Anar Rzayev — MAS473: Homework Set 2 — ID: 20190788

2)

Class 1	Class 2
0110	1011
1010	0000
0011	0100
1111	1110

Denote the given 4 binary attributes A_1, A_2, A_3, A_4, and visualize them as follows:

A_1	A_2	A_3	A_4	
0	1	1	0	class 1
1	0	1	0	class 1
0	0	1	1	class 1
1	1	1	1	class 1
1	0	1	1	class 2
0	0	0	0	class 2
0	1	0	0	class 2
1	1	1	0	class 2

Our first strategy is to determine Information Gain in order to obtain the desired root node

In the part depth=0 ⟹ we wish to evaluate total entropy, in which there are 8 data points and 4 respective attributes with 2 classes $\rightsquigarrow -\frac{4}{8}\log\left(\frac{4}{8}\right)-\frac{4}{8}\log\left(\frac{4}{8}\right)=-\frac{1}{2}\log\left(\frac{1}{2}\right)\times 2=$

$=-\log(2^{-1})=\underline{1}$

Now, counting 1's for each correspondent attribute where measuring # of 1's in each class will yield the following Information Gain results for A_1, A_2, A_3, A_4:

$$\underline{A_1} \rightsquigarrow 1-\frac{4}{8}\times\left(-\frac{2}{4}\log\left(\frac{2}{4}\right)-\frac{2}{4}\log\left(\frac{2}{4}\right)\right)-\frac{4}{8}\times\left(-\frac{2}{4}\log\left(\frac{2}{4}\right)-\frac{2}{4}\log\left(\frac{2}{4}\right)\right)$$

$$=1-\frac{1}{2}\times\left(-\frac{1}{2}\log(2^{-1})-\frac{1}{2}\log(2^{-1})\right)-\frac{1}{2}\left(-\frac{1}{2}\log(2^{-1})\times 2\right)=$$

$$=1-\frac{1}{2}\times\left[-\log(2^{-1})\right]-\frac{1}{2}\times\left[-\log(2^{-1})\right]=1-\frac{1}{2}-\frac{1}{2}=\underline{0}$$

$$A_2 \rightsquigarrow 1-\frac{4}{8}\left(-\frac{2}{4}\log\left(\frac{2}{4}\right)-\frac{2}{4}\log\left(\frac{2}{4}\right)\right)-\frac{4}{8}\times\left(-\frac{2}{4}\log\left(\frac{2}{4}\right)-\frac{2}{4}\log\left(\frac{2}{4}\right)\right)=1-\frac{4}{8}\times\left(-\frac{1}{2}\log\left(2^{-1}\right)\times 2\right)$$

$$-\frac{4}{8}\times\left(-\frac{1}{2}\log\left(2^{-1}\right)\times 2\right)=1-\frac{4}{8}-\frac{4}{8}=\underline{0}$$

$$A_3 \rightsquigarrow 1-\frac{6}{8}\left(-\frac{4}{6}\log\left(\frac{4}{6}\right)-\frac{2}{6}\log\left(\frac{2}{6}\right)\right)-\frac{2}{8}\times\left(-\frac{0}{2}\log\left(\frac{0}{2}\right)-\frac{2}{2}\log\left(\frac{2}{2}\right)\right)=1-\frac{6}{8}\times\left(-\frac{2}{3}\log\left(\frac{2}{3}\right)-\frac{1}{3}\log\left(\frac{1}{3}\right)\right)$$

$$-\frac{1}{4}\times 0=1-\frac{3}{4}\times\overbrace{\left(\underbrace{-\frac{2}{3}\log\left(\frac{2}{3}\right)}_{-0.389975}\underbrace{-\frac{1}{3}\log\left(\frac{1}{3}\right)}_{-0.52832083357}\right)}^{0.68872187554}\rightsquigarrow \underline{0.311278124}$$

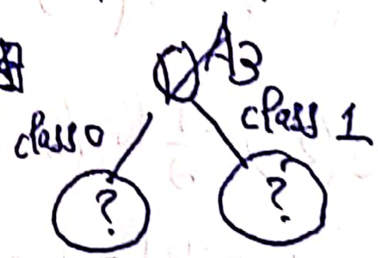
$$A_4 \rightsquigarrow 1-\frac{3}{8}\times\left(-\frac{2}{3}\log\left(\frac{2}{3}\right)-\frac{1}{3}\cdot\log\left(\frac{1}{3}\right)\right)-\frac{5}{8}\left(-\frac{2}{5}\log\left(\frac{2}{5}\right)-\frac{3}{5}\times\log\left(\frac{3}{5}\right)\right)$$

$$=1-\frac{3}{8}\times\left(0.389975+0.52832083357\right)-\frac{5}{8}\left(0.9709505945\right)\approx$$

