CS 6375

ASSIGNMENT -1

Names of students in your group:

Dafne Navita Maria Joseph – dxm172530

Yuxin Xi -

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.

2. False Positive - 10 %

False Negative - 20 %

3.

a. Selecting the most specific hypothesis (S) based on a training data.

Pros :

->Most specific hypothesis has less false positive errors.

->If new data comes in and S is still consistent, then we would have a precise knowledge of the concept we need to learn.

Cons :

->Selecting the most specific hypothesis can result in false negative errors .

-> S is not likely to be consistent with all the new data that comes in.

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

Pros :

->Most general hypothesis has less false negative errors.

-> G is likely to be consistent with the new data coming in

Cons :

-> Selecting the most general hypothesis can lead to false positive errors.

4.**Consistent hypothesis :** Any hypothesis which matches with the training data is called consistent hypothesis .

A hypothesis h is consistent with a set of training examples D if and only if h(x) = c(x) for each example hx, c(x)i in D.

Consistent(h, D) ≡ (∀hx, c(x)i ∈ D)h(x) = c(x)

**Version Space** : Version space represents the entire space of consistent hypotheses .

The version space, denoted V SH,D, with respect to hypothesis space H and training examples D, is the subset of hypotheses from H consistent with the training examples in D.

***V S H,D*** = *{h* **E *H | Consistent(h, D)]***

5. The most general hypothesis has **Don’t care (?)** value for each attribute.

6.

a. The number of possible instances for |X| is : **24  = 16** since there is a total of 4 attributes with each attribute having 2 choices .

b. The number of possible hypotheses is **: 216 = 65536**

c. The number of possible labellings with 3 choices **: 281**

d. Fully-grown decision trees of depth 2 using only 2 attributes : 6 \*2 = **12**

e. No. of labels each tree can have : 2 \*2 = **4**

7. After applying the Find – S algorithm , the following hypotheses can be arrived .

i) h0 = (∅,∅,∅,∅,∅)

ii)h1 = (1,1,0,1,1)

iii)h2= (1,1,0,1,1) – No change since it’s a negative example

iv)h3=(1,1,?,1,?)

v)h4=(1,1,?,1,?) – no change : negative example

vi)h5= (1,1,?,1,?)

Result : ( xGPA > 3.5 , xWorkEx > 2, xