CS 6375
ASSIGNMENT 1

Names of students in your group:

1. Rutuja Kaushike (rnk170000)
2. Akhila Kancharana (axk180025)

CS 6375.502

Number of free late days used: 0

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

1. Machine Learning (Book by Tom Michelle)

1.

Python code for question 1 is as follows:

```
def question1(theta0, theta1, alpha, examples, m):
    length=len(examples)
    for i in range(m):
        err= (theta0 + sum(((theta1*examples[i][0])-examples[i][1])**2 for i in
range(length)))/(2*length)
        print("Theta 0 is ", theta0, "Theta 1 is ", theta1, "and Error is ",err)
        der1=[theta0 + (theta1*examples[i][0]) for i in range(length)]
        newTheta0=theta0-(alpha*sum(der1))/length
        der2=[theta1*examples[i][0] for i in range(length)]
        newTheta1=theta1-alpha*sum(der2)/length

        theta0=newTheta0
        theta1=newTheta1

examples=[(3,2),(1,2),(0,1),(4,3)]
question1(0,1,0.0245,examples,5)
```

The output is as follows:

```
>>>
====== RESTART: /Users/rutujakaushike/Documents/assignment1question1.py
======
Theta 0 is  0 Theta 1 is  1 and Error is  0.5
Theta 0 is  -0.049 Theta 1 is  0.951 and Error is  0.4281782499999999
Theta 0 is  -0.0943985 Theta 1 is  0.904401 and Error is  0.37450398610325
Theta 0 is  -0.13640138575 Theta 1 is  0.860085351 and Error is
0.33670020754682795
Theta 0 is  -0.175203733998125 Theta 1 is  0.8179411688010001 and Error is
0.3127338950087588
>>>
```

Hence, error goes down as iterations increase.

Q2.
→

	test positive	test negative
Has disease	0.8	
Doesn't have disease		0.9

Here, as given in the example, a testing machine can identify the disease in 80% of the cases i.e true positive here is 80% i.e 0.8
Also, here true negative is 90% i.e a machine is able to correctly predict for those who do not have disease 90% of the time.
∴ true negative is 90% i.e. 0.9.
Now, as true positive is 0.8, ~~false~~ ~~tve~~ false negative should be 0.2 which is 1 − true positive
As,

True Positive = 1 − False Negative

0.8 = 1 − false negative

∴ false Negative is 0.2 i.e. 20% (20 percent)

Similarly,

False positive = 1 − true negative

= 1 − 0.9

= 0.1

Hence, false positive is 0.1 i.e. 10% (10 percent)

Q3. PROS AND CONS OF SPECIFIC AND GENERAL HYPO.

→

a. Selecting the most specific hypothesis on a training data.

→

PROS:

1. Result is not affected due to the order of examples which we proceed.
2. works fine with examples with only positive data.
3. Represents only specific hypothesis out of all consistent ones. Only specific boundaries are shown/represented.

CONS:

1. Avoids negative examples.
2. No room for generalization.
3. If noisy data comes, result gets affected.

b. Selecting the most general hypothesis (G) based on a training data

→

PROS

1. Works on both positive and negative data
2. ~~not affected~~ Result is not affected due to

CONS

1. Noisy data misleads the hypothesis (General Hypothesis)

the order of examples.	2. Incosistent data leads to exhaustion.
3. Finds most ~~spe~~ general hypothesis based on training data	

Q.4. Consistent hypothesis :

A hypothesis h is consistent with a set of training example D if and only if $h(x) = c(x)$ for each example $\{x, c(x)\}$ in D.

$$\text{Consistent}(h, D) = (\forall (x, c(x)) \in D)\ h(x) = c(x)$$

Version Space : The version space, denoted $VS_{H.D}$ with respect to hypothesis space H and training examples D, is the subset of hypothesis from H consistent the training examples in D.

$$VS_{H,D} = \{h \in H \mid \text{Consistent}(h, D)\}$$

Q5. The most general hypothesis has ? value for each attribute.

Q6 $f : X \to Y$, $X = (x_1, x_2, x_3, x_4)$

(a) How many instances ~~of~~ i.e. $|X|$ are possible?

→ x_1, x_2, x_3 and x_4 are boolean values. Hence, assuming that each attribute is included, number of instances = 2 for each attribute

Hence

$= 2 \times 2 \times 2 \times 2 = 2^4 = 16$ instances.

(b) Let's assume that there are four literals.

$\therefore$ total instance space will be $2^4 = 16$.

~~Now~~, (for every literal, there is a positive i.e. attribute and a negation of that attribute $\therefore$ 2 for each. $\therefore 2 \times 2 \times 2 \times 2 = 2^4$.)

For every attribute and negation of that, we have 2 values, 0 or 1. $\therefore$ we will have

$2^{|x|} = 2^{2^4} = 2^{16}$ in total.

(c) For each attribute, there are three labellings i.e. (1, 0, ?) hence, the $|X|$ i.e. instance space will be 3 for each attribute i.e. 3^4 in total.

Now, for every hypothesis we propose, we have positive or a negative choice.

i.e. again 2 choices for every hypothesis

Hence, we will have

$2^{|x|} = 2^{3^4} = 2^{81}$ in total.

(d) Here, we have depth 2 and 2 attributes.
Hence, we can choose ~~4 C~~ 2 out of 4 attributes for depth 2.
Hence, for a combination problem (~~two~~ (order is not important),
we will get

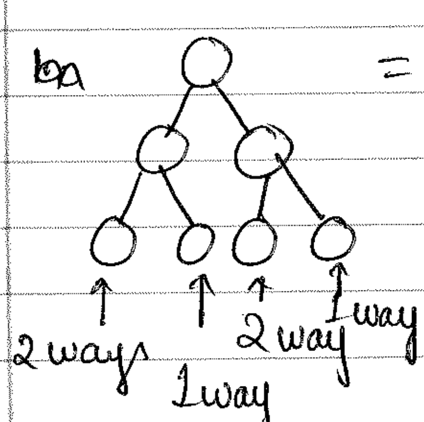
$$^4C_2 = \frac{4!}{2!\,2!} = \frac{\cancel{4}^{2}\times 3\times\cancel{2}\times 1}{2\times\cancel{2}} = 6 \text{ trees.}$$

$$\left[\,^nC_r = \frac{n!}{(n-r)!\,r!}\right]$$

If ordering is important and ~~if is diff~~ newly ordered tree is considered as new tree
we will have

$$^4P_2 = \frac{4!}{2!} = 12 \text{ trees.} \quad \left[\,^nP_r = \frac{n!}{(n-r)!}\right]$$

(e) ~~ba~~



$= 2\times1\times2\times1 = 4$ ways to choose.
we have 6 distinct trees from Ans (d), and we can label them in 4 diff ways.
Hence, total ways
$= 4\times 6 = 24$.

~~Q6.~~ Find 'S' algorithm.

Q7.

→ Let's start 'h' with

$S_0 = \{\phi, \phi, \phi, \phi, \phi\}$

Example 1 is a positive example ∴ After applying it on S_0, S_1 becomes.

$(\langle 1, 1, 0, 1, 1 \rangle, 1)$

$S_1 = \{1, 1, 0, 1, 1\}$

example 2 is a negative example ∴ we will exclude that.

∴ S_1 and S_2 will be same.

$S_2 = \{1, 1, 0, 1, 1\}$

example 3 is a positive example, ∴ we will consider that.

∴ $(\langle 1, 1, 1, 1, 0 \rangle, 1)$

$S_3 = \{1, 1, ?, 1, ?\}$

example 4 is a negative example ∴ we will not consider that. ∴ S_3 and S_4 will be the same.

∴ $S_4 = \{1, 1, ?, 1, ?\}$

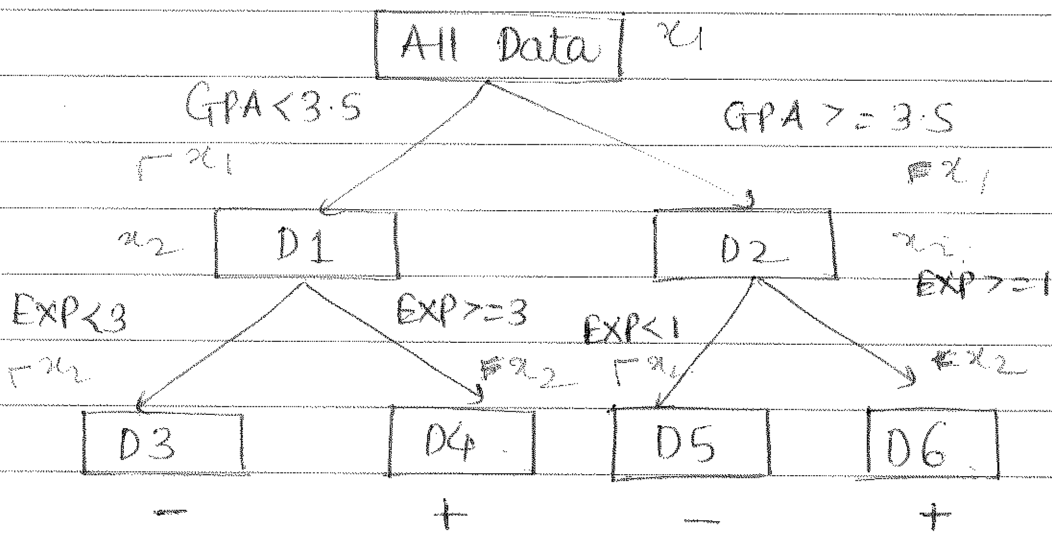
example 5 is a positive example and we will consider it.

$(\langle 1, 1, 1, 1, 1 \rangle, 1)$

Hence,

$S_5 = \langle 1, 1, ?, 1, ? \rangle$

Q8.



Final hypothesis shown by this decision tree in the form of Disjunctive Normal Form (DNF).

$(\neg x_1 \wedge x_2) \vee (x_1 \wedge x_2)$

→ We will consider only positive examples for this specific hypothesis.

∴ for D4, CNF is : (GPA < 3.5 ∧ EXP>=3)

∴ for D6, CNF is : (GPA>=3.5 ∧ EXP>=1)

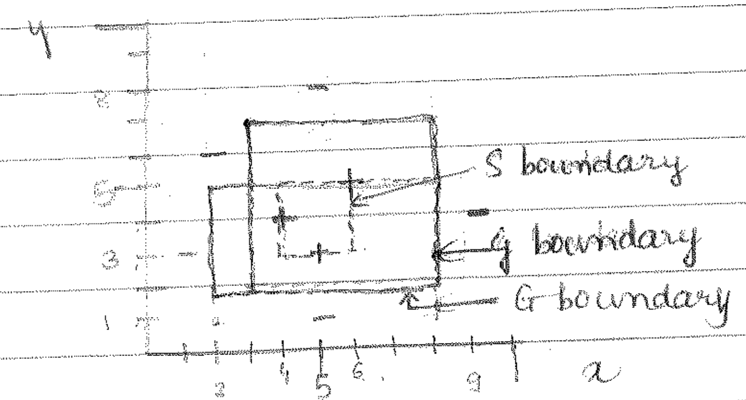
Hence, DNF for this hypothesis will be :

D4 ∨ D6.

∴ Final hypothesis is

(GPA < 3.5 ∧ EXP>=3) ∨ (GPA>=3.5 ∧ EXP>=1)

Q.9



ⓐ Hypotheses for S boundary is: (4, 6, 3, 5)

i.e. for S(4,6,3,5) $4 \leq x \leq 6$ and $3 \leq y \leq 5$

3, 8, 2, 7

i.e. Hypothesis for S is

$S = (4 \leq x \leq 6,\ 3 \leq y \leq 5)$

ⓑ G boundary of this version space.

For G boundary,

(3, 8, 2, 7) and (2, 8, 2, 5)

As these two hypotheses sets only contains positive examples. Hence, it is the G boundary.

