, which is a mistake.

Hence,  $w = w + x_4 y_4$  gives  $w = [0 \ 1]$

$w^T x_3$  predicts +1 which is correct.

$w^T x_2$  predicts +1 which is a mistake.

Hence  $w = w + x_2 y_2$  gives  $w = [1 \ 1]$

$w^T x_1$  predicts -1 which is correct.

Once again we check  $x_4, x_3, x_2, x_1$  in that order, and they all are predicted correctly.

Hence final  $w = [1 \ 1]$  and  $w$  had been updated twice.

# **Graded**

---

This document has 8 questions.

## Question-1 [1 point]

---

### Statement

Consider a linearly separable dataset for a binary classification problem in  $\mathbb{R}^d$ . Three linear classifiers have been trained on this dataset. All three pass through the origin and have the following property:

$$(\mathbf{w}_j^T \mathbf{x}_i) \cdot y_i \geq 1, \quad 1 \leq i \leq n$$

Here,  $\mathbf{w}_j$  is the weight vector corresponding to the  $j^{th}$  classifier. Note that the above property is satisfied for each of the  $n$  data-points. If  $\mathbf{w}_1$  is the weight vector corresponding to a hard-margin SVM, which of the following statements is always true? You can assume that the norms of all three weights are different from each other.

### Options

(a)

$$\|\mathbf{w}_1\| > \|\mathbf{w}_2\| > \|\mathbf{w}_3\|$$

(b)

$$\|\mathbf{w}_1\| < \|\mathbf{w}_2\| < \|\mathbf{w}_3\|$$

(c)

$$\|\mathbf{w}_1\| < \|\mathbf{w}_2\| \text{ and } \|\mathbf{w}_1\| < \|\mathbf{w}_3\|$$

(d)

$$\|\mathbf{w}_1\| > \|\mathbf{w}_2\| \text{ and } \|\mathbf{w}_1\| > \|\mathbf{w}_3\|$$

### Answer

(c)

### Solution

$\mathbf{w}_1$  will have the smallest norm (maximum margin) among the three classifiers. The three weight vectors are feasible points for the primal. Among them,  $\mathbf{w}_1$  is optimal.

## Common Data for questions (2) to (4)

---

### Statement

#### Common data for questions (2) to (4)

Consider the following training dataset for a binary classification problem in  $\mathbb{R}^2$ . Each data-point is represented by  $\mathbf{x} = [x_1 \quad x_2]^T$  whose label is  $y$ .

Index	$x_1$	$x_2$	$y$
1	1	0	1
2	-1	0	-1
3	5	4	1
4	-5	-4	-1

We wish to train a hard-margin SVM for this problem.  $\mathbf{w} = [w_1 \quad w_2]^T$  represents the weight vector. The index  $i$  is for the  $i^{th}$  data-point.  $\alpha_i$  is the Lagrange multiplier for the  $i^{th}$  data-point.

## Question-2 [1 point]

---

### Statement

Select all primal constraints from the options given below.

### Options

(a)

$$w_1 \geq 1$$

(b)

$$w_1 \leq 1$$

(c)

$$5w_1 + 4w_2 \geq 1$$

(d)

$$5w_1 + 4w_2 \leq 1$$

### Answer

(a), (c)

### Solution

Because of the symmetry in the problem, we effectively have only two constraints even though there are 4 data-points:

$$\begin{aligned} w_1 &\geq 1 \\ 5w_1 + 4w_2 &\geq 1 \end{aligned}$$

But in order to remain consistent with our formulation, let us list it down in the following manner:

$$\begin{aligned} w_1 &\geq 1 & (1) \\ w_1 &\geq 1 & (2) \\ 5w_1 + 4w_2 &\geq 1 & (3) \\ 5w_1 + 4w_2 &\geq 1 & (4) \end{aligned}$$

## Question-3 [2 points]

### Statement

Which of the following is the objective function of the dual problem? In all options,  $\alpha = [\alpha_1 \quad \alpha_2 \quad \alpha_3 \quad \alpha_4]^T$  and  $\mathbf{1} = [1 \quad 1 \quad 1 \quad 1]^T$ .

### Options

(a)

$$\alpha^T \mathbf{1} - \frac{1}{2} \cdot \alpha^T \alpha$$

(b)

$$\alpha^T \mathbf{1} - \frac{1}{2} \cdot \alpha^T \begin{bmatrix} 1 & 1 & 5 & 5 \\ 1 & 1 & 5 & 5 \\ 5 & 5 & 41 & 41 \\ 5 & 5 & 41 & 41 \end{bmatrix} \alpha$$

(c)

$$\alpha^T \mathbf{1} - \frac{1}{2} \cdot \alpha^T \begin{bmatrix} 5 & 5 & 1 & 1 \\ 5 & 5 & 1 & 1 \\ 1 & 1 & 10 & 41 \\ 1 & 1 & 41 & 10 \end{bmatrix} \alpha$$

(d)

$$\alpha^T \mathbf{1} - \frac{1}{2} \cdot \alpha^T \begin{bmatrix} 1 & 1 & 30 & 30 \\ 1 & 1 & 30 & 30 \\ 5 & 5 & 10 & 10 \\ 5 & 5 & 10 & 10 \end{bmatrix} \alpha$$

### Answer

(b)

### Solution

The objective function corresponding to the dual is:

$$\alpha^T \mathbf{1} - \frac{1}{2} \cdot \alpha^T (\mathbf{Y}^T \mathbf{X}^T \mathbf{X} \mathbf{Y}) \alpha$$

We have:

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & 5 & -5 \\ 0 & 0 & 4 & -4 \end{bmatrix}, \mathbf{Y} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

Therefore:

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} 1 & -1 & 5 & -5 \\ -1 & 1 & -5 & 5 \\ 5 & -5 & 41 & -41 \\ -5 & 5 & -41 & 41 \end{bmatrix}$$

And:

$$\mathbf{X}^T \mathbf{X} \mathbf{Y} = \begin{bmatrix} 1 & 1 & 5 & 5 \\ -1 & -1 & -5 & -5 \\ 5 & 5 & 41 & 41 \\ -5 & -5 & -41 & -41 \end{bmatrix}$$

Finally:

$$\mathbf{Y}^T \mathbf{X}^T \mathbf{X} \mathbf{Y} = \begin{bmatrix} 1 & 1 & 5 & 5 \\ 1 & 1 & 5 & 5 \\ 5 & 5 & 41 & 41 \\ 5 & 5 & 41 & 41 \end{bmatrix}$$

## Question-4 [1 point]

---

### Statement

What is the optimal weight vector,  $\mathbf{w}^*$ ?

**Hint:** Plot the points and try to compute the answer using geometry; do not try to solve the dual algebraically!

### Options

(a)

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

(b)

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

(c)

$$\begin{bmatrix} -1 \\ 0 \end{bmatrix}$$

(d)

$$\begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

### Answer

(a)

### Solution

We see that the  $x_2$  axis has to be the optimal separator as that is the one which has maximum margin. So, the equation of the decision boundary is  $x_1 = 0$ . This implies,  $w_2 = 0$ . Therefore, the weight vector becomes  $[w_1 \ 0]^T$ . The choice of  $w_1$  can be found out by noticing that  $x_1 = 1$  and  $x_1 = -1$  are the supporting hyperplanes. Therefore,  $w_1 \cdot 1 = 1 \implies w_1 = 1$ .

## Question-5 [1 point]

---

### Statement

Consider a kernel-SVM trained on a dataset of 100 points with polynomial kernel of degree 2. If  $\alpha^*$  is the optimal dual solution, what is the predicted label for a test-point  $\mathbf{x}_{\text{test}}$ ?

### Options

(a)

$$\sum_{i=1}^{100} \alpha_i^* \cdot \mathbf{x}_{\text{test}}^T \mathbf{x}_i \cdot y_i$$

(b)

$$\text{sign} \left( \sum_{i=1}^{100} \alpha_i^* \cdot \mathbf{x}_{\text{test}}^T \mathbf{x}_i \cdot y_i \right)$$

(c)

$$\sum_{i=1}^{100} \alpha_i^* \cdot (1 + \mathbf{x}_{\text{test}}^T \mathbf{x}_i)^2 \cdot y_i$$

(d)

$$\text{sign} \left( \sum_{i=1}^{100} \alpha_i^* \cdot (1 + \mathbf{x}_{\text{test}}^T \mathbf{x}_i)^2 \cdot y_i \right)$$

### Answer

(d)

### Solution

The optimal weight vector is given by:

$$\mathbf{w}^* = \sum_{i=1}^{100} \alpha_i^* \cdot \phi(\mathbf{x}_i) \cdot y_i$$

Here,  $\phi(\mathbf{x}_i)$  is the vector in the transformed space. First we compute the dot-product:

$$\begin{aligned} \mathbf{w}^{*T} \phi(\mathbf{x}_{\text{test}}) &= \sum_{i=1}^{100} \alpha_i^* \cdot \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_{\text{test}}) \cdot y_i \\ &= \sum_{i=1}^{100} \alpha_i^* \cdot k(\mathbf{x}_i, \mathbf{x}_{\text{test}}) \cdot y_i \\ &= \sum_{i=1}^{100} \alpha_i^* \cdot (1 + \mathbf{x}_{\text{test}}^T \mathbf{x}_i) \cdot y_i \end{aligned}$$

Finally, the prediction is:

$$\text{sign} \left( \mathbf{w}^{*T} \phi(\mathbf{x}_{\text{test}}) \right) = \text{sign} \left( \sum_{i=1}^{100} \alpha_i^* \cdot (1 + \mathbf{x}_{\text{test}}^T \mathbf{x}_i)^2 \cdot y_i \right)$$

## Common Data for questions (6) and (7)

---

### Statement

#### Common data for questions (6) and (7)

Consider the transformation  $\phi : \mathbb{R}^2 \rightarrow \mathbb{R}^6$  associated with the polynomial kernel with degree 2:

$$\phi(\mathbf{x}) = \begin{bmatrix} 1 \\ x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \end{bmatrix}$$

A kernel-SVM is trained on a dataset with the above kernel. The optimal weight vector is as follows:

$$\mathbf{w}^* = \begin{bmatrix} -25 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

You can assume that the dataset is linearly separable in the transformed space.

