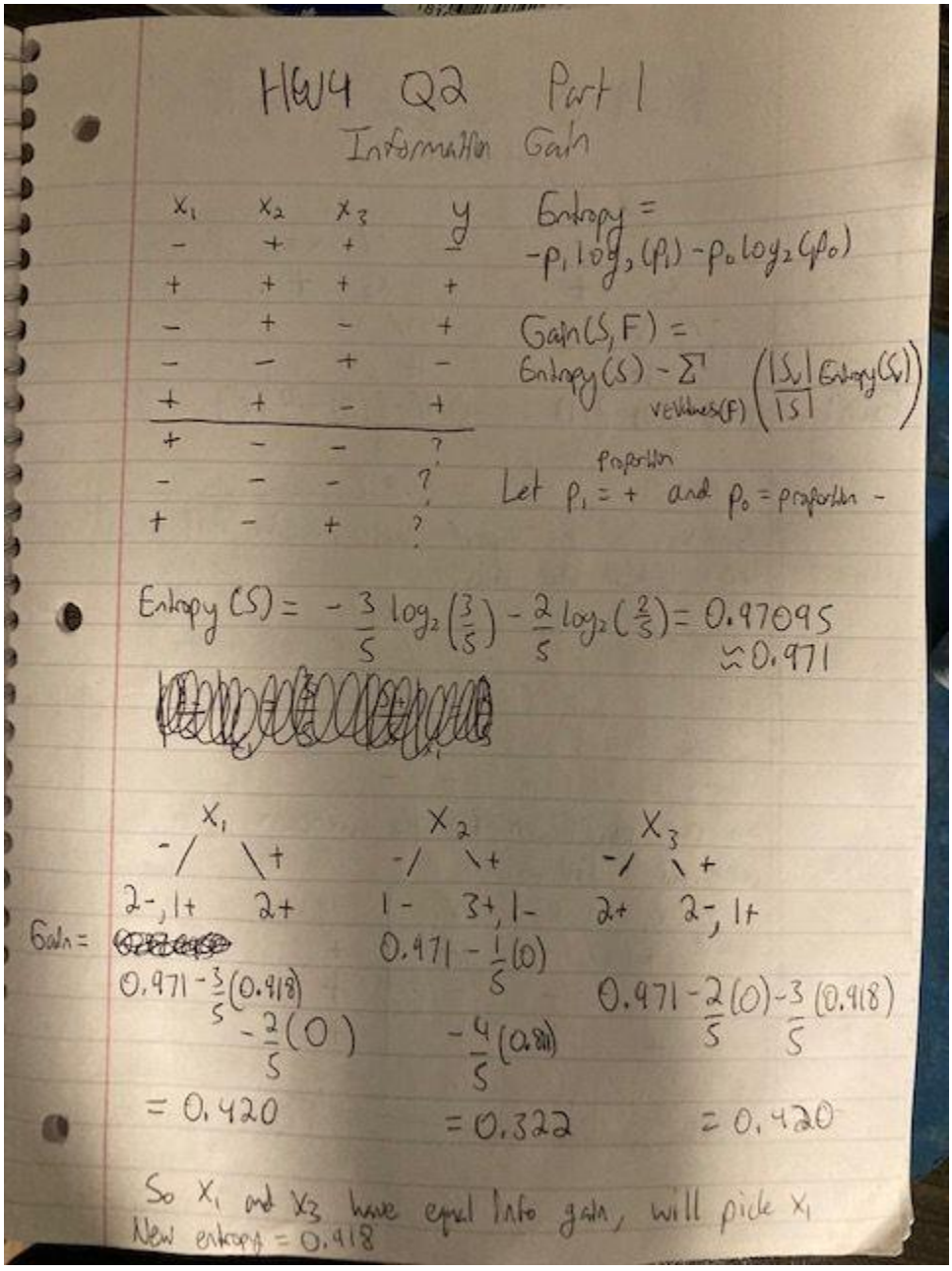
Question 2:



HW4 Q2 Part 1

Information Gain

X_1	X_2	X_3	y
−	+	+	−
+	+	+	+
−	+	−	+
−	−	+	−
+	+	−	+
+	−	−	?
−	−	−	?
+	−	+	?