Hence,

G boundary (3, 8, 2, 7)

$G = (3 \leq x \leq 8,\ 2 \leq y \leq 7)$

and (2, 8, 2, 5)

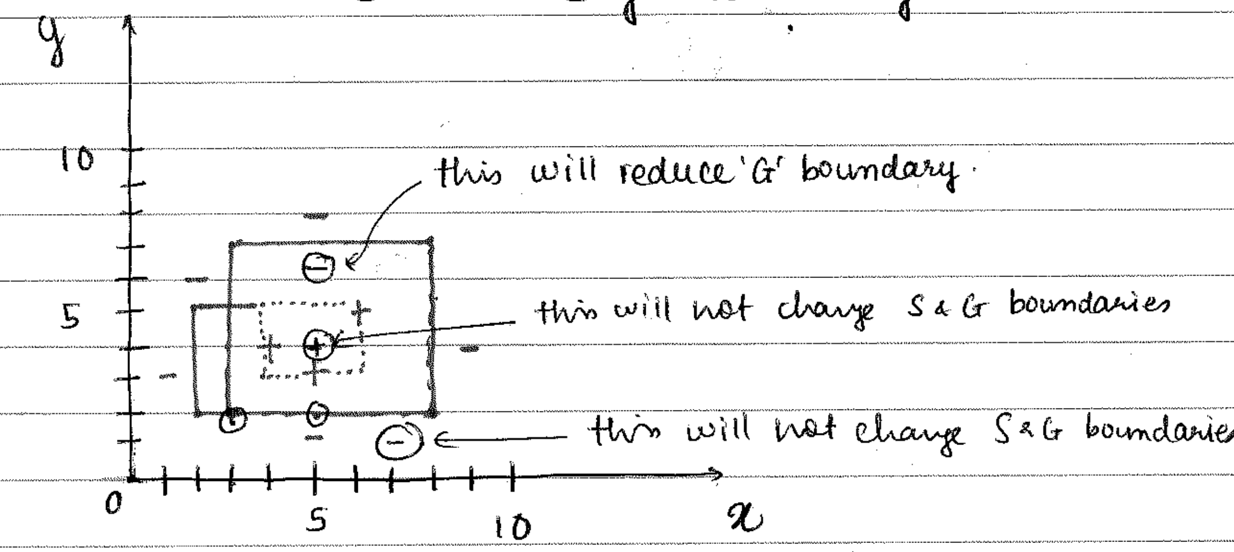
$G = (\cancel{2\leq8}\ 2 \leq x \leq 8,\ 2 \leq y \leq 5)$

Q9

(c) Query to reduce the size of the version space,

(i) If it is a -ve example, if we put it anywhere in between S and G boundary, it will reduce the hypothesis space of 'G'

for example, if (5,6) is a -ve example, it will reduce the 'G' boundary



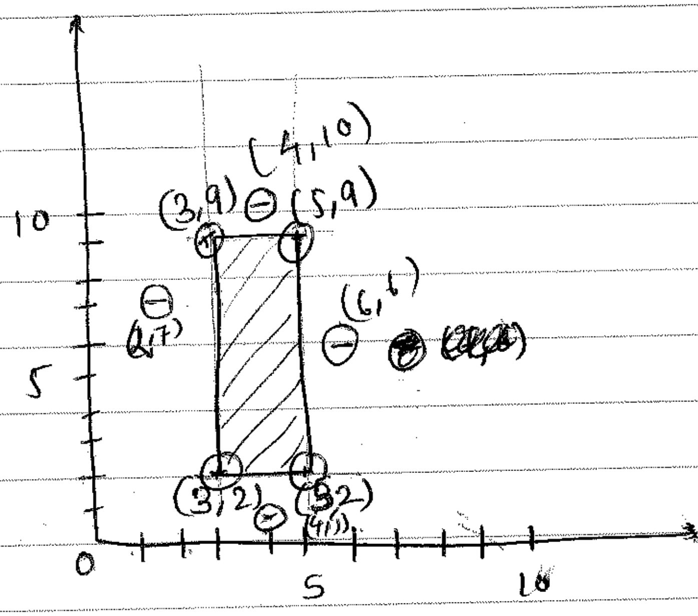
Now, for no change boundaries,

(i) if we define any positive point in ~~between~~ 'S' region / boundary, it will not change the existing 'S' & 'G' boundaries.

for example, if (5,4) is a positive example, it won't change S & G boundaries.