## **Question-6 [1 point]**

---

### **Statement**

What is the shape of the decision boundary in  $\mathbb{R}^2$ ?

### **Options**

**(a)**

It is a parabola of the form  $x_2 = 25x_1^2$

**(b)**

It is a parabola of the form  $x_2 = -25x_1^2$

**(c)**

It is a straight line

**(d)**

It is a circle

### **Answer**

(d)

### **Solution**

The decision boundary in  $\mathbb{R}^2$  is given by:

$$\mathbf{w}^{*T} \phi(\mathbf{x}) = 0$$

$$x_1^2 + x_2^2 = 25$$

It is a circle centered at the origin with radius 5.

## Question-7 [2 points]

---

### Statement

Which of the following training data-points are certainly **not** support vectors?

### Options

(a)

$$[-1 \quad 5]^T$$

(b)

$$[3 \quad 5]^T$$

(c)

$$[-4 \quad -5]^T$$

(d)

$$[2\sqrt{5} \quad -2]^T$$

(e)

$$[\sqrt{30} \quad \sqrt{6}]^T$$

### Answer

(b), (c), (e)

### Solution

All the support vectors will lie on the two supporting hyperplanes:

$$\mathbf{w}^* \phi(\mathbf{x}) = \pm 1$$

This gives two curves:

$$x_1^2 + x_2^2 = 24 \quad \text{OR} \quad x_1^2 + x_2^2 = 26$$

These curves are two circles, one smaller than the decision boundary and one larger than the decision boundary. From the points given here, there are two points that could be support vectors:

- $[-1 \quad 5]^T$ : this point lies on  $x_1^2 + x_2^2 = 26$
- $[2\sqrt{5} \quad -2]^T$ : this point lies on  $x_1^2 + x_2^2 = 24$

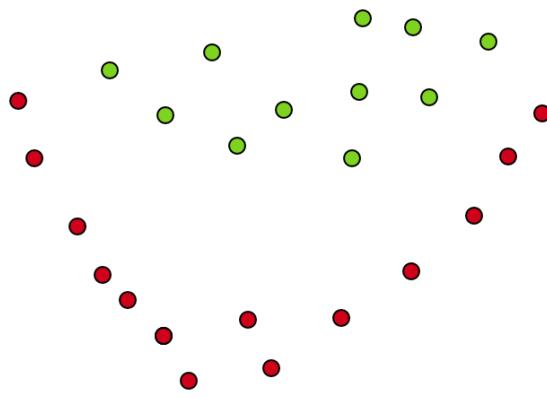
Note that these two are "potential" support vectors. Every support vector lies on the supporting hyperplanes. But every point on the supporting hyperplanes need not be a support vector.

## Question-8 [1 point]

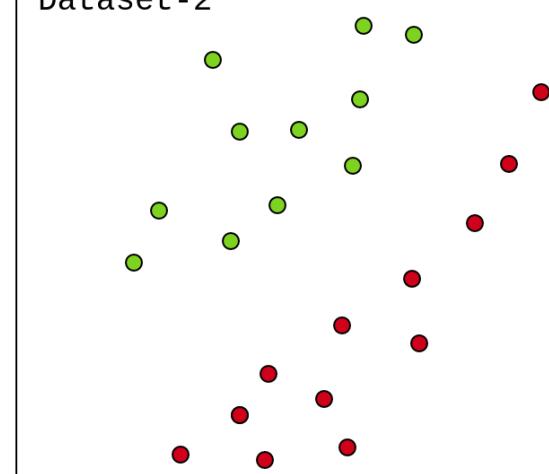
### Statement

Match the following classification datasets with the most appropriate choice of adjectives:

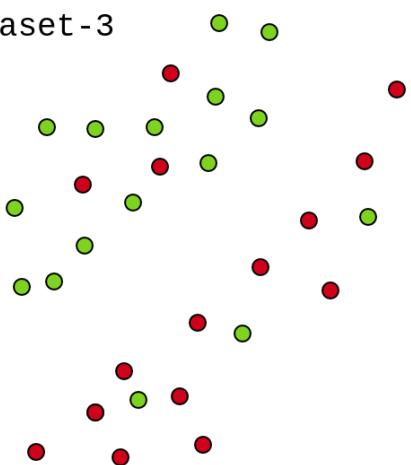
Dataset-1



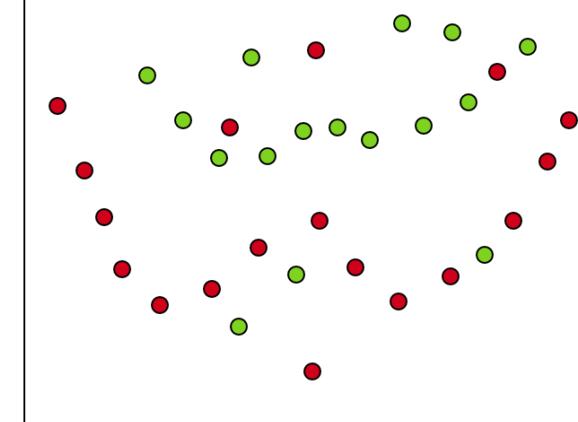
Dataset-2



Dataset-3



Dataset-4



### Options

(a)

Dataset-1: kernel, hard-margin

Dataset-2: hard-margin

Dataset-3: soft-margin

Dataset-4: hard-margin

(b)

Dataset-1: kernel, hard-margin

Dataset-2: hard-margin

Dataset-3: soft-margin

Dataset-4: kernel, soft-margin

**(c)**

Dataset-1: soft-margin

Dataset-2: hard-margin

Dataset-3: kernel

Dataset-4: soft-margin

## **Answer**

(b)

## **Solution**

- Dataset-1: the decision boundary is non-linear. The structure is non-linear and the problem is linearly separable in some high dimensional space.
- Dataset-2: this is a clear case of hard-margin SVM
- Dataset-3: the boundary is linear. The presence of outliers suggests that this should be solved using a soft-margin SVM
- Dataset-4: The boundary is non-linear. In addition, the dataset has outliers. So, this should involve both a kernel and a soft-margin formulation.

# Week 11

---

## Question 1

---

### Statement

In each round of AdaBoost, the weight for a particular training observation is increased going from round  $t$  to round  $t + 1$  if the observation was...

### Options

(a)

classified incorrectly by the weak learner trained in round  $t$

(b)

classified correctly by the weak learner trained in round  $t$

(c)

classified incorrectly by a majority of the weak learners trained up to round  $t$

(d)

classified correctly by a majority of the weak learners trained up to round  $t$

### Answer

(a)

### solution

Since in AdaBoost, we increase the weight of the incorrectly classified points for the next bag, the option (a) is correct.

---

## Question 2

---

### Statement

Which of the following statements are true about bagging?

### Options

(a)

In general, the final model has a higher bias than the individual learners.

**(b)**

In general, the final model has less bias than the individual learners.

**(c)**

In general, the final model has a higher variance than the individual learners.

**(d)**

In general, the final model has less variance than the individual learners.

## Answer

(d)

## Solution

Bagging on high variance models will reduce the variance without increasing the bias.

There is always a tradeoff between bias and variance. And reducing variance may cost increment in the bias. But bagging on high variance and low bias models reduces the variance without making the predictions biased.

## Question 3

---

### Statement

Is the following statement true or false?

If a point lies between the supporting hyperplanes in the soft-margin SVM problem, it always pays a positive bribe and plays a role in defining  $w^*$ .

### Options

**(a)**

True

**(b)**

False

## Answer

(a)

## Solution

If a point lies between the supporting hyperplanes, it satisfies the following:

$$w^{*T} x_i y_i < 1 \quad (1)$$

Using the 1<sup>st</sup> constraint,

$$\begin{aligned}
1 - w^{*T} x_i y_i - \xi_i^* &\leq 0 \\
\Rightarrow \xi_i^* &\geq 1 - w^{*T} x_i y_i
\end{aligned}$$

from 1, we can conclude that

$$\xi_i^* > 0$$

It implies that if a point lies between the supporting hyperplanes in the soft-margin SVM problem, it always pays a positive bribe

Using the CS 2,

$$\begin{aligned}
\beta_i^* \xi_i^* &= 0 \\
\Rightarrow \beta_i^* &= 0 \quad (\text{as } \xi_i^* > 0)
\end{aligned}$$

It implies that  $\alpha_i^* = C$  and therefore If a point lies between the supporting hyperplanes in the soft-margin SVM problem, it plays a role in defining  $w^*$ .

## Question 4

### Statement

Is the following statement true or false?

If  $i^{th}$  point in soft-margin SVM pays a non-zero bribe ( $\xi_i > 0$ ), then the value of  $\alpha_i$  is  $C$ .