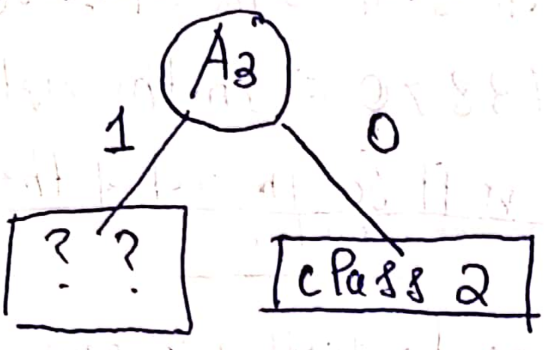
$$\approx 1-0.34436083758-0.60684412153\rightsquigarrow \underline{0.048794941}$$

Since Information Gain is largest for A_3, where $0.311278124 > 0.048794941$

$\Rightarrow \boxed{A_3 \text{ will become root node}}$



depth = 1 According to class table, A_3 will become zero merely for the class 2^{nd}, it's evident fact that the constructed decision tree will be depicted as follows for zero:



After this, we are moving to $A_3 = 1$ subtree and let's evaluate total entropy wrt $A_3 = 1$ subcase:

\# of 1's $\rightsquigarrow$ 6

\# of 1's in class 1 $\rightsquigarrow$ 4

$$-\frac{4}{6}\log_2\left(\frac{4}{6}\right)-\frac{2}{6}\log_2\left(\frac{2}{6}\right)=-\frac{2}{3}\log_2\left(\frac{2}{3}\right)-\frac{1}{3}\log_2\left(\frac{1}{3}\right)\approx$$

$$\approx 0.9182958 3405$$

Now, we are evaluating information gain of A_1, A_2, A_4 with respect to $A_3 = 1$ subtree:

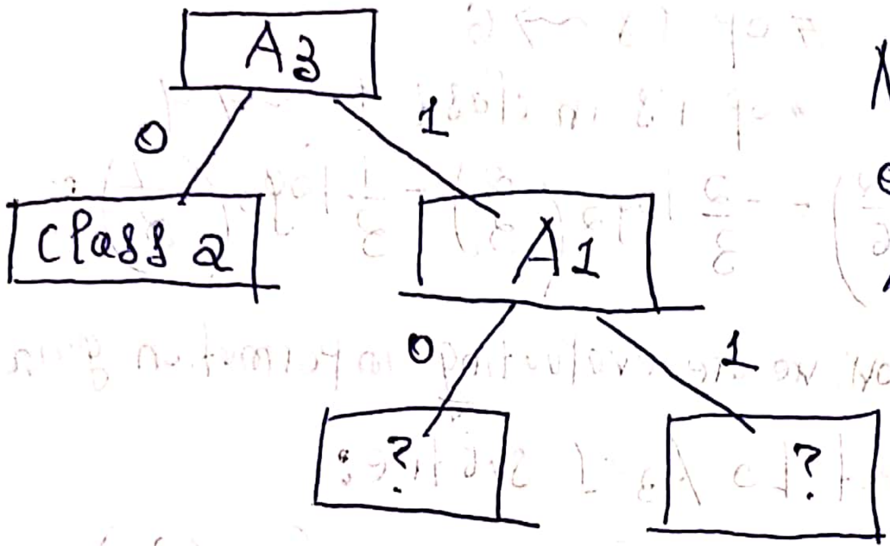
$$A_1 \rightsquigarrow 0.91829583405-\frac{4}{6}\times\left(-\frac{2}{4}\log\left(\frac{2}{4}\right)-\frac{2}{4}\log\left(\frac{2}{4}\right)\right)-\frac{2}{6}\times\left(-\frac{2}{2}\log\left(\frac{2}{2}\right)-0\right)\approx 0.91829583405-\frac{2}{3}\times\left(-\frac{1}{2}\log\left(\frac{1}{2}\right)\times 2\right)-\frac{1}{3}\times 0 = 0.91829583405-\frac{2}{3}\approx \underline{0.25162916738}$$