(ii) if we define any -ve point outside G boundary for example, (7,1) which is a -ve entry, G boundary and S boundary will not change.

(d)



Let's take the target concept as
$(3 \leq x \leq 5,\ 2 \leq y \leq 9)$.
the version space is as shown above.
for this, any two diagonal +ve boundary points will suffice for perfect learning.
say, (3,2) and (5,9) are enough to make positive boundaries work.
For negative points (G should exclude those), any 4 -ve points outside version space will be enough
i.e. (2,7), (4,1), (6,6) and (4,10) will be sufficient for drawing General Boundaries.
Hence, any 2 positive points like (3,2), (5,9) and 4 negative points like (2,7)(4,1)(6,6)(4,10)
i.e. 6 points min. are required.

Q10.

(a) $S_0 = \langle(\phi, \phi, \phi, \phi)(\phi, \phi, \phi, \phi)\rangle$

$G_0 = \langle(?, ?, ?, ?)(?, ?, ?, ?)\rangle$

first example is a positive one.

$\therefore$ $S_1 = \langle(ug, se, l, hs), (gr, cs, h, hs)\rangle$

$G_1 = \langle(?, ?, ?, ?), (?, ?, ?, ?)\rangle$

second example is also a positive one.

$\therefore$

$S_2 = \langle(ug, se, ?, ?)(gr, cs, h, hs)\rangle$

$G_2 = \langle(?, ?, ?, ?)(?, ?, ?, ?)\rangle$

third example is -ve though.

$\therefore$ we won't change S_2 for S_3

$\therefore S_3 = \langle(ug, se, ?, ?), (gr, cs, h, hs)\rangle$

& $G_3 = \langle(ug, ?, ?, ?)(?, ?, ?, ?)\rangle$

$\langle(?, ?, ?, ?)(?, ?, ?, hs)\rangle$.

Now, example 4 is +ve.

$\therefore$ $S_4 = \langle(ug, se, ?, ?)(gr, ?, h, ?)\rangle$

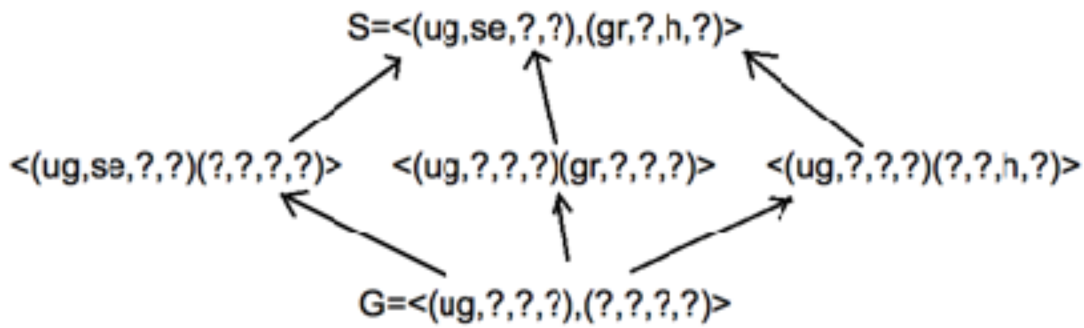
and $G_4 = \langle(ug, ?, ?, ?)(?, ?, ?, ?)\rangle$

10.
b)

From 10.a) we have,

S=<(ug,se,?,?),(gr,?,h,?)>

G=<(ug,?,?,?),(?,?,?,?)>



Hence, there are total 6 consistent hypotheses present.
Given hypotheses <(ug, cs, h, do), (gr, ma, l, se)> with class label '+' is then, consistent with three and inconsistent with three.