Entropy = $-p_1 \log_2(p_1) - p_0 \log_2(p_0)$

$Gain(S, F) = Entropy(S) - \sum_{v \in Values(F)} \left(\frac{|S_v|}{|S|} Entropy(S_v) \right)$

Let p_1 = proportion + and p_0 = proportion −

$Entropy(S) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) = 0.97095 \approx 0.971$

X_1: − / + → 2−, 1+ and 2+

Gain = $0.971 - \frac{3}{5}(0.918) - \frac{2}{5}(0) = 0.420$

X_2: − / + → 1− and 3+, 1−

$0.971 - \frac{1}{5}(0) - \frac{4}{5}(0.811) = 0.322$

X_3: − / + → 2+ and 2−, 1+

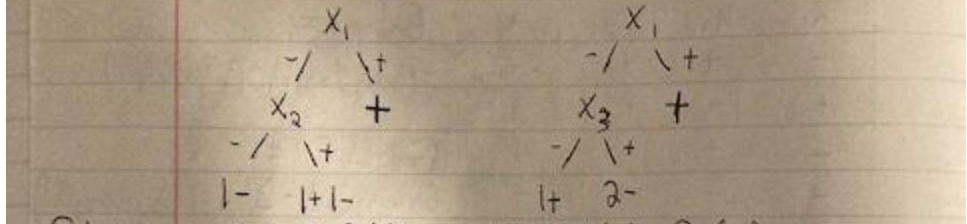
$0.971 - \frac{2}{5}(0) - \frac{3}{5}(0.918) = 0.420$

So X_1 and X_3 have equal info gain, will pick X_1

New entropy = 0.918

HW4 Q2 continued (Part 1)

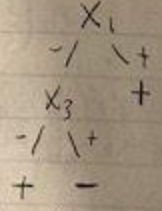
Information Gain



Gain: $0.918 - \frac{1}{3}(0) - \frac{2}{3}(1)$ $\quad$ $0.918 - \frac{1}{3}(0) - \frac{2}{3}(0)$

$= 0.251$ $\quad$ $= 0.918$

So X_3 is the superior choice. Our final tree looks like this:



So our classification for the unknown data points are as follows:

X_1	X_2	X_3	y
+	-	-	+
-	-	-	+
+	-	+	+

HW4 Q2 Part 2 AdaBoost

Label the datapoints as follows:

	X_1	X_2	X_3	y
a)	-1	1	1	-1
b)	1	1	1	1
c)	-1	1	-1	1
d)	-1	-1	1	-1
e)	1	1	-1	1

$$W = \begin{bmatrix} 1/5 \\ 1/5 \\ 1/5 \\ 1/5 \\ 1/5 \end{bmatrix} = \begin{bmatrix} w_a \\ w_b \\ w_c \\ w_d \\ w_e \end{bmatrix}$$

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$

Step 1: Using X_1 as classifier gives $\frac{4}{5}$ accuracy, X_2 gives 4/5 accuracy, and X_3 gives 1/5 accuracy. So we'll pick X_1 as our decision stump.

~~So triple the weight of C_1 which was classified incorrectly.~~ So $W = \begin{bmatrix} w_a \\ w_b \\ w_c \\ w_d \\ w_e \end{bmatrix} = \begin{bmatrix} 1/5 \\ 1/5 \\ 1/5 \\ 1/5 \\ 1/5 \end{bmatrix}$