### Options

(a)

True

(b)

False

### Answer

(a)

### Solution

Using the CS 2,

$$\begin{aligned}
\beta_i^* \xi_i^* &= 0 \\
\Rightarrow \beta_i^* &= 0 \quad (\text{as } \xi_i^* > 0)
\end{aligned}$$

It implies that  $\alpha_i^* = C$

## Common data for question 5 and 6

---

## Statement

Consider that an AdaBoost model is trained on the following binary classification dataset.

$x_1$	$x_2$	Label ( $y$ )
3.7	2	0
2.0	2	0
5	4	1
2.9	5	0
4.1	6	1

The first decision stump was created using the question  $x_2 < 4$  or not. The error of a decision stump is defined as the proportion of misclassified points.

## Question 5

---

### Statement

Find the value of  $\alpha_0$ . Notation is defined as per lecture.

### Options

(a)

$\ln 2$

(b)

$\ln 4$

(c)

$\ln(4/5)$

(c)

$\ln(\sqrt{3/2})$

### Answer

(a)

### Solution

If we split the root node as per question  $x_2 < 4$  or not, the left node will contain the points  $(3.7, 2), (2.0, 0)$  and the labels of these points are 0, 0 respectively. Therefore, the prediction in left node will be 0 (the majority class).

Similarly, in the right nodes, labels will be 1, 0, and 1 and the prediction will be 1 (the majority class).

Only one point  $(2.9, 5)$  is misclassified.

Therefore, error is  $e = \frac{1}{5}$

$$\begin{aligned}\alpha_0 &= \ln \sqrt{\frac{1-e}{e}} \\ &= \ln \sqrt{\frac{1-1/5}{1/5}} \\ &= \ln 2\end{aligned}$$

## Question 6

---

### Statement

How will the weight corresponding to the last example change for creating the next stump?

### Options

(a)

It will increase

(b)

It will decrease

### Answer

(b)

Since the last example is correctly classified, its weight will decrease.

## Question 7

---

### Statement

A strong learner  $L$  is formed as per the AdaBoost algorithm by three weak learners  $L_1, L_2$ , and  $L_3$ . Their performance/weights ( $\alpha$ ) are 1, 0.4, and 1.6, respectively. For a particular point,  $L_1$  and  $L_2$  predict that its label is positive, and  $L_3$  predicts that it's negative. What is the final prediction the learner  $L$  makes on this point? Enter 1 or  $-1$ .

### Answer

$-1$  No range is required

### Solution

For the final prediction, we have

$$\alpha_1 h_1(x) + \alpha_2 h_2(x) + \alpha_3 h_3(x) = 1 + 0.4 + 1.6(-1) = -0.2 < 0$$

Therefore, prediction will be  $-1$

## Question 8

---

## Statement

Which of the following options is correct? Select all that apply.

## Options

(a)

In bagging, typically around  $\frac{1}{3}rd$  data points remain unselected in bags if the number of data points is large.

(b)

Each weak learner has equal importance in making the final prediction in Bagging.

(c)

Each weak learner has equal importance in making the final prediction in AdaBoost.

(d)

Generally, weak learners in the random forest tend to overfit.

## Answer

(a), (b), (d)

## Solution

The probability that a point will not be selected in any one pick will be  $(1 - 1/n)$ .

The probability that a point will not be selected in  $n$  picks will be  $(1 - 1/n)^n$ .

as  $n \rightarrow \infty$ ,  $(1 - 1/n)^n \rightarrow 0.33$

that is In bagging, typically around  $\frac{1}{3}rd$  data points remain unselected in bags if the number of data points is large.

In bagging, each learner has equal importance in making the final prediction as the majority of all the predictions are taken into account, and therefore each prediction counts.

But in AdaBoost, the weighted average is taken into account, and the estimator which has a higher value of  $\alpha$  will have a higher importance in making the final prediction.

In the random forest, overfit models are preferred as they have high variance and low bias.

---

## Common data for questions 9, 10, and 11

---

## Statement

We have trained four models in the same dataset with different hyperparameters. In the following table, we have recorded the training and testing errors for each of the models.

Model	Training error	Test error
1	0.2	1.8

Model	Training error	Test error
2	1.0	1.1
3	0.5	0.7
4	1.9	2.3

## Question 9

---

### Statement

Which model tends to underfit?

### Options

(a)

Model 1

(b)

Model 2

(c)

Model 3

(d)

Model 4

### Answer

(d)

### Solution

Model 4 has high training error as well as high test error. It means that model 4 has high variance and high bias and tends to underfit.

---

## Question 10

---

### Statement

Which model tends to overfit?

### Options

**(a)**

Model 1

**(b)**

Model 2

**(c)**

Model 3

**(d)**

Model 4

## **Answer**

(a)

Model 1 has less training error as well as high test error. It means that model 4 has high variance and low bias and tends to overfit.

---

## **Question 11**

### **Statement**

Which model would you choose?

### **Options**

**(a)**

Model 1

**(b)**

Model 2

**(c)**

Model 3

**(d)**

Model 4

## **Answer**

(c)

Model 3 has less training and test error and therefore, it is most preferred.

