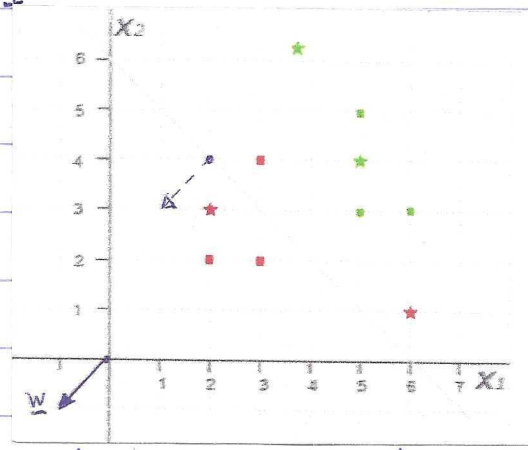
Question 1 (as intended)

a. Equation of separating line:

$w_1 x_1 + w_2 x_2 - w_0 = 0, \quad w_0 = -6$

$x_1 = 0 \Rightarrow x_2 = 6$, therefore $\quad w_2 = -6/6 = -1$

$x_2 = 0 \Rightarrow x_1 = 6$, therefore $\quad w_1 = -6/6 = -1$

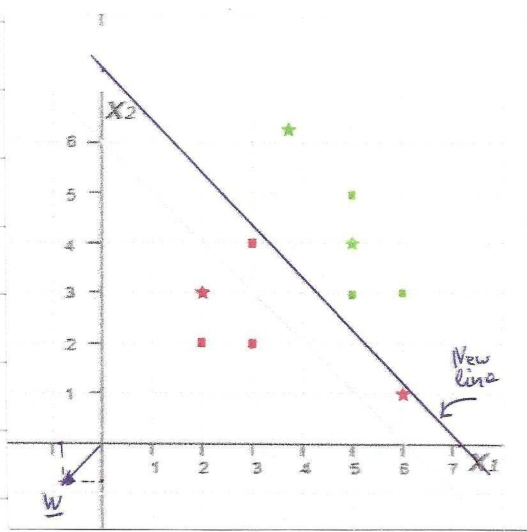


b.

w_1	w_2	w_0	tx_1	tx_2	tx_0	$t\underline{w}\cdot\underline{x}$	Upd.	Δw_1	Δw_2	Δw_0
-1	-1	-6	2	2	-1	+2	No	0	0	0
-1	-1	-6	3	2	-1	+1	No	0	0	0
-1	-1	-6	3	4	-1	-1	Yes	0,15	0,20	-0,05
-0,85	-0,80	-6,05	-5	-3	1	+0,6	No	0	0	0
-0,85	-0,80	-6,05	-6	-3	1	+1,45	No	0	0	0
-0,85	-0,80	-6,05	-5	-5	1	+2,20	No	0	0	0

c. Separating line: $\quad -0{,}85 x_1 - 0{,}80 x_2 + 6{,}05 = 0$

$x_1 = 0 \Rightarrow x_2 = \dfrac{6{,}05}{0{,}80} = 7{,}56, \quad x_2 = 0 \Rightarrow x_1 = \dfrac{6{,}05}{0{,}85} = 7{,}12$



d. Before: Training set: TP = 2, FP = 0, TN = 3, FN = 1.

Precision: $\dfrac{TP}{TP+FP} = \dfrac{2}{2+0} = 1$ (100%)

Recall: $\dfrac{TP}{TP+FN} = \dfrac{2}{2+1} = \dfrac{2}{3}$ (66,7%)

