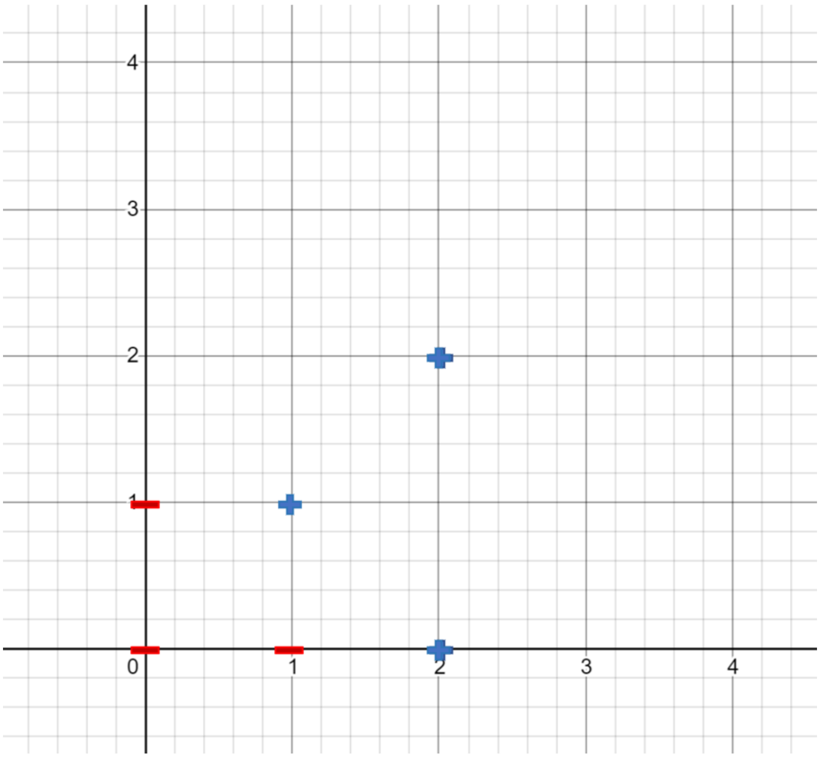
Data Mining – Assignment 3

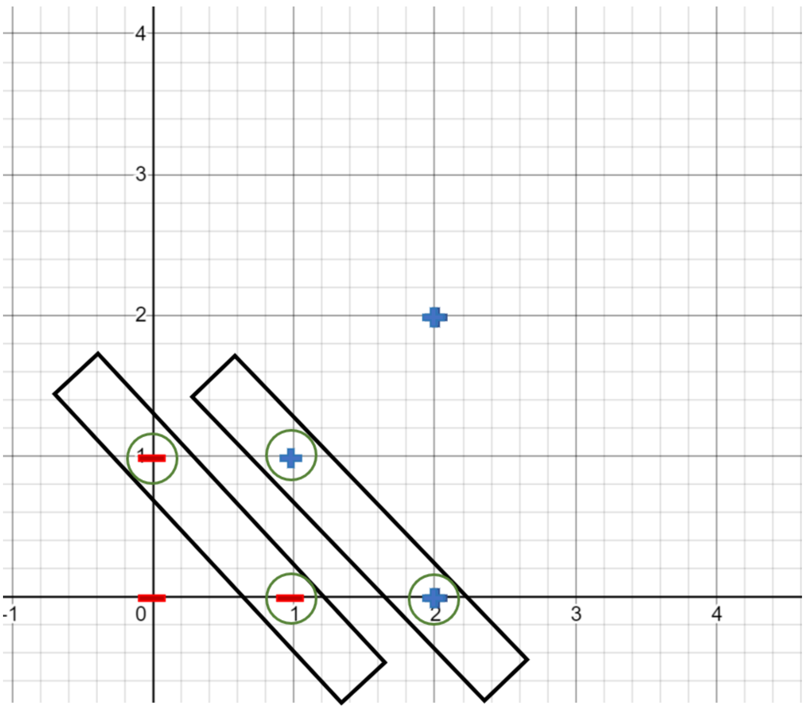
薛劭杰　　1930026143

Q1.

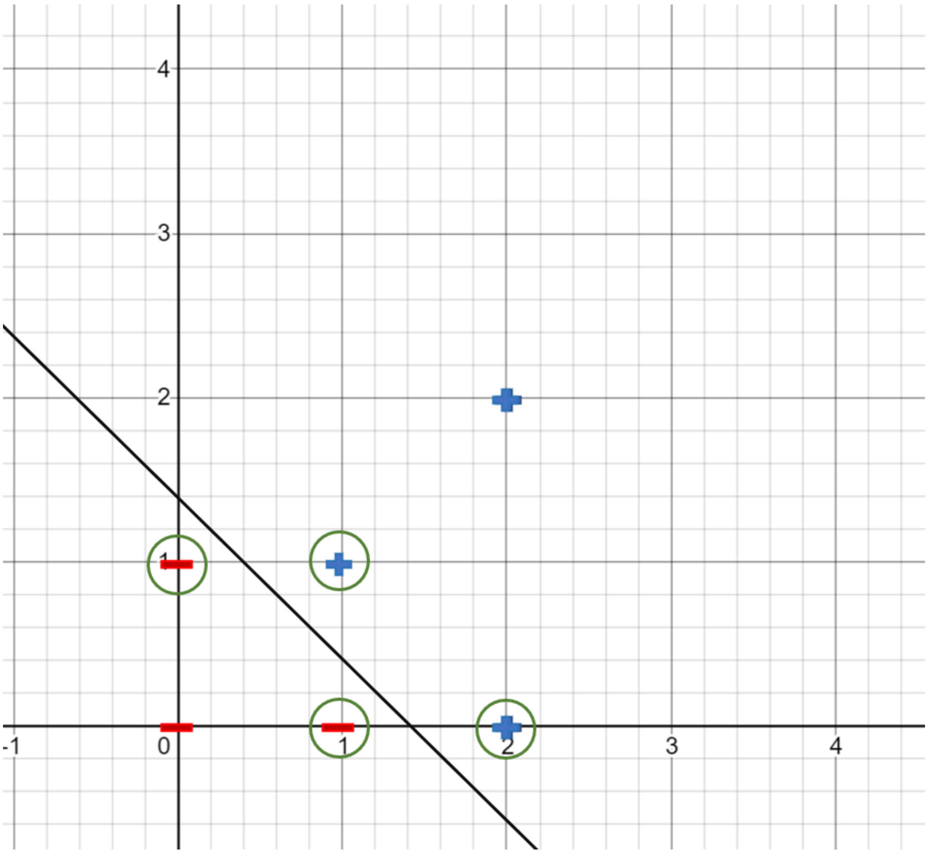
(a)



(2)



(3)



(d)

In the vectors, we select the point1: (1, 0), point3: (1, 1) and point5: (2, 0), and we can the three functions: $\begin{cases} 1 * W_1 + 0 * W_2 + b = -1 \\ 1 * W_1 + 1 * W_2 + b = 1 \\ 2 * W_1 + 0 * W_2 + b = 1 \end{cases}$, there are three unknowns and three the equations, so we can get all results: $W_1 = 2,\ W_2 = 2,\ b = -3$. So, the optional separating hyperplane of this system is: $2x_1 + 2x_2 - 3 = 0$, the bias is -3.

The margin for each point:

- For $P_1(1,1)$: $d_1 = \frac{|2x_1+2x_2-3|}{\sqrt{2^2+2^2}} = \frac{\sqrt{2}}{4}$.
- For $P_2(2,2)$: $d_2 = \frac{|2x_1+2x_2-3|}{\sqrt{2^2+2^2}} = \frac{5\sqrt{2}}{2}$.
- For $P_3(2,0)$: $d_3 = \frac{|2x_1+2x_2-3|}{\sqrt{2^2+2^2}} = \frac{\sqrt{2}}{4}$.
- For $P_4(0,0)$: $d_4 = \frac{|2x_1+2x_2-3|}{\sqrt{2^2+2^2}} = \frac{3\sqrt{2}}{4}$.
- For $P_5(1,0)$: $d_5 = \frac{|2x_1+2x_2-3|}{\sqrt{2^2+2^2}} = \frac{\sqrt{2}}{4}$.
- For $P_6(0,1)$: $d_6 = \frac{|2x_1+2x_2-3|}{\sqrt{2^2+2^2}} = \frac{\sqrt{2}}{4}$.

For the minimum margin $\gamma = 2 * \min_{0 \le i \le 6}(d_i) = 2 * \frac{\sqrt{2}}{4} = \frac{\sqrt{2}}{2}$.

Q2.

(a)

Jaccard coefficient: $JC = \frac{f_{11}}{f_{01}+f_{10}+f_{11}}$. We set $'yes'$ to 1 and $'No'$ to 0

Adopt Jaccard similarity to measure the closeness between samples:

$JC_1 = \frac{2}{3}$ $\quad JC_2 = \frac{0}{2} = 0$ $\quad JC_3 = \frac{1}{3}$ $\quad JC_4 = \frac{2}{4} = \frac{1}{2}$

$JC_5 = \frac{2}{3}$ $\quad JC_6 = \frac{1}{4}$ $\quad JC_7 = \frac{2}{4} = \frac{1}{2}$ $\quad JC_8 = \frac{2}{3}$

With $k\ value$ set to 5:

The largest five Jaccard similarity is $JC_1, JC_5, JC_8, JC_4, JC_7$.

Then for their classes:

$1 - Mammals$

$4 - Mammals$

$5 - Non\text{-}Mammals$

$7 - Mammals$

$8 - Mammals$

There are 4 Mammals and 1 Non-mammals, so classify the test to Mammals.

(b)

We have to compute $Gain(D, A)$ for each attribute A. First of all, calculate the entropy for the category.

$- \ I(D) = -\frac{4}{8} * \log\left(\frac{4}{8}\right) - \frac{4}{8} * \log\left(\frac{4}{8}\right) = 1$

Then focus on $A = Give\ Birth$

$- \ I(D_{yes}) = -\frac{4}{4} * \log\left(\frac{4}{4}\right) - \frac{0}{4} * \log\left(\frac{0}{4}\right) = 0$

$- \ I(D_{No}) = -\frac{4}{4} * \log\left(\frac{4}{4}\right) - \frac{0}{4} * \log\left(\frac{0}{4}\right) = 0$

$- \ Weight\ average = \frac{1}{2} * I(D_{yes}) + \frac{1}{2} * (D_{No}) = 0$

$- \ Gain(D_{Give\ Birth}) = I(D) - \ Weight\ average = 1$

Then focus on $A = Can\ Fly$

$- \ I(D_{yes}) = -\frac{2}{3} * \log\left(\frac{2}{3}\right) - \frac{1}{3} * \log\left(\frac{1}{3}\right) = 0.9183$

$- \ I(D_{No}) = -\frac{2}{5} * \log\left(\frac{2}{5}\right) - \frac{3}{5} * \log\left(\frac{3}{5}\right) = 0.9710$

$- \ Weight\ average = \frac{3}{8} * I(D_{yes}) + \frac{5}{8} * I(D_{No}) = 0.9512$