# **Graded**

---

# Question-1

## Statement

Consider the following data and the hypothesis function  $h(x) = \text{sign}(g(x))$ :

<b>g(x)</b>	<b>y</b>
+30	+1
-20	-1
-1	-1
+1	+1

Which of the following will be true?

## Options

(a)

The values of 0-1 loss and squared loss will be same, which will be equal to zero.

(b)

The values of 0-1 loss and squared loss will be same, which will be equal to some large positive quantity.

(c)

The value of 0-1 loss will be zero, while the value of squared loss will be some large positive quantity.

(d)

The value of squared loss will be zero, while the value of 0-1 loss will be some large positive quantity.

## Answer

(c)

## Solution

There is an ambiguity in this question. The given  $h(x) = \text{sign}(g(x))$  is supposed to be explain the 0-1 loss and not the squared loss. In this case,

<b>g(x)</b>	<b>y</b>	<b>sign(g(x))</b>	<b>0-1 loss</b>	<b>Squared loss ((g(x)-y)^2)</b>
+30	+1	+1	0	$(29)^2$
-20	-1	-1	0	$(-19)^2$
-1	-1	-1	0	0

<b>g(x)</b>	<b>y</b>	<b>sign(g(x))</b>	<b>0-1 loss</b>	<b>Squared loss ((g(x)-y)^2)</b>
+1	+1	-1	0	0

The value of 0-1 loss is zero, while the value of squared loss is a large positive quantity. Hence, option (c) should be correct.

However, since  $h(x)$  is stated to be  $\text{sign}(g(x))$  and if squared loss is computed on the column 2 and 3 of the above table, squared loss will also come out to be 0, resulting in option (a) to be correct.

Hence, both options (a) and (c) will fetch marks.

## **Common instructions for questions 2-5**

Consider a neural network with the number of neurons in different layers as mentioned in the below list for a regression task:

[5, 5, 4, 3, 1]

### **Question-2**

#### **Statement**

How many hidden layers are there in the network?

#### **Answer**

3 (No range required)

#### **Solution**

The first layer is input layer and the last one is hidden layer. The intermediate 3 layers are hidden layers.

## Question-3

### Statement

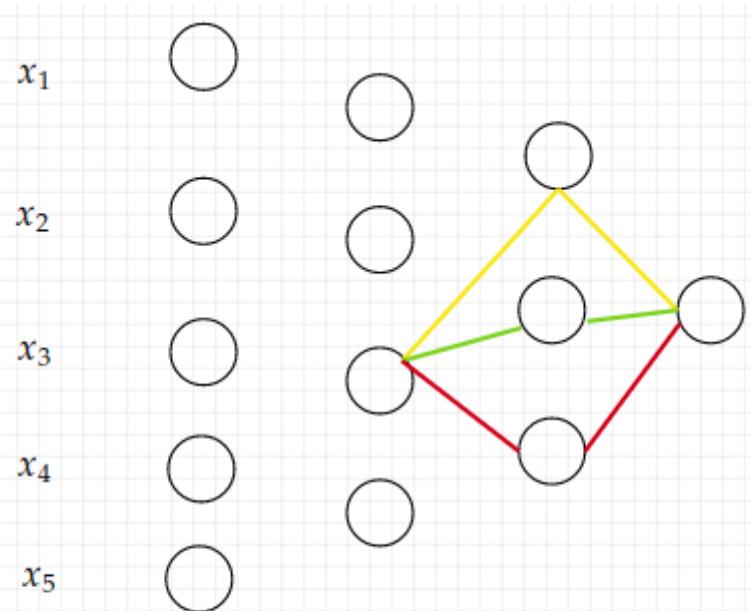
Through how many paths can the 3rd neuron in the 2nd hidden layer affect the final output?

### Answer

3 (No range required)

### Solution

The three paths are shown below in different colors.



input layer	Hidden layer 1	Hidden layer 2	Hidden layer 3	Output layer
-------------	----------------	----------------	----------------	--------------

## **Question-4**

### **Statement**

Assuming that there is a bias associated with each neuron, how many total parameters need to be computed?

### **Answer**

73 (No range required)

### **Solution**

#weights from input to hidden layer 1:  $5 \times 5 = 25$

#weights from hidden layer 1 to hidden layer 2:  $5 \times 4 = 20$

#weights from hidden layer 2 to hidden layer 3:  $4 \times 3 = 12$

#weights from hidden layer 3 to output layer:  $3 \times 1 = 3$

There will be a bias associated with each neuron except the input layer neurons. (Input layer is used to simply pass on the inputs to the subsequent layers)

#biases =  $5 + 4 + 3 + 1 = 13$

Total number of parameters =  $25 + 20 + 12 + 3 + 13 = 73$

## **Question-5**

### **Statement**

What will be an appropriate activation function for the output layer?