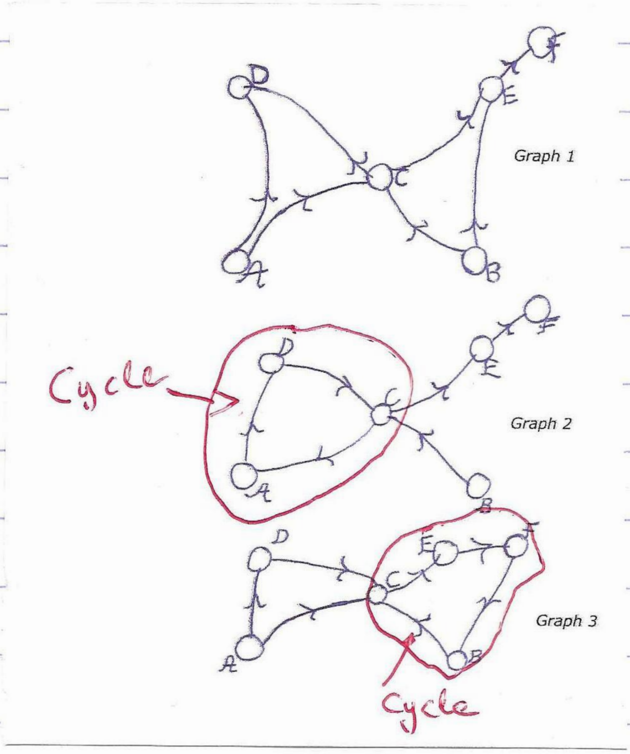
Test set: TP = 1, FP = 0, TN = 2, FN = 1.

Precision: 100%, Recall: 50%.

After: All instances classified correctly.
All success measures equal to 100%.

Question 2

a. Graph 1 is a directed acyclic graph, while graphs 2 and 3 contain cycles as shown:



Therefore, only graph 1 can represent a Bayesian network.

b. For graph 1 we need:

$P(A_1)$

$P(B_1)$

$P(C_1|A_i, B_j)$, $i=0,1, j=0,1$ (4 quantities).

$P(E_1|B_i, C_j)$, $i=0,1$, $j=0,1$.

$P(D_1|A_i C_j)$, $i=0,1$, $j=0,1$

$P(F_1|E_i)$, $i=0,1$.

c. $P(A,B,C,D,E,F) = P(F|E)\cdot P(E|C,B)\cdot P(D|A,C)\cdot P(C|A,B)\cdot P(B)\cdot P(A)$

d. d1) $P(F_1|A_1) = P(A_1,F_1)/P(A_1) = \sum_{\text{All } B,C,E} P(A_1,B,C,E,F_1)/P(A_1)$

$$= \frac{\sum_{\text{All } B,C,E} P(F_1|E)\cdot P(E|B,C)\cdot P(C|A_1,B)\cdot P(B)\cdot \cancel{P(A_1)}}{\cancel{P(A_1)}}$$

$$= \sum_{\text{All } B,C,E} P(F_1|E)\cdot P(E,B,C)\cdot P(C|A_1,B)\cdot P(B).$$

d2) $$P(A_1|F_1) = P(A_1,F_1)/P(F_1) = \frac{\sum_{\text{All } B,C,E} P(F_1|E)\cdot P(E|B,C)\cdot P(C|A_1,B)\cdot P(B)\cdot P(A_1)}{\sum_{\text{All } A,B,C,E} P(F_1|E)\cdot P(E|B,C)\cdot P(C|A,B)\cdot P(B)\cdot P(A)}$$

Alternatively, the denominator: $P(F_1) = P(F_1|A_1)\cdot P(A_1) + P(F_1|A_0)\cdot P(A_0)$
$= P(F_1|A_1)\cdot P(A_1) + P(F_1|A_0)\cdot(1-P(A_1))$, with $P(F_1|A_0)$ calculated as in d1.

