Introduction to Machine Learning

Assignment 3

Important Instructions:

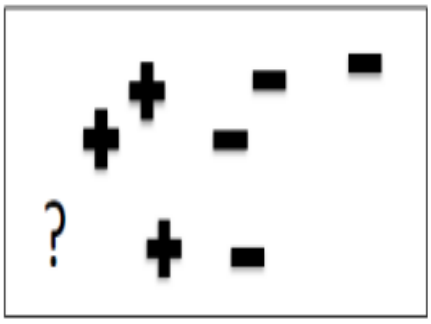
- This is an individual assignment. Each student must work alone, without help from any other person.
- Plagiarism is NOT allowed. Copying material from the web or from documents and submitting it for this assignment is plagiarism. The copied assignment will be marked ZERO.
- Explain solution to the questions properly in your report. (Be concise and precise in your answers)
- A bad presentation usually earns bad credits

1 K-Nearest Neighbor (KNN)

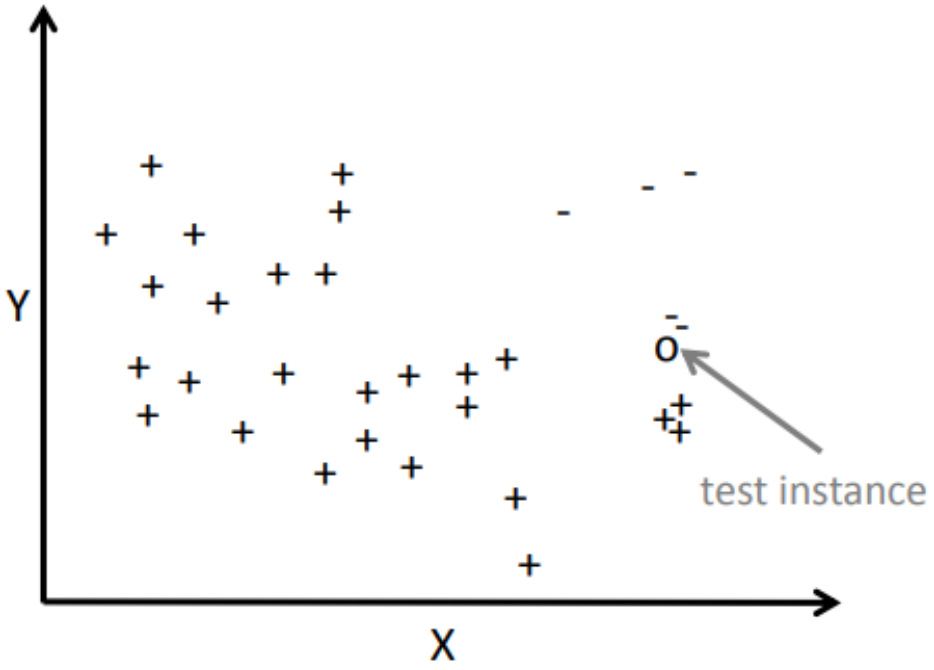
In this part of the exercise, you will be solving questions related to KNN.

1.1 For what minimal value of k will the query point "?" be negative? You also must find label of query point for k = 1 to 7.

Following is the given labeled dataset.



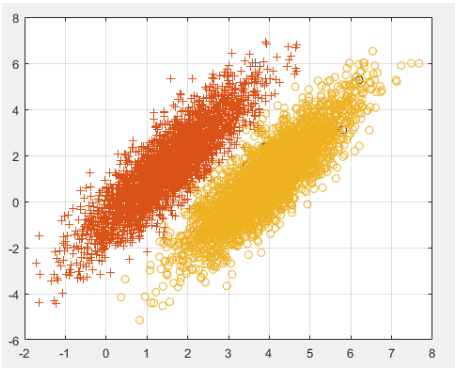
1.2 What value of k you would recommend for the given below dataset? Also explain your approach for selecting the value.



1.3 Implementation:

Your task is to implement KNN algorithm in MATLAB. MATLAB file "Data.m" is given in this assignment. It will generate data which you will use to test implementation of your KNN.