$- \ Gain(D_{Can\ Fly}) = I(D) - \ Weight\ average = 0.0488$

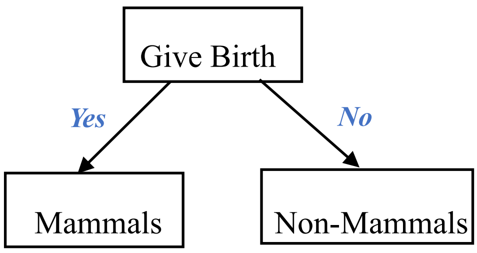
Then focus on $A = \ Live\ in\ Water$

- $I(D_{yes}) = -\frac{2}{3} * \log\left(\frac{2}{3}\right) - \frac{1}{3} * \log\left(\frac{1}{3}\right) = 0.9183$
- $I(D_{No}) = -\frac{2}{5} * \log\left(\frac{2}{5}\right) - \frac{3}{5} * \log\left(\frac{3}{5}\right) = 0.9710$
- $Weight\ average = \frac{3}{8} * I(D_{yes}) + \frac{5}{8} * I(D_{No}) = 0.9512$
- $Gain(D_{Live\ in\ Water}) = I(D) - \ Weight\ average = 0.0488$

Then focus on $A = \ Have\ Legs$
- $I(D_{yes}) = -\frac{2}{5} * \log\left(\frac{2}{5}\right) - \frac{3}{5} * \log\left(\frac{3}{5}\right) = 0.9710$
- $I(D_{No}) = -\frac{2}{3} * \log\left(\frac{2}{3}\right) - \frac{1}{3} * \log\left(\frac{1}{3}\right) = 0.9183$
- $Weight\ average = \frac{5}{8} * I(D_{yes}) + \frac{3}{8} * I(D_{No}) = 0.9512$
- $Gain(D_{Have\ Legs}) = I(D) - \ Weight\ average = 0.0488$

$Gain(D_{Give\ Birth}) > Gain(D_{Have\ Legs}) \geq Gain(D_{Can\ Fly}) > \ Gain(D_{Live\ in\ Water})$

So, split using the attribute $Give\ Birth$.



It can get the result and it not need to split again.
The Give birth of the test one is $'Yes'$, so classify the test to Mammals.

Set the set is split on an attribution A into subset yes and no
Gini index:

$gini(D) = 1 - \sum_i^m {p_i}^2 = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$

Focus on $Give\ Birth$:

$gini(D_{Yes}) = 1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2 = 0$

$gini(D_{No}) = 1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2 = 0$

$gini_{Give\ Birth}(D) = \frac{|D_{yes}|}{|D|} * gini(D_{yes}) + \frac{|D_{No}|}{|D|} * gini(D_{No}) = 0 + 0 = 0$

$\Delta gini(Give\ Birth) = gini(D) - gini_{Give\ Birth}(D) = \frac{1}{2} - 0 = \frac{1}{2}$

Focus on $Can\ Fly$:

$gini(D_{Yes}) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 \approx 0.4444$

$gini(D_{No}) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$

$gini_{Can\ Fly}(D) = \frac{|D_{yes}|}{|D|} * gini(D_{yes}) + \frac{|D_{No}|}{|D|} * gini(D_{No}) = \frac{3}{8} * 0.4444 + \frac{5}{8} * 0.48 =$

0.46665

$\Delta gini(Can\ Fly) = gini(D) - gini_{Can\ Fly}(D) = \frac{1}{2} - 0.46665 = 0.03335$

Focus on $Can\ Fly$:

$gini(D_{Yes}) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 \approx 0.4444$

$gini(D_{No}) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$

$gini_{Can\ Fly}(D) = \frac{|D_{yes}|}{|D|} * gini(D_{yes}) + \frac{|D_{No}|}{|D|} * gini(D_{No}) = \frac{3}{8} * 0.4444 + \frac{5}{8} * 0.48 =$

0.46665

$\Delta gini(Can\ Fly) = gini(D) - gini_{Can\ Fly}(D) = \frac{1}{2} - 0.46665 = 0.03335$

Focus on $Live\ in\ Water$:

$gini(D_{Yes}) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 \approx 0.4444$

$gini(D_{No}) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$

$gini_{Live\ in\ Water}(D) = \frac{|D_{yes}|}{|D|} * gini(D_{yes}) + \frac{|D_{No}|}{|D|} * gini(D_{No}) = \frac{3}{8} * 0.4444 + \frac{5}{8} *$

$0.48 = 0.46665$

$\Delta gini(Live\ in\ Water) = gini(D) - gini_{Live\ in\ Water}(D) = \frac{1}{2} - 0.46665 =$