Question 3

a. The isolevel curves for the two pdfs $p(\underline{x}|A)$ and $p(\underline{x}|B)$ have different shapes, signifying that the corresponding covariance matrices for the two Gaussians are not equal. Therefore, this is a non-linear classifier.

b. Both sets of ellipses have their principal axes along the x_1 and x_2 coordinate axes. Therefore the corresponding covariance matrices are diagonal, signifying that the two features are statistically independent. [This will become more apparent in the following subsections]. This is a naive Bayes classifier.

c. The eigenvalues are proportional to the squares of the length of the principal axes of the ellipses.
For class A (bottom left of the diagram), we get
$\lambda_{1A} = 1, \ \lambda_{2A} = 2^2 = 4.$
For class B (top right), we get $\lambda_{1B} = 2^2 = 4, \ \lambda_{2B} = 1.$
The eigenvectors follow the direction of the principal axes:
$\underline{v}_{1A} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \ \underline{v}_{2A} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \ \underline{v}_{1B} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \ \underline{v}_{2B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$

d. The covariance matrices: $\Sigma_A = 1 \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix}[1 \ 0] + 4\begin{bmatrix} 0 \\ 1 \end{bmatrix}[0 \ 1] = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}.$

$$\Sigma_B = 4\begin{bmatrix} 1 \\ 0 \end{bmatrix}[1 \ 0] + 1\begin{bmatrix} 0 \\ 1 \end{bmatrix}[0 \ 1] = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$

e. Since the two classes are equiprobable, the separating curve is given by $p(\underline{x}|A) = p(\underline{x}|B)$.
With $\Sigma_A^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1/4 \end{bmatrix}$ and $\Sigma_B^{-1} = \begin{bmatrix} 1/4 & 0 \\ 0 & 1 \end{bmatrix}$ and equating the logarithms of $p(\underline{x}|A)$ and $p(\underline{x}|B)$ we obtain:

$$[x_1 \ x_2]\begin{bmatrix} 1 & 0 \\ 0 & 1/4 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \cancel{+ \det \Sigma_1} = [x_1-4, x_2-4]\begin{bmatrix} 1/4 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_1-4 \\ x_2-4 \end{bmatrix} \cancel{+ \det \Sigma_2} \Rightarrow$$

$$\Rightarrow x_1^2 + \frac{1}{4}x_2^2 = \frac{1}{4}(x_1-4)^2 + (x_2-4)^2.$$

Note that there are no terms involving $x_1 \cdot x_2$, signifying again feature independence and verifying that this is indeed equivalent to a naive Bayes classifier.

f. For point $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$, $x_1^2 + \frac{1}{4}x_2^2 = 9 + \frac{1}{4} = 9{,}25.$

$$\frac{1}{4}(x_1-4)^2 + (x_2-4)^2 = \frac{1}{4} + (1-4)^2 = \frac{1}{4} + 9 = 9{,}25.$$

Our point is on the border between the two classes.

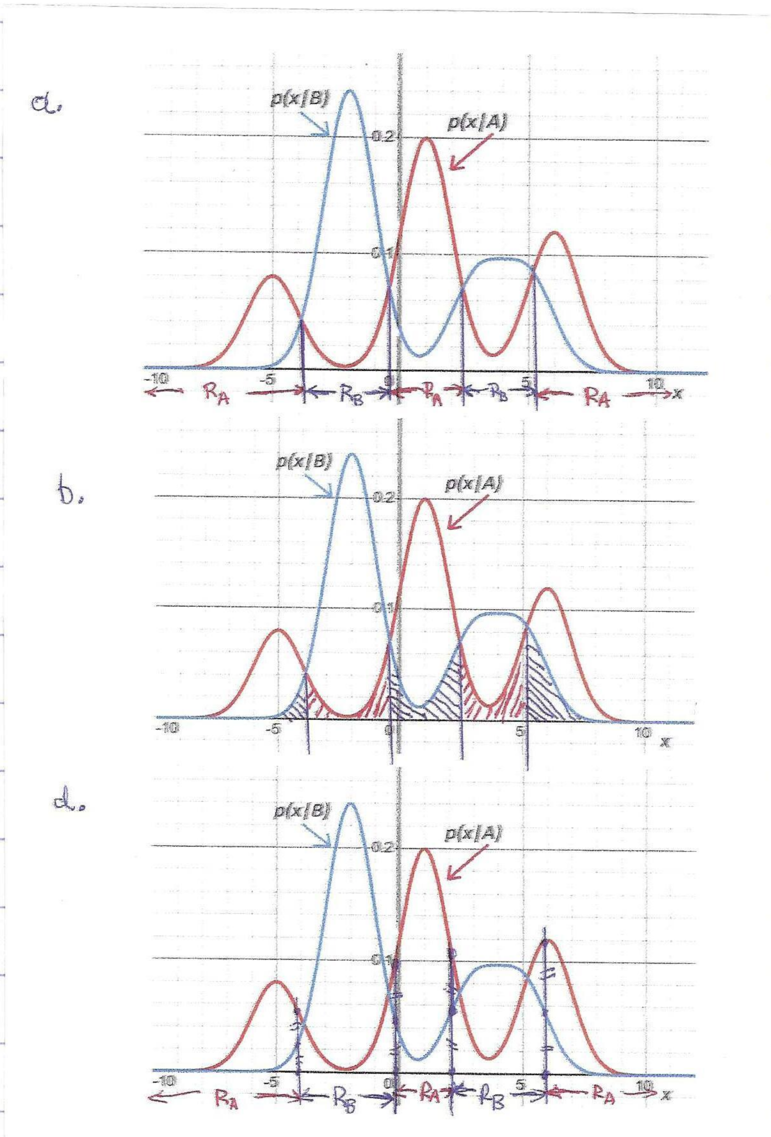
Question 4.

a. R_A consists of regions where $p(x|A) > p(x|B)$
and R_B consists of regions where $p(x|B) > p(x|A)$.

b. Shaded areas as shown in the diagram.

c. There are approximately 10 blue-shaded squares
and 9 red-shaded squares.
Therefore $P_e \cong 0{,}5 \cdot 10 \cdot 0{,}02 + 0{,}5 \cdot 9 \cdot 0{,}02 = 0{,}19 = 19\%$.

d. Now R_A consists of regions where $p(x|A) > 2p(x|B)$
and R_B consists of regions where $2p(x|B) > p(x|A)$.
The designated segments on the diagram are equal. The blue "tails" are smaller than before.



Question 5. $x_1 = 1,\ x_2 = 2,\ x_3 = 3,\ x_4 = 6$, true value: $\theta_0 = 2,\ N = 4$

a. $\mu = \dfrac{1+2+3+6}{4} = 3$

$$\sigma^2 = \frac{\sum_i x_i^2}{4} - \mu^2 = \frac{1+4+9+36}{4} - 9 = \frac{14}{4} = 3{,}5 \Rightarrow \sigma = 1{,}87.$$

b. Squared bias: $(\mu - \theta_0)^2 = (3-2)^2 = 1.$

Variance: $\sigma^2 = 3{,}5.$

MSE = Squared bias + Variance = $1 + 3{,}5 = 4{,}5.$

c. We need to maximize the a posteriori probability:

$$p(X|\mu,\sigma^2)\cdot p(\mu) = \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]\right] \cdot \frac{1}{\mu_0} \exp(-\mu/\mu_0),\ \mu > 0.$$

We take the logarithm:

$$\mathcal{L} = -\frac{1}{2\sigma^3} \sum_{i=1}^{N} (x_i-\mu)^2 - \frac{\mu}{\mu_0} + C,$$ assuming that σ is known.

Maximize:

$$\frac{\partial \mathcal{L}}{\partial \mu} = 0 \Rightarrow \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \mu) - \frac{1}{\mu_0} = 0 \Rightarrow \sum_{i=1}^{N} x_i - N\mu = \frac{\sigma^2}{\mu_0}$$

$$\Rightarrow \mu_{MAP} = \frac{\sum_{i=1}^{N} x_i}{N} - \frac{\sigma^2}{N\mu_0} = 3 - \frac{3.5}{4} = 2{,}125.$$

The new evaluation for the squared bias is:

$$(\mu_{MAP} - \theta_0)^2 = (2{,}125 - 2)^2 \simeq 0.016$$