Following figure shows the data which is provided.



2 Decision Tree:

The following table contains training examples that help predict whether a patient is likely to have a heart attack. Use information theory to construct a minimal decision tree that predicts whether a patient is likely to have a heart attack. Show each step for computation.

Patient ID	Chest Pain	Male	Smokes	Exercise	Heart Attack
1	Yes	Yes	No	Yes	Yes
2	Yes	Yes	Yes	No	Yes
3	No	No	Yes	No	Yes
4	No	Yes	No	Yes	No
5	Yes	No	Yes	Yes	No
6	No	Yes	Yes	Yes	No

Good Luck

Rafay Aamir (bscs 19047)

Q1

(1)

For 1-NN ? will be assigned + label

2-NN ? will be assigned + label

3-NN " " " " − "

4-NN " " " " −

5 NN +

6 NN −

7 NN −

so, K = 3 will label the query point "?" as "−"

(2) Here if K = 5, two closer (−)ves can be considered as the out liers. there are three methods that I think, can be used to to find the best (K)

$K = \log_2(32) = 5$

$K = \sqrt[2]{32} = 5.6 = 5 \text{ or } 6$

K = can be found by using error and accuracy plot / graph.

Q8-2

ID	Chest Pain CP	Male M	Smokes S	Exercise E	Heart Attack HA
1	Y	Y	N	Y	Y
2	Y	Y	Y	N	Y
3	N	N	Y	N	Y
4	N	Y	N	Y	N
5	Y	N	Y	Y	N
6	N	Y	Y	Y	N

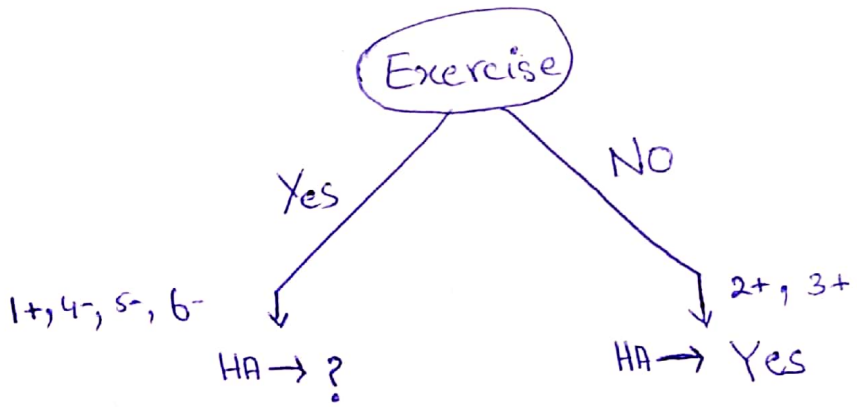
$$\text{Entropy}[HA] = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} \Rightarrow 1 \quad \text{——— (1)}$$

$$IG[CP] = \left(\frac{3}{6}\left[-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}\right] + \frac{3}{6}\left[-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}\right]\right) + 1 \Rightarrow 0.333 \quad \text{——— (2)}$$

$$IG[M] = 1 - \left[\frac{4}{6}\left[-\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}\right] + \frac{2}{6}\left[-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right]\right] \Rightarrow 0 \quad \text{——— (3)}$$

$$IG[S] = 1 - \left[\frac{4}{6}\left[\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}\right] + \frac{2}{6}\left[-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right]\right] \Rightarrow 0 \quad \text{——— (4)}$$

$$IG(E) = 1 - \left[\frac{4}{6}\left[-\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4}\right] + \frac{2}{6}\left[\frac{2}{2}\log\frac{2}{2} - \frac{0}{2}\log\frac{0}{2}\right]\right] \Rightarrow 0.45915 \quad \text{——— (5)}$$



	CP	M	S	HA
ID	chest pain	Male	Smokes	Heart Attack
1	Y	Y	N	Y
4	N	Y	N	N
5	Y	N	Y	N
6	N	Y	Y	N