### **Options**

**(a)**

Sigmoid

**(b)**

Linear

**(c)**

ReLU

### **Answer**

(b)

### **Solution**

Sigmoid is used for binary classification at the output layer.

ReLU is mostly used at the hidden layers. (It does not make sense to be used at the output layer.)

Linear is an appropriate activation function for output layer in a regression problem.

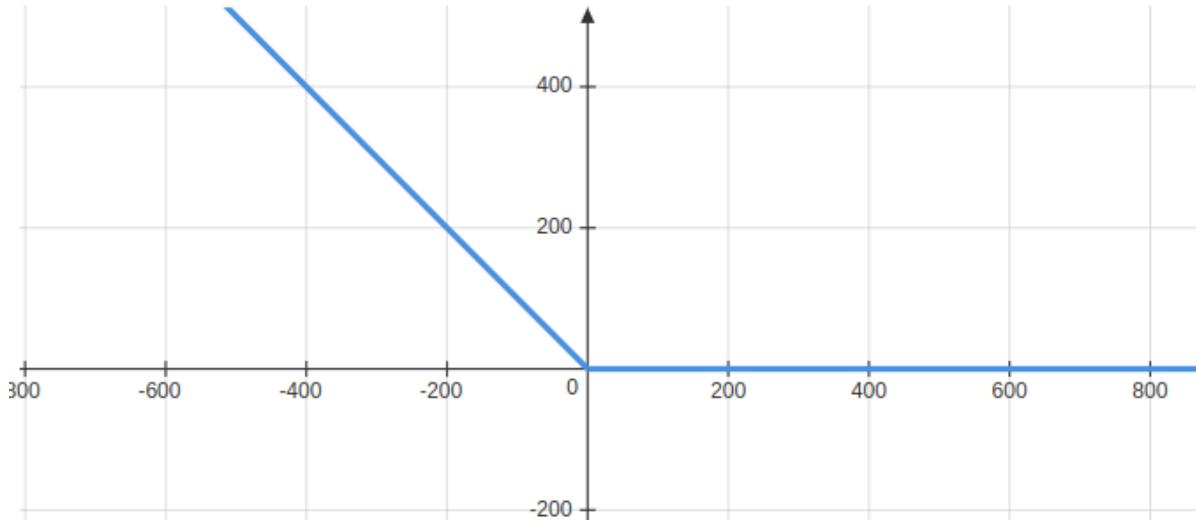
## Question-6

### Statement

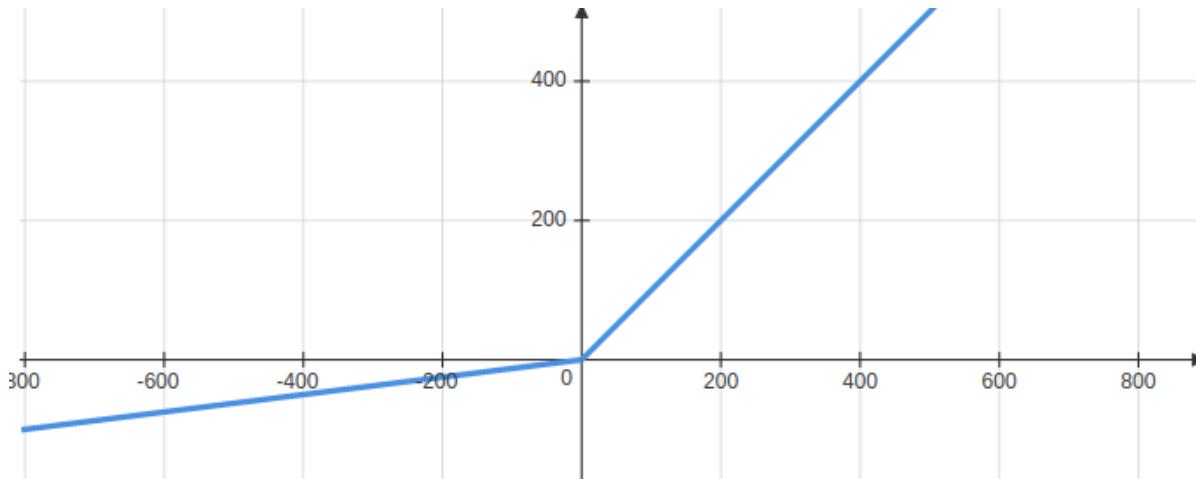
Which of the following represents ReLU activation function?