Weight vector; weights at each point

$$\varepsilon_t = \frac{1}{5}$$

$$\alpha_t = \frac{1}{2}\ln\left(\frac{4/5}{1/5}\right) = 0.693$$

$$w_i = \frac{w_i \cdot e^{-\alpha_t y_i h_t(x_i)}}{Z}$$

$y_i h_t(x_i) = 1$ if correct, -1 if incorrect

So $W = \begin{bmatrix} 0.161 \\ 0.161 \\ 0.356 \\ 0.161 \\ 0.161 \end{bmatrix}$

HW4 Q2 Part 2 AdaBoost continued

Step 2:

x_1 accuracy = 0.644 $\varepsilon = .356$

x_2 accuracy = 0.839 $\varepsilon = .161$

x_3 accuracy = 0.161 $\varepsilon = .839$

Select x_2 as classifier

$\alpha_t = \frac{1}{2} \ln\left(\frac{0.839}{0.161}\right) = 0.8254$

$w_i = \frac{w_i \cdot e^{-\alpha_t y_i h_t(x_i)}}{Z}$ Renormalization =

$$\begin{bmatrix} 0.3675 \\ 0.1114 \\ 0.3184 \\ 0.1114 \\ 0.1114 \end{bmatrix}$$

$$w = \begin{bmatrix} 0.360 \\ 0.109 \\ 0.312 \\ 0.109 \\ 0.109 \end{bmatrix}$$

x_1 accuracy = 0.688

x_2 accuracy = 0.64

x_3 accuracy = 0.109

Step 3: $-1 \cdot x_3$ accuracy = 0.891

Select $-1 \cdot x_3$ as classifier, $\varepsilon = 0.109$

$\alpha_t = \frac{1}{2} \ln\left(\frac{0.891}{0.109}\right) = 1.05$

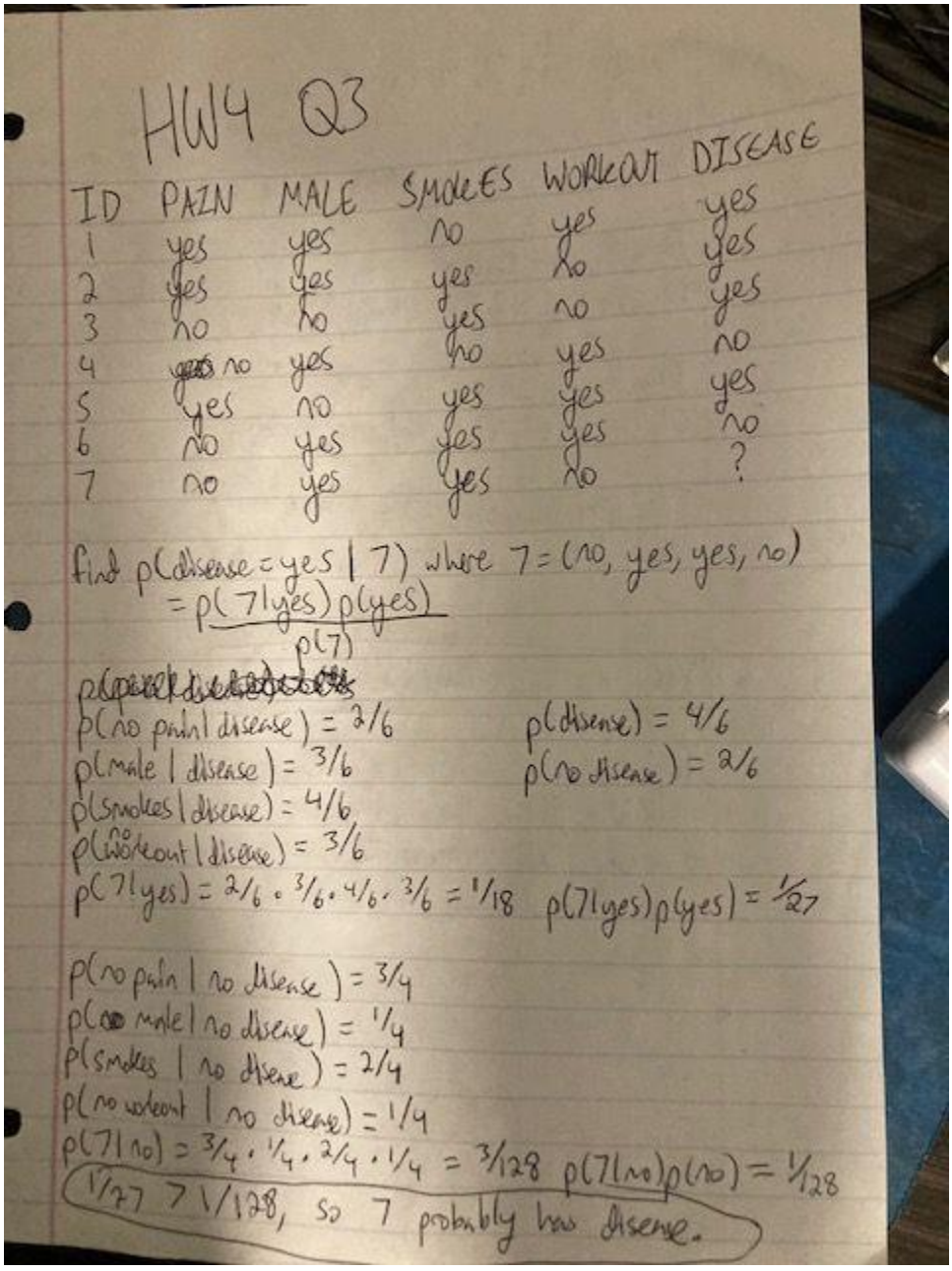
Don't need final weight vector, since this is the last iteration.