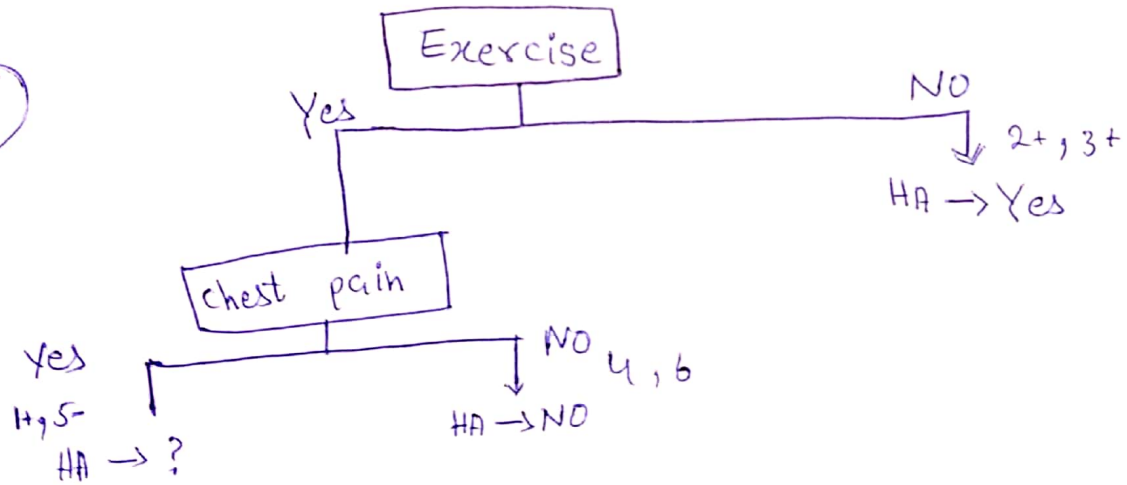
Entropy (HA) = 0.8113 — (6)

$$IG(CP) = \frac{2}{4}(0.8113) - \left[\frac{2}{4}\left[-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right] + \frac{2}{4}\left[\frac{2}{2}\log\frac{2}{2} + 0\right]\right] \Rightarrow 0.3113 \quad (7)$$

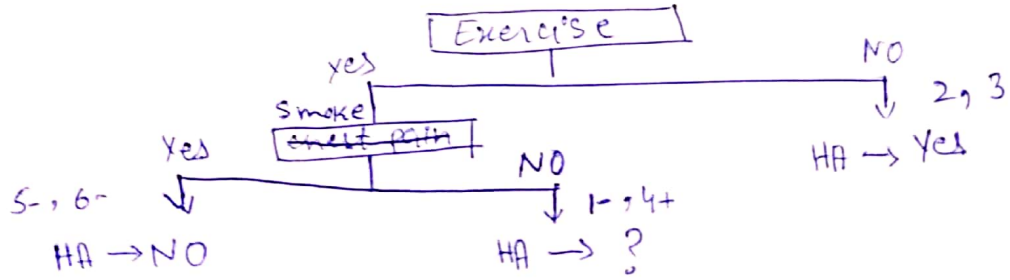
$$IG(M) = 0.8113 - \left[\frac{3}{4}\left[-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}\right] + \frac{1}{4}\left[-1\log 1 - 0\right]\right] \Rightarrow 0.12258 \quad (8)$$

$$IG(S) = 0.8113 - \left[\frac{2}{4}\left[1\log 1\right] + \frac{2}{4}\left[\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right]\right] \Rightarrow 0.3113 \quad (9)$$