### Options

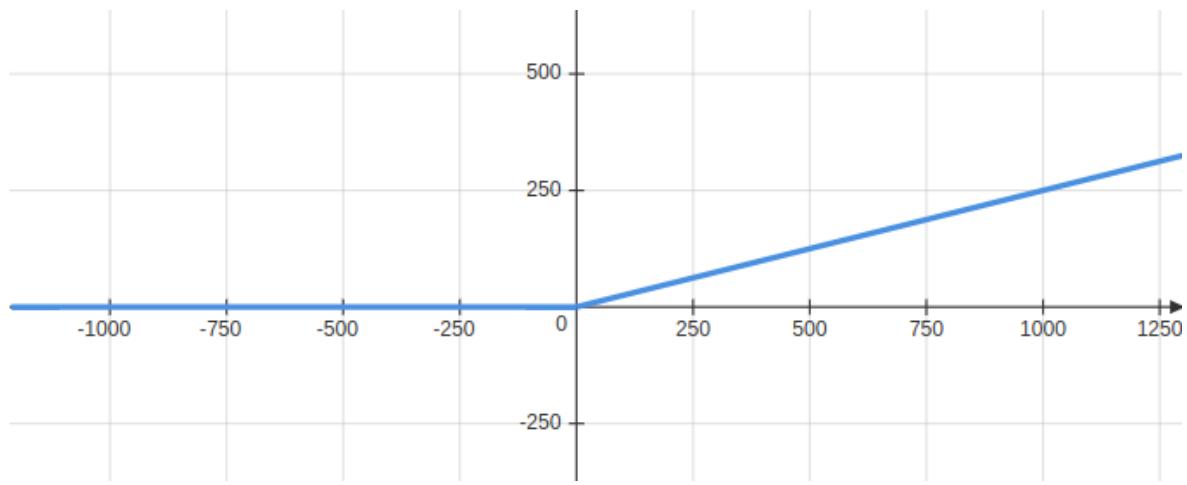
(a)



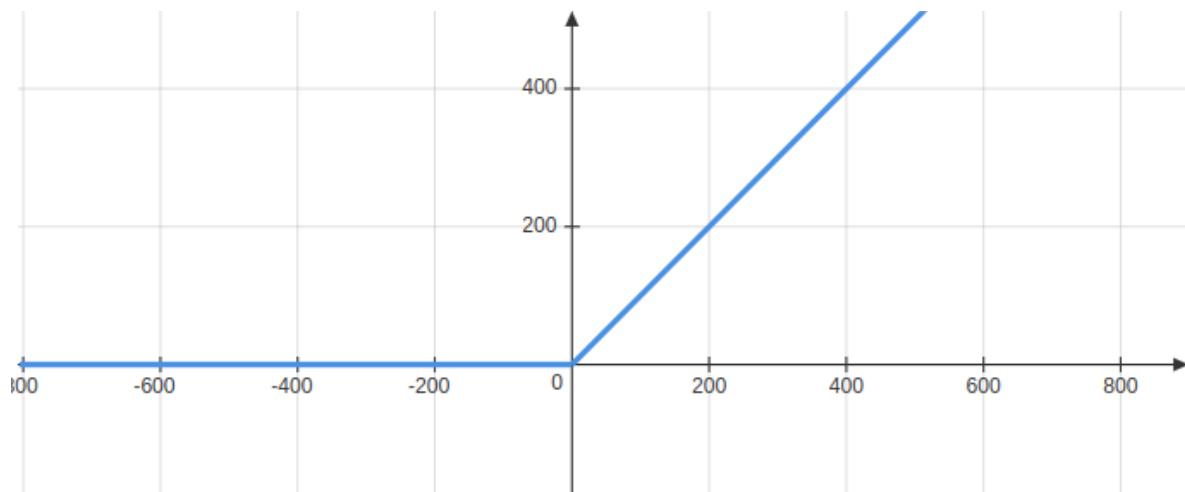
(b)



(c)



(d)



**Answer**

(d)

**Solution**

ReLU converts the negative inputs to zero, and keeps the positive inputs same.

## Question-7

---

### Statement

Suppose we build a neural network for a 5-class classification task. Suppose for a single training example, the true label is  $[1 \ 0 \ 0 \ 0 \ 0]$  while the predictions by the neural network are  $[0.1 \ 0.5 \ 0.1 \ 0.1 \ 0.2]$ . What would be the value of cross entropy loss for this example?

### Answer

3.322 (Range: 3.2 to 3.4)

### Solution

$$\begin{aligned}\text{Cross entropy} &= - \sum y_i \log_2 \hat{y}_i \\ &= -1 * \log_2(0.1) - 0 * \log_2(0.5) - 0 * \log_2(0.1) - 0 * \log_2(0.1) - 0 * \log_2(0.2) \\ &= -\log_2(0.1) = 3.322\end{aligned}$$

## **Question-8**

---

### **Statement**

State True or False:

If  $CE(y_1, y_2)$  represents the value of cross entropy loss, then  $CE(y_1, y_2) = CE(y_2, y_1)$  always.

### **Options**

**(a)**

True

**(b)**

False

### **Answer**

(b)

### **Solution**

Cross entropy ( $CE(y_1, y_2)$ ) =  $-\sum y_1 \log_2 y_2$  which is not commutative.