$H(x) = \text{sign}(0.693 \cdot x_1(x) + 0.8254 \cdot x_2(x) - 1.05 \cdot x_3(x))$

Predictions

x_1	x_2	x_3	y
+	−	−	+
−	−	−	−
+	−	+	−

Question 3:



HW4 Q3

ID	PAIN	MALE	SMOKES	WORKOUT	DISEASE
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	yes
6	no	yes	yes	yes	no
7	no	yes	yes	no	?

find $p(\text{disease} = \text{yes} \mid 7)$ where $7 = (\text{no}, \text{yes}, \text{yes}, \text{no})$

$= \frac{p(7 \mid \text{yes})\, p(\text{yes})}{p(7)}$

$p(\text{no pain} \mid \text{disease}) = 2/6$

$p(\text{male} \mid \text{disease}) = 3/6$

$p(\text{smokes} \mid \text{disease}) = 4/6$

$p(\text{no workout} \mid \text{disease}) = 3/6$

$p(\text{disease}) = 4/6$

$p(\text{no disease}) = 2/6$

$p(7 \mid \text{yes}) = 2/6 \cdot 3/6 \cdot 4/6 \cdot 3/6 = 1/18$ $\quad p(7 \mid \text{yes})\, p(\text{yes}) = 1/27$

$p(\text{no pain} \mid \text{no disease}) = 3/4$

