1. Part A, Qn(a)

Solution:

$$\begin{aligned} B(B-I) &= 0 \\ B^2 - B &= 0 \\ B^2 &= B \end{aligned}$$

We can find eigenvalues in the following way:

$$\begin{aligned} B^2\mathbf{x} &= B(B\mathbf{x}) \\ &= B(\lambda\mathbf{x}) \\ &= \lambda(B\mathbf{x}) \\ &= \lambda^2\mathbf{x} \end{aligned}$$

But, $B^2\mathbf{x} = B\mathbf{x} = \lambda\mathbf{x}$

$$\implies \lambda\mathbf{x} = \lambda^2\mathbf{x}$$

Since x is an eigenvector (non-zero),

$$\begin{aligned} \lambda &= \lambda^2 \\ \lambda(\lambda - 1) &= 0 \end{aligned}$$

This gives us eigenvalues of B as either 1 or 0.

2. Part A, Qn(b)

Solution: Let the random variable X indicate at least one heads observed. Similarly, let random variable Y indicate at most two heads observed. The support for X is $\mathbb{R}_X = \{1, 2, 3\}$. The support for Y is $\mathbb{R}_Y = \{0, 1, 2\}$. We know $P(X) = \frac{7}{8}$ and $P(Y) = \frac{7}{8}$. We have to find $p_{Y|X}(y|x)$ and we do in the following way:

$$\begin{aligned} p_{Y|X}(y|x) &= \frac{p_{XY}(x, y)}{p_X(x)} \\ p_{Y|X}(y|x) &= \frac{6/8}{7/8} \\ p_{Y|X}(y|x) &= \frac{6}{7} \end{aligned}$$

3. Part A, Qn(c)

Solution: Let the likelihood function to estimate $\hat{a}_{ML}$ be defined as follows:

$$L(x_1, \ldots, x_6, \hat{a}) = P(x_1, \ldots, x_6; \hat{a})$$

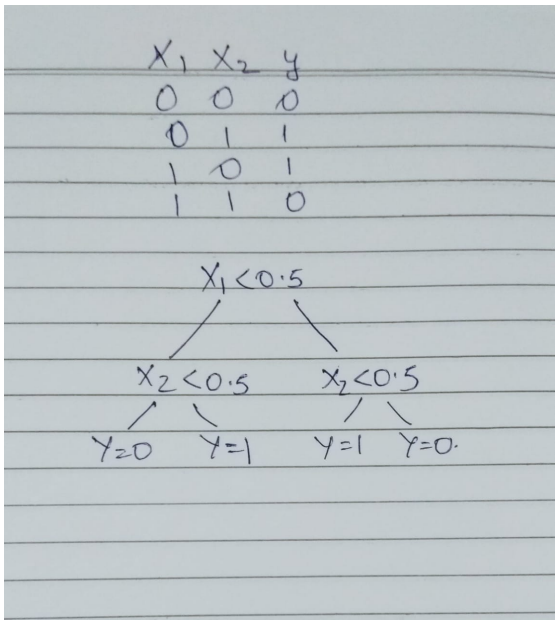
$\because$ data-points come from a continuous distribution,

$$\begin{aligned} &= f_{x_1,\ldots,x_6}(x_1, \ldots, x_6; \hat{a}) \\ &= f_{x_1}(x_1; \hat{a}) \ldots f_{x_6}(x_6; \hat{a}) \\ &= \prod_{i=1}^{6} (1 \cdot \frac{1}{\hat{a}} \mid x_i \in [0, \hat{a}]) \end{aligned}$$

Here, we can see that even if one of the data-points false out of distribution range, the likelihood will be zero. Hence, to maximize the likelihood, all the points must exists in between the uniform distribution range. For this to happen, $\hat{a}_{ML} = \max\{x_1, \ldots, x_6\} = 30.2$.

4. Part A, Qn(d)

Solution: This statement is false. Consider the XOR problem, where we the two features are: $X_1 = \{0, 0, 1, 1\}$ and $X_2 = \{0, 1, 0, 1\}$. Also, $y = \{0, 1, 1, 0\}$. As per the claim given, max height of the tree will be $log(2)$. However, we can see a decision tree take two steps to classify all the points correctly. Hence, the claim is wrong.