$$A_2 \rightsquigarrow 0.91829583405-\frac{3}{6}\times\left(-\frac{2}{3}\log\left(\frac{2}{3}\right)-\frac{1}{3}\log\left(\frac{1}{3}\right)\right)-\frac{3}{6}\times\left(-\frac{2}{3}\log\left(\frac{2}{3}\right)-\frac{1}{3}\log\left(\frac{1}{3}\right)\right)=0.91829583405-\left(-\frac{2}{3}\log\left(\frac{2}{3}\right)-\frac{1}{3}\log\left(\frac{1}{3}\right)\right)=0.91829583405-0.91829583405=0$$

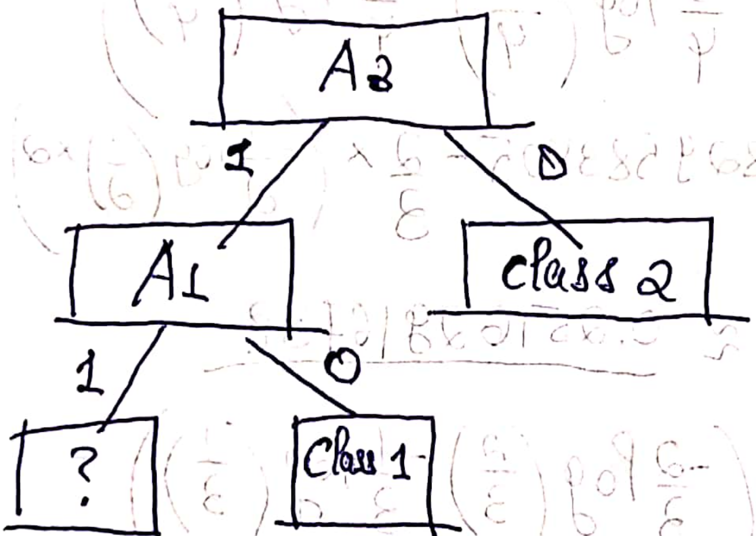
$$A_4 \rightsquigarrow 0.91829583405-\frac{3}{6}\times\left(-\frac{2}{3}\log\left(\frac{2}{3}\right)-\frac{1}{3}\log\left(\frac{1}{3}\right)\right)-\frac{3}{6}\left(-\frac{2}{3}\log\left(\frac{2}{3}\right)-\frac{1}{3}\log\left(\frac{1}{3}\right)\right.$$

$$0.9182958\,3405 - \left(\underbrace{-\frac{2}{3}\log\left(\frac{2}{3}\right)-\frac{1}{3}\log\left(\frac{1}{3}\right)}_{0.9182958\,3405}\right) = \underline{0}$$

Since the largest value for Information Gain turns out to be A_1, where the value is 0.251629 167387 0 ⇒ In the depth 1 and for the subtree $A_3 = \underline{1}$ ⇒ $\underline{A_1 \text{ will be the root of that subtree}}$

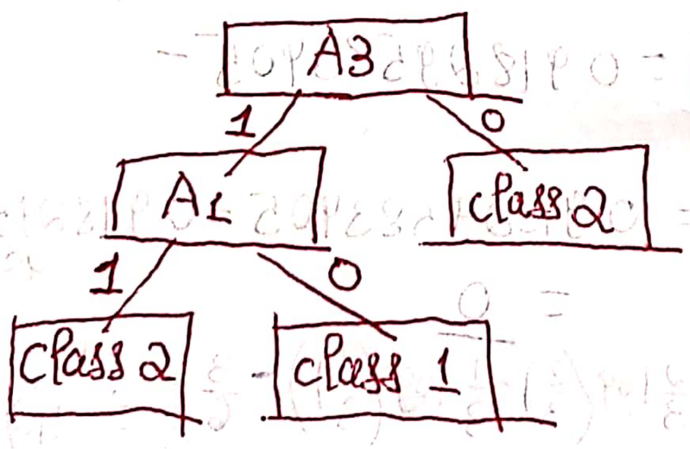


Moving to depth 2, we have to observe the possibilities for $A_1 = \{0,1\}$ ⇒ According to the table, $\underline{A_1 = 0 \text{ will only correspond to class } 1}$



In the final case, let's tackle how $A_1 = A_3 = \underline{1}$ will correspond to which class (1 or 2)

As $A_3 = \underline{1}$ and $A_1 = 0$ is correspondent to class 1, it's evident fact that the last missing is $\underline{\text{class } 2}$ due to the conventions of decision tree



V ▪ ⊕

3) According the given formula, loss function is determined by

$$E(w) = -\sum_{i=1}^{N} y_i \times \left(\log p(y_i = 1 | x_i, w) \right) -$$

$$- \sum_{i=1}^{N} (1 - y_i) \times \log \left(1 - p(y_i = 1 | x_i, w) \right)$$

Plugging in the formula $p(y_j = 1 | x_j, w) = \dfrac{1}{1 + \dfrac{1}{e^{w^T x_j}}}$ in the summation mentioned before, we conclude that

$$-y_j \times \log p(y_j = 1 | x_j, w) - (1 - y_j) \times \log \left(1 - p(y_j = 1 | x_j, w) \right) =$$

$$= -y_j \times \log \left(\frac{e^{w^T x_j}}{e^{w^T x_j} + 1} \right) - (1 - y_j) \times \log \left(\frac{1}{1 + e^{w^T x_j}} \right) =$$

$$= -y_j \times \left[(w^T x_j) - \log (e^{w^T x_j} + 1) \right] + (1 - y_j) \times \log \left(1 + e^{w^T x_j} \right) =$$

$= -y_j \times w^T x_j + \log \left(1 + e^{w^T x_j} \right)$. If we take first derivative wrt w in the final result

$$\Rightarrow \frac{1}{1 + e^{w^T x_j}} \cdot e^{w^T x_j} \cdot x_j^T - y_j \times x_j^T = \frac{1}{1 + e^{-w^T x_j}} x_j^T - y_j \cdot x_j^T$$

$\boxed{\text{First derivative} \rightsquigarrow \dfrac{1}{1 + e^{-w^T x_j}} x_j^T - y_j x_j^T}$ After this, we move for evaluating the 2^{nd} derivative $\rightsquigarrow \dfrac{-1}{(1 + e^{-w^T x_j})^2} \times e^{-w^T x_j} \times (-x_j) x_j^T = \dfrac{e^{-w^T x_j}}{(1 + e^{-w^T x_j})^2} (x_j x_j^T)$

In conclusion, $\frac{\partial^2 E(w)}{\partial w^2} = \sum_{j=1}^{N} \frac{e^{-w^T x_j}}{(1+e^{-w^T x_j})^2} (x_j$

$= \sum_{j=1}^{N} \frac{e^{-w^T x_j}}{(1+e^{-w^T x_j})^2} \|x_j\|_2^2 > 0$, considering the fact that given training set is not completely consisted of zeros

$\frac{\partial^2 E(w)}{\partial w^2}$ is valid and $> 0 \Rightarrow$ From the well-known rule, $\boxed{E(w) \text{ will be convex}}$ $\boxed{V}$ +

1) a) We suppose w has a prior distribution $p(w) = \mathcal{N}(0, \alpha^{-1} I)$ with $\alpha > 0$. Considering the Bayes' formula

$P(w|D) = \frac{P(W,D)}{P(D)} \Rightarrow$ it's easily seen that $P(w|X,y) = \frac{P(y|X,w)\,P(w)}{P(y|X)}$ where it's clear that

$P(w,D) \propto P(y|X,w)\,P(w) \Rightarrow \log(P(w|D)) = \log(P(y|X)) + \log(P(w)) + \text{constant}$

Given that $P(y|X,w) = \prod p(y_i|x_i,w)$ where $y_i = w^T \Phi(x_i) + \varepsilon \Rightarrow y_i \sim \mathcal{N}(w^T \Phi(x_i), \beta_0^{-1})$

Plugging in the formula for normal distribution, we can easily see $P(y|X,w) = (2\pi)^{-n/2} (\beta)^{-n/2} \exp\left(\frac{-1}{2\beta^{-2}} \sum_{j=1}^{N} (y_j - w^T\phi(x_j))^2\right)$

Scanned with CamScanner

4) Taking into account the inequalities $y_i(w^T x_i + b) \geq 1$ for all $i = 1, 2, \dots, 6 \Rightarrow$

The key idea is to minimize the value $\frac{1}{2} w^T w = \frac{1}{2} \|w\|$ meaning we subsequently minimize the value $\|w\|$

$$\begin{cases}
(-1)(-2w_1 - 2w_2 + b) = \underline{2w_1 + 2w_2 - b \geq 1} & (1) \\
(-1)(-w_1 + w_2 + b) = \underline{w_1 - w_2 - b \geq 1} & (2) \\
(-1)(3w_1 - w_2 + b) = \underline{-3w_1 + w_2 - b \geq 1} & (3) \\
(1)(2w_1 + 5w_2 + b) = \underline{2w_1 + 5w_2 + b \geq 1} & (4) \\
(1)(3w_1 + 4w_2 + b) = \underline{3w_1 + 4w_2 + b \geq 1} & (5) \\
(1)(3w_1 + 5w_2 + b) = \underline{3w_1 + 5w_2 + b \geq 1} & (6)
\end{cases}$$

$-3w_1 + w_2 - b + 2w_1 + 5w_2 + b + 3w_1 + 4w_2 + b = -w_1 + 3w_1 + 6w_2 + 4w_2 + b \geq 3$ and specifically, $-3w_1 + w_2 - b + 3w_1 + 4w_2 + b = 5w_2 \geq 1 + 1 = 2$, therefore $\Rightarrow \boxed{w_2 \geq \frac{2}{5}}$ becomes true

$2w_1 + 10w_2 + b \geq 3$, $10w_2 \geq 4 \Rightarrow$ Considering $1 \leq w_1 - w_2 - b$ with $w_2 \geq \frac{2}{5} \Rightarrow 1 \leq w_1 - w_2 - b \leq w_1 - \frac{2}{5} - b$ where $\boxed{w_1 \geq b + \frac{7}{5}}$

from (4), $1 \leq 2w_1 + 5w_2 + b \leq 2w_1 + 5w_2 + w_1 - \frac{7}{5} = 3w_1 + 5w_2 - \frac{7}{5}$

$\frac{12}{5} \leq 3w_1 + 5w_2$ from (5) and (6), it's seen that

$5 \rightsquigarrow 1 \leq 3w_1 + 4w_2 + b \leq 3w_1 + 4w_2 + w_1 - \frac{7}{5} = 4w_1 + 4w_2 - \frac{7}{5}$

$6 \rightsquigarrow 1 \leq 3w_1 + 5w_2 + b \leq 3w_1 + 5w_2 + w_1 - \frac{7}{5} = 4w_1 + 5w_2 - \frac{7}{5}$

$\Rightarrow \frac{12}{5} \leq 4(w_1 + w_2)$ and $\frac{12}{5} \leq 4w_1 + 5w_2$. If we take the first result $\Rightarrow \frac{3}{5} \leq w_1 + w_2$ becomes satisfied

Assume $w_2 = \frac{2}{5} + x_2$, where $x_2 \geq 0$ ⊛
$w_1 = b + \frac{7}{5} + x_1$, where $x_1 \geq 0$ } Considering final result (with the common factor 4 being eliminated in the first expressed result $\frac{12}{5} \leq 4(w_1+w_2)$)

It's natural to deduce $\underbrace{\frac{2}{5} + x_2}_{w_2} + \underbrace{b + \frac{7}{5} + x_1}_{w_1} \geq \frac{3}{5}$ ⊛

$\frac{12}{5} - \frac{3}{5} + x_1 + x_2 + b = \frac{9}{5} + x_1 + x_2 + b \geq \frac{3}{5} \Rightarrow$ $\boxed{\frac{6}{5} + x_1 + x_2 + b \geq 0}$

$\frac{1}{2} w^T w = \frac{1}{2}\left(w_1^2 + w_2^2\right) = \frac{1}{2}\left(\left(b + \frac{7}{5} + x_1\right)^2 + \left(\frac{2}{5} + x_2\right)^2\right) =$

$= \frac{1}{2}\left(\left(b + \frac{7}{5} + x_1\right)^2 + x_2^2 + \frac{4x_2}{5} + \frac{4}{25}\right) =$

$= \frac{1}{2}\left[\left(b + \frac{6}{5} + x_1 + \frac{1}{5}\right)^2 + x_2^2 + \frac{4x_2}{5} + \frac{4}{25}\right] =$

$= \frac{1}{2}\left[\left(b + \frac{6}{5} + x_1\right)^2 + \frac{1}{25} + \frac{2}{5}\left(b + \frac{6}{5} + x_1\right) + x_2^2 + \frac{4x_2}{5} + \frac{4}{25}\right]$

$= \frac{1}{2}\left[\underbrace{\left(b + \frac{6}{5} + x_1\right)^2}_{\geq 0} + \frac{1}{5} + \underbrace{x_2^2}_{\geq 0} + \frac{2}{5}\underbrace{\left(x_1 + \frac{6}{5} + 2x_2 + b\right)}_{\geq 0}\right]$

where the last inequality holds since $\left(\frac{6}{5} + x_1 + x_2 + b\right) + x_2 \geq 0 + 0 = 0$ from red-starred inequalities on this page $\Rightarrow$
$\|w\|^2 = \left(b + \frac{6}{5} + x_1\right)^2 + x_2^2 + \frac{2}{5}\left(b + \frac{6}{5} + x_1 + x_2 + x_2\right) + \frac{1}{5} \geq \frac{1}{5}$

Where inequality becomes $=0 \Leftrightarrow b+\frac{6}{5}+x_1=0$ and moreover, $\boxed{\|w\|^2 \geq \frac{1}{5}}$ $x_2=0$ $\frac{1}{2}\|w\|^2$ becomes the smallest value $\Leftrightarrow \boxed{\begin{array}{c} b+\frac{6}{5}+x_1=0 \\ x_2=0 \end{array}}$

$x_2=0 \Rightarrow w_2=\frac{2}{5}$

$w_1=b+\frac{6}{5}+x_1+\frac{1}{5}=\frac{1}{5}$

$\boxed{\begin{array}{c} w_1=1/5 \\ w_2=2/5 \end{array}}$ $\rightsquigarrow \|w\|^2$ will receive its minimum for those w_1, w_2 values

(1) $\rightsquigarrow 2w_1+2w_2-b=\frac{2}{5}+\frac{4}{5}-b=\boxed{\frac{6}{5}-b\geq 1} \Rightarrow \boxed{\frac{1}{5}\geq b}$

(2) $\rightsquigarrow w_1-w_2-b=\frac{1}{5}-\frac{2}{5}-b=-\frac{1}{5}-b\geq 1 \Rightarrow \boxed{b\leq -\frac{6}{5}}$

(3) $\rightsquigarrow -3w_1+w_2-b=-\frac{3}{5}+\frac{2}{5}-b=-\frac{1}{5}-b\geq 1 \Rightarrow \boxed{b\leq -\frac{6}{5}}$

(4) $\rightsquigarrow 2w_1+5w_2+b=\frac{2}{5}+2+b=\frac{12}{5}+b\geq 1 \Rightarrow \boxed{b\leq -\frac{7}{5}}$

(5) $\rightsquigarrow 3w_1+4w_2+b=\frac{3}{5}+\frac{8}{5}+b=\frac{11}{5}+b\geq 1 \Rightarrow \boxed{b\geq -\frac{6}{5}}$

(6) $\rightsquigarrow 3w_1+5w_2+b=\frac{3}{5}+2+b=\frac{13}{5}+b\geq 1 \Rightarrow \boxed{b\leq -\frac{8}{5}}$

It's an obvious fact that the resultant values of support vectors originating from first, fourth, and sixth will not correctly satisfy $y_i(w^T x_i+b)\geq 1$ (more specifically, they are satisfying $y_i(w^T x_i+b)>1$ which should not) In fact, if any of the resultant values from the remaining options (second, fifth, sixth) is true, then all of these aforementioned inequalities will remain satisfactory $\Rightarrow \boxed{b=-\frac{6}{5}}$

Scanned with CamScanner

$w_1 = 1/5$
$w_2 = 2/5$
$b = -6/5$

Equation for the decision line could be derived in the following manner: $\frac{1}{5}x + \frac{2}{5}y - \frac{6}{5} = 0 \Rightarrow$

$$x + 2y - 6 = 0 \Leftrightarrow \boxed{y = \frac{6-x}{2} = \frac{-1}{2}x + 3}$$

From the given dataset, it's easily derived that
and previous implications

$[-1, 1]^T$
$[3, -1]^T$
$[3, 4]^T$
support vectors

The first, fourth, and sixth will not be support vectors, considering the inequality $y_i(w^T x_i + b) > 1$ (strictly > 1)

Equality $\Rightarrow$ support vectors

Note: Just comparing ② and ⑤ will yield $\boxed{b = \frac{-6}{5}}$

Equality holds

Scanned with CamScanner