T1



T2



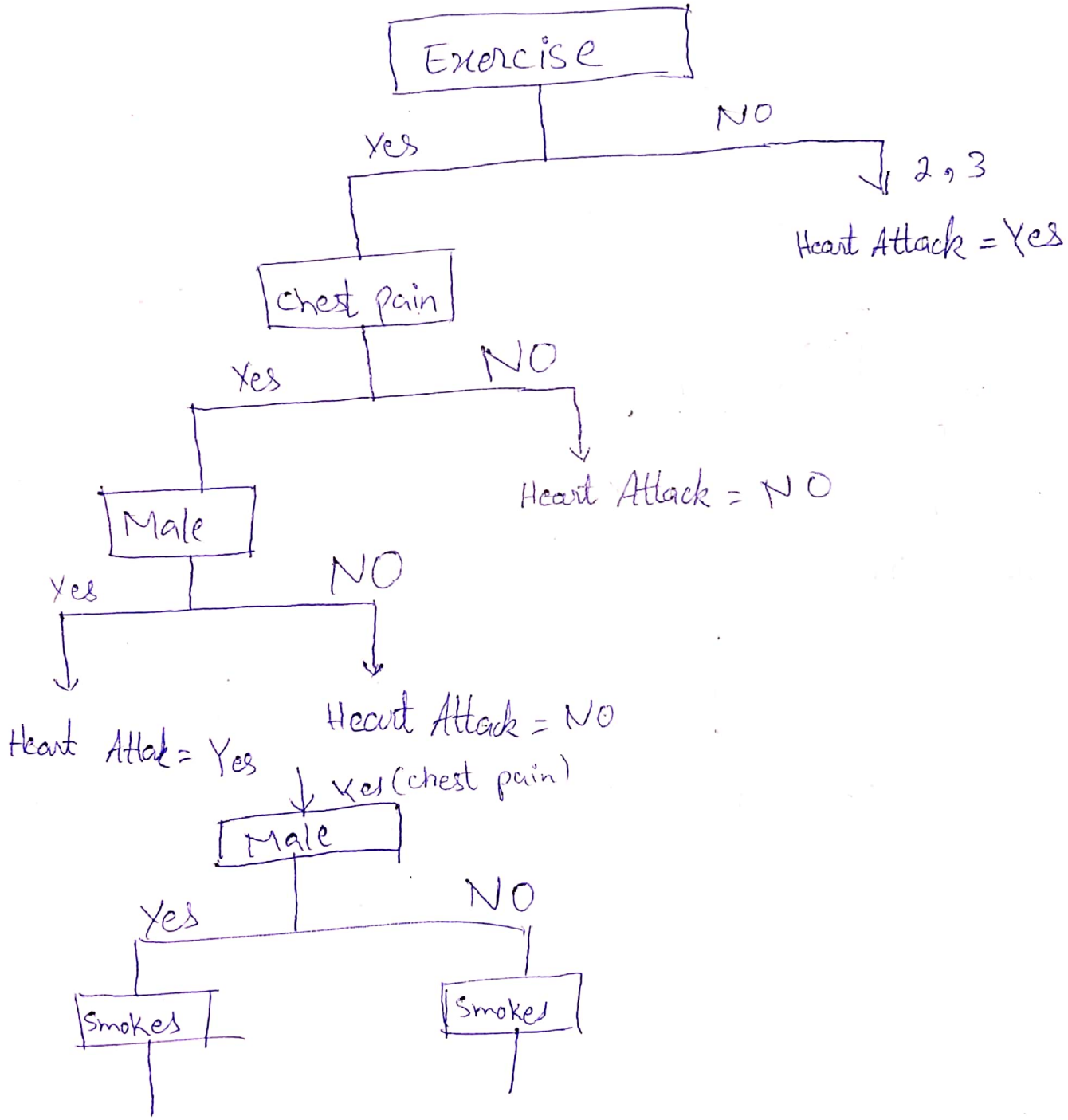
For Tree 1 (T1)

M	S	HA
Y	N	Y
N	Y	N

Entropy(HA) = 1

$IG(M) = 1 - \left[\frac{1}{2}\left(-\frac{1}{1}\log\frac{1}{2} - 0\right)\right. = 1$

$IG(S) = 1 - \left[\frac{1}{2}\right.$



For T2

	CP	M	HA
	~~Y~~	~~Y~~	
1	Y	Y	Y
4	N	Y	N

entropy(HA) = 1

IG(CP) = 1 − [½{log₂(1) − 0} + ½{ = 1

IG(M) = 1 − [1{−½ log ½} ½ log ½] = 0