$p(\text{male} \mid \text{no disease}) = 1/4$

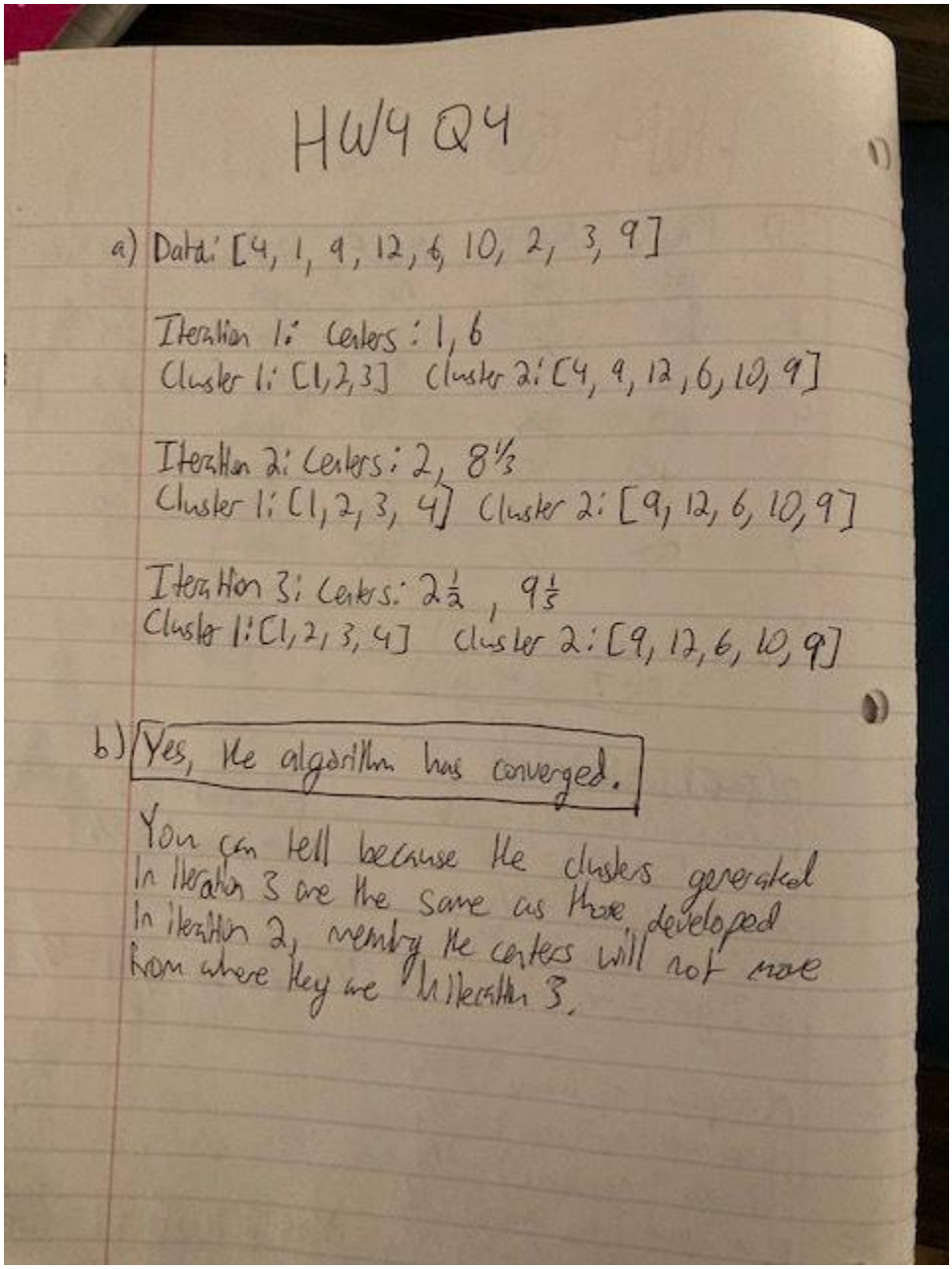
$p(\text{smokes} \mid \text{no disease}) = 2/4$

$p(\text{no workout} \mid \text{no disease}) = 1/4$

$p(7 \mid \text{no}) = 3/4 \cdot 1/4 \cdot 2/4 \cdot 1/4 = 3/128$ $\quad p(7 \mid \text{no})\, p(\text{no}) = 1/128$

$1/27 > 1/128$, so 7 probably has disease.

Question 4:



HW4 Q4

a) Data: [4, 1, 9, 12, 6, 10, 2, 3, 9]

Iteration 1: Centers: 1, 6
Cluster 1: [1, 2, 3] Cluster 2: [4, 9, 12, 6, 10, 9]

Iteration 2: Centers: 2, 8⅓
Cluster 1: [1, 2, 3, 4] Cluster 2: [9, 12, 6, 10, 9]

Iteration 3: Centers: 2½, 9⅓
Cluster 1: [1, 2, 3, 4] Cluster 2: [9, 12, 6, 10, 9]

b) Yes, the algorithm has converged.

You can tell because the clusters generated in iteration 3 are the same as those developed in iteration 2, meaning the centers will not move from where they are in iteration 3.