0.03335

Focus on $Have\ Legs$:

$gini(D_{Yes}) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$

$gini(D_{No}) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 \approx 0.4444$

$gini_{Have\ Legs}(D) = \frac{|D_{yes}|}{|D|} * gini(D_{yes}) + \frac{|D_{No}|}{|D|} * gini(D_{No}) = \frac{5}{8} * 0.48 + \frac{3}{8} *$

$0.4444 = 0.46665$

$\Delta gini(Have\ Legs) = gini(D) - gini_{Have\ Legs}(D) = \frac{1}{2} - 0.46665 = 0.03335$

(c)

Firstly, we compute the prior probability for each class:

Set $C_1 = 'Mammals'$, $C_2 =' Non\text{-}Mammals'$

$- \ P(C_1) = \frac{4}{8} = 0.5,\ P(C_2) = \frac{4}{8} = 0.5$

To derive $P(x|C_i)$ for $i = 1,2$, we need to compute the following:

$- \ P(Give\ Birth = Yes|C_1) = \frac{4+1}{4+2} = \frac{5}{6},\ P(Give\ Birth = Yes|C_2) = \frac{0+1}{4+2} = \frac{1}{6}$

$- \ P(Can\ Fly = No|C_1) = \frac{2+1}{4+2} = \frac{1}{2},\ P(Can\ Fly = No|C_2) = \frac{3+1}{4+2} = \frac{2}{3}$

$- \ P(Live\ in\ Water = Yes|C_1) = \frac{1+1}{4+2} = \frac{1}{3},\ P(Live\ in\ Water = Yes|C_2) = \frac{2+1}{4+2} = \frac{1}{2}$

$- \ P(Have\ Legs = yes|C_1) = \frac{3+1}{4+2} = \frac{2}{3},\ P(Have\ Legs = Yes|C_2) = \frac{2+1}{4+2} = \frac{1}{2}$

Given the previous probabilities, we obtain:

$- \ P(x|C_1) = P(GiveBirth = Yes|C_1) * P(CanFly = No|C_1) *$

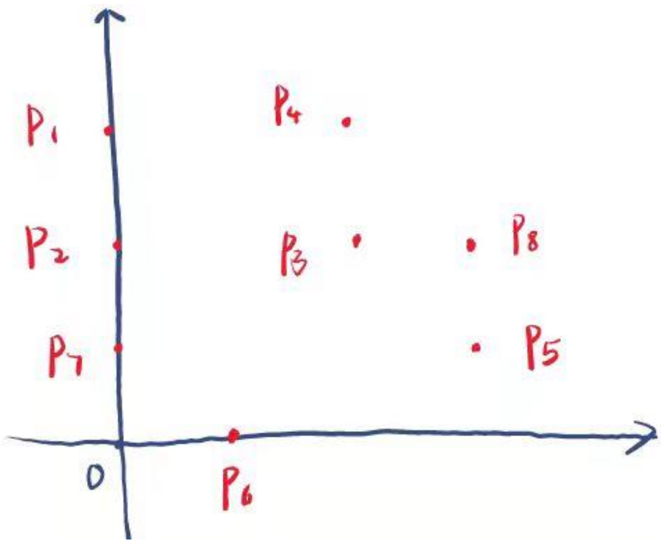
$P(Livein\ Water = Yes|\ C_1) * P(Have\ Legs = yes|C_1) = \frac{5}{6} * \frac{1}{2} * \frac{1}{3} * \frac{2}{3} * \frac{1}{2} \approx 0.0462965$

$- \ P(x|C_2) = P(GiveBirth = Yes|C_2) * P(CanFly = No|C_2) *$

$P(Livein\ Water = Yes|\ C_2) * P(HaveLegs = yes|C_2) = \frac{1}{6} * \frac{2}{3} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} = 0.013889$

$P(x|C_1) > P(x|C_2)$, so classify the test to (C_1) Mammals.

Q3.