5. Part A, Qn(e)

Solution: We needn't calculate the co-variance matrix. We can see that x_2 and x_3 are scalar multiples of x_1. Hence, λ_2 and λ_3 will be 0. Hence, the value of the fraction becomes 1.

6. Part A, Qn(f)

Solution: We compute MLE estimator $\hat{\mu}_{ML}$ as follows:

$$\max_{\mu} L(D_1, D_2; \mu) = \prod_{i=1}^{n} f_X(x_i; \mu, 1) \cdot \prod_{i=1}^{n} f_Y(y_i; \mu, 2)$$

$$\max_{\mu} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \cdot e^{-(x_i-\mu)^2/2} \cdot \prod_{i=1}^{n} \frac{1}{2\sqrt{2\pi}} e^{-(y_i-\mu)^2/4}$$

Taking log on both sides and removing constant terms, we get

$$\max_{\mu} = -\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2} + \frac{(y_i-\mu)^2}{4}$$

$$\min_{\mu} = \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2} + \frac{(y_i-\mu)^2}{4}$$

$$\min_{\mu} = \frac{1}{2}\left(\sum_{i=1}^{n} x_i^2 + \mu^2 - 2x_i\mu + 0.5y_i^2 + 0.5\mu^2 - y_i\mu\right)$$

Differentiating w.r.t μ and equating to 0, we get,

$$\implies \sum_{i=1}^{n} 2\mu - 2x_i + \mu - y_i = 0$$

$$\implies 3\mu - 2\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = 0$$

$$\implies \mu = \frac{2\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}{3}$$

$$\implies \mu = \frac{2\bar{x} + \bar{y}}{3}$$

$$\implies \mu = \frac{2}{3}\bar{x} + \frac{1}{3}\bar{y}$$

7. Part A, Qn(g)

Solution: We have given the radius of all our datapoints as $R = 1$. We have also been given the margin $\gamma = 0.5$ with which all datapoints must be linearly separable. Number of mistakes the perceptron makes with these parameters is given by:

$$\begin{aligned} &\#\text{ mistakes} < \frac{R^2}{\gamma^2} \\ \implies &\#\text{ mistakes} < \frac{1}{0.25} \\ \implies &\#\text{ mistakes} < 4 \end{aligned}$$

Maximum number of mistakes is equal to 3.

8. Part A, Qn(h)

Solution: For any dataset S and a hypothesis class $\mathcal{H}$, the empirical risk minimizer is:

$$h_S = \arg\min_{h \in \mathcal{H}} \text{er}_S[h]$$

For instance, for a squared loss function, and linear functions $y = w^T x$ as $\mathcal{H}$, linear regression is the learning algorithm. However, it isn't necessarily the learning algorithm for other hypothesis classes such as Logistic regression, SVM, etc. Hence, the statement is False.

9. Part A, Qn(i)

Solution: VC dimension of such a hypothesis class is 3. Any circle can classify a single data-point correctly. Similarly, when we have two datapoints, any circle can classify the data-points, regardless of its labeling. Now, when we have 3 data-points and we place it in the form of a triangle, where each axis has one data-point, then yet again, regardless of labeling permutations, any circle can classify all points correctly. However, when we have 4 data-points, regardless of how I position my data-points, my adversary can always choose a specific permutation which will lead to misclassification by my classifier. Hence, the VC dimension is 3.

10. Part A, Qn(j)

Solution: Our label probability $\hat{p} = 0.5$ as there are equal number of positive labeled and negative labeled datapoints. We can calculate the probability of a data-point j appearing either in positive or negative class using the following formula:

$$\hat{p}_j^{-1} = \frac{\sum_{i=1}^{n} 1 \cdot (f_j, y_i = -1)}{\sum_{i=1}^{n} 1 \cdot (y_i = -1)}$$

$$\hat{p}_j^{+1} = \frac{\sum_{i=1}^{n} 1 \cdot (f_j, y_i = 1)}{\sum_{i=1}^{n} 1 \cdot (y_i = 1)}$$

p_1^{-1}	p_2^{-1}	p_3^{-1}	p_4^{-1}	p_5^{-1}	p_6^{-1}
$\frac{3}{3}$	$\frac{3}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{3}{3}$

p_1^{1}	p_2^{1}	p_3^{1}	p_4^{1}	p_5^{1}	p_6^{1}
$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{3}{3}$	$\frac{3}{3}$	$\frac{3}{3}$

Next, we calculate the following:

$$P(y_t = -1|x_t) = \left(\prod_{j=1}^{n} (\hat{p}_j^{-1})^{f_j^t} \ (1 - \hat{p}_j^{-1})^{1-f_j^t} \right) \cdot \hat{p}$$

$$P(y_t = -1|x_t) = \frac{1}{54}$$

Also,

$$P(y_t = +1|x_t) = \left(\prod_{j=1}^{n} (\hat{p}_j^{+1})^{f_j^t} \ (1 - \hat{p}_j^{+1})^{1-f_j^t} \right) \cdot \hat{p}$$

$$P(y_t = +1|x_t) = \frac{1}{27}$$

Since $P(y_t = +1|x_t) > P(y_t = -1|x_t)$, the classifier will predict $y_{test} = +1$.

11. Part B(Qn1)

Solution: We know that $C = XX^T$ is a symmetric positive semi-definite matrix. This can be seen in the following way. Let $u \in \mathbb{R}^d$. Then, u^T

$$u^T C u = u^T X X^T u = (u^T X)(u^T X)^T = (u^T X)^2 \geq 0$$

Similarly, for our kernel matrix K,

$$\begin{aligned} K &= \mathbf{x}^T C \mathbf{y} \\ K &= \mathbf{x}^T X X^T \mathbf{y} \\ K &= (X^T \mathbf{x})^T (X^T \mathbf{y}) \end{aligned}$$

From the above result, we can define the function $\phi(x) = X^T x = u$ and $\phi(z) = X^T z = v$. Therefore, this gives us:

$$\begin{aligned} K &= \phi(x)^T \phi(z) \\ K &= u^T v \end{aligned}$$

This makes K a linear kernel which is a valid kernel since eigenvalues of K are non-negative.

12. Part B(Qn2)

Solution:
Qn (i)
Yes

Qn (ii)
After projecting original datapoints onto the first PC, which is approx. $(-1, 0)$, the new dataset is $\{5, 1, 3, -1, -3, -5\}$.

Qn (iii)
No, the data is not linearly separable.

13. Part B(Qn3)

Solution:
Qn (i)
For $K = 1$ and class -1, x_2 and x_4 should have more points closer to itself labeled as -1 than points labeled as $+1$. For x_2, this can happen when x_2 falls in the intersection of region $A \implies (x, y) : y = x \cap y = -x \cap y \geq 0$. Similarly, for x_4, region $B \implies (x, y) : y = x \cap y = -x \cap y \leq 0$. Hence, for class -1, the total decision region will be $A \cup B$.

Qn (ii)
For x_2, the labels will flip since for $K = 3$, x_1 and x_3, labeled as $+1$ is closer than x_4. Similarly, for x_4, x_1 and x_3, labeled as $+1$ is closer than x_2. Hence, for class -1, the total decision region will be $(A \cup B)^C$.

14. Part B(Qn4)

Solution:
Qn (i)
We want a weight vector w for which all datapoints satisfy the condition: $(w^T x_i)y_i \geq 1$. One such $w = [1, 1]$. Bias will be equal to 0.

Qn (ii)
Initialize $w = [0, 0]$. We predict $+1$ if $w^T x > 0$ and -1 if $w^T x \leq 0$. With the initial weights, the first two data-points are classified correctly since $w^T x_1 \leq 0$ and $w^T x_2 \leq 0$. For x_3, $w^T x_3 = 0$ and hence, prediction is -1. Since the true prediction is $+1$, the new weight vector is:

$$w = [0, 0] + x_3 * y_3 = [0, 1]$$

Using this new weight, all points are classified in further iterations. Hence, total mistake is 1 and the weight vector is $[0, 1]$.

Qn (iii)
Initialize $w = [0, 0]$. We predict $+1$ if $w^T x > 0$ and -1 if $w^T x \leq 0$. With the initial weights, x_4 is predicted correctly since $w^T x_4 \leq 0$. For x_3, $w^T x_3 = 0$ and hence, prediction is -1. Since the true prediction is $+1$, the new weight vector is:

$$w = [0, 0] + x_3 * y_3 = [0, 1]$$

Using this new weight, all points are classified in further iterations. Hence, total mistake is 1 and the weight vector is $[0, 1]$.

15. Part B(Qn5)

Solution:

Qn (i)

We know that the Bayes error is the expected loss of the Bayes classifier when prediction is $+1$ since $P(Y = 1|X) > P(Y = -1|X)$. For the above problem, when samples come from $[0, 1]$, Bayes classifier predicts $+1$ and no error is made. Similarly, when samples come from $(2, 4]$, the classifier correctly predicts -1 without any incorrect classification. Therefore, for both these regions, Bayes error is 0. Now, since in the region $[1, 2]$, there is a overlap, we need to see whether the classifier predicts $+1$ or -1. To check this, we calculate,

$$\begin{aligned} P(Y = 1|X) &\propto f(X|Y = 1)P(Y = 1) \\ &\propto \frac{1}{2} \cdot \frac{1}{2} \\ &\propto \frac{1}{4} \end{aligned}$$

Now, for $P(Y = -1|X)$,

$$\begin{aligned} P(Y = -1|X) &\propto f(X|Y = -1)P(Y = -1) \\ &\propto \frac{1}{3} \cdot \frac{1}{2} \\ &\propto \frac{1}{6} \end{aligned}$$

We see that $P(Y = 1|X) > P(Y = -1|X)$. Therefore, in the region $[1, 2]$, the classifier will always predict $+1$. Thus,

$$\begin{aligned} \text{Bayes Error} &= \int_1^2 P(Y = -1|X) \cdot f_X(x) \cdot dx \\ \text{Bayes Error} &= \int_1^2 \frac{1}{2} \cdot f_X(x) \cdot dx \\ \text{Bayes Error} &= \frac{1}{2} \cdot \frac{2-1}{4-1} \\ \text{Bayes Error} &= \frac{1}{6} \end{aligned}$$

Qn (ii)

Given $P(Y = 1) = 0.75$,

$$\begin{aligned} P(Y = 1|X) &\propto f(X|Y = 1)P(Y = 1) \\ &\propto \frac{1}{2} \cdot \frac{3}{4} \\ &\propto \frac{3}{8} \end{aligned}$$

Still, we see that $P(Y=1|X) > P(Y=-1|X)$. Therefore, in the region $[1,2]$, the classifier will always predict $+1$. Thus,

$$\text{Bayes Error} = \int_1^2 P(Y=-1|X) \cdot f_X(x) \cdot dx$$
$$\text{Bayes Error} = \int_1^2 \frac{1}{4} \cdot f_X(x) \cdot dx$$
$$\text{Bayes Error} = \frac{1}{4} \cdot \frac{2-1}{4-1}$$
$$\text{Bayes Error} = \frac{1}{12}$$

We see that Bayes error hasn't increased.

Qn (iii)
Given $P(Y=1) = p$ and $x = 1.5$, for the classifier to predict -1, the following condition must hold,

$$\begin{aligned} & P(Y=-1|X=x) > P(Y=1|X=x) \\ \implies & f(X|Y=-1)P(Y=-1) > f(X|Y=1)P(Y=1) \\ \implies & \frac{1}{3} \cdot (1-p) > \frac{1}{2} \cdot p \\ \implies & \frac{1-p}{3} > \frac{p}{2} \\ \implies & 2-2p > 3p \\ \implies & p < \frac{2}{5} \\ \implies & p < 0.4 \end{aligned}$$

16. Part B(Qn6)

Solution:
Qn (i)
At step 1, $\mu_1^1 = \frac{f_3+f_4+f_7+f_8}{4} = -1.5$ and $\mu_2^1 = \frac{f_1+f_2+f_5+f_6}{4} = 1.5$. After step 1, we see that all negative points are close to μ_1^1 and all positive points are closer to μ_2^1. Hence, in step 2, cluster 1 will be $\{f_3, f_4, f_6, f_8\}$ and cluster 2 will be $\{f_1, f_2, f_5, f_7\}$.

Qn (ii)
Initially, we have equal number of friends in each clusters. Hence, probability assign-

ment to each cluster depends only on the densities provided in the table. Without calculating, we can compare densities for the initial means, $\mu_1^1 = -1.5$ and $\mu_2^1 = 1.5$. Friends with higher densities can be assigned to their respective clusters. Hence, after one step, cluster 1 will be $\{f_3, f_4, f_6, f_8\}$ since they have higher densities at μ_1^1 than with densities at μ_2^1.