The eight points in the coordinate system like above picture:

Firstly, we can calculate the Euclidean distance for each point:

For point P_1: $dis(1{,}2) = 1$, $dis(1{,}3) = \sqrt{5}$, $dis(1{,}4) = 3$, $dis(1{,}5) = \sqrt{13}$, $dis(1{,}6) = \sqrt{10}$, $dis(1{,}7) = 2$, $dis(1{,}8) = \sqrt{10}$.

For point P_2: $dis(2{,}3) = 2$, $dis(2{,}4) = \sqrt{5}$, $dis(2{,}5) = \sqrt{10}$, $dis(2{,}6) = \sqrt{5}$, $dis(2{,}7) = 1$, $dis(2{,}8) = 3$.

For point P_3: $dis(3{,}4) = 1$, $dis(3,5) = \sqrt{2}$, $dis(3,6) = \sqrt{5}$, $dis(3{,}7) = \sqrt{5}$, $dis(3{,}8) = 1$.

For point P_4: $dis(4{,}5) = 5$, $dis(4,6) = \sqrt{10}$, $dis(4,7) = 2\sqrt{2}$, $dis(4,8) = \sqrt{2}$.

For point P_5: $dis(5,6) = \sqrt{5}$, $dis(5,7) = \sqrt{3}$, $dis(5,8) = 1$.

For point P_6: $dis(6,7) = \sqrt{2}$, $dis(6,8) = 2\sqrt{2}$.

For point P_7: $dis(7,8) = \sqrt{10}$.

$\epsilon = 1$ and $minPt = 3$. For each point, select the other points which distance is smaller ϵ

$- N(P_1) = \{P_1, P_2\}$

$- N(P_2) = \{P_1, P_2, P_7\}$

$- N(P_3) = \{P_3, P_4, P_8\}$

$- N(P_4) = \{P_3, P_4\}$

$- N(P_5) = \{P_5, P_8\}$

$- N(P_6) = \{P_6\}$

$- N(P_7) = \{P_2, P_7\}$

$- N(P_8) = \{P_3, P_5, P_8\}$

As $MinPt = 3$

$-$ P_2, P_3, P_8 are core points

$-$ P_1, P_4, P_5, P_7 are border points

$-$ P_6 is a noise point

Now run DBSCAN algorithm:

For each visited points, support start with P_2 (P_2 is a core point and a cluster is form C_1), then mark P_1 as the visited and retrieve ϵ-$neighborhood$, $N(P_2) = \{P_1, P_2, P_7\}$

$-$ Next add P_1 (visited), as $N(P_1) = \{P_1, P_2\}$, no append.

- Next add P_7 (visited), as $N(P_7) = \{P_2, P_7\}$, no append.
- Finish for $C_1 = \{P_1, P_2, P_7\}$

Now $\{P_1, P_2, P_7\}$ are visited

For each unvisited point, suppose continue with P_3 (P_2 is a core point and a cluster is form C_2).

mark P_3 as the visited and retrieve $\epsilon\text{-}neighborhood$, $N(P_3) = \{P_3, P_4, P_8\}$

- Next add P_4 (visited), as $N(P_4) = \{P_3, P_4\}$, no append.
- Next add P_8 (visited), as $N(P_8) = \{P_3, P_5, P_8\}$, append $N(P_5)$ to $N(P_3)$.
- Next add P_5 (visited), as $N(P_5) = \{P_5, P_8\}$, no append.
- Finish for $C_2 = \{P_3, P_4, P_5, P_8\}$

Now $\{P_1, P_2, P_3, P_4, P_5, P_7, P_8\}$ are visited.

For each unvisited point, suppose continue with P_6.

mark P_1 as the visited and retrieve $\epsilon\text{-}neighborhood$, $N(P_6) = \{P_6\}$.

- P_6 is not a core point, mark the P_6 as noise point.

All points are visited:

- $C_1 = \{P_1, P_2, P_7\}$
- $C_2 = \{P_3, P_4, P_5, P_8\}$
- $Noise = \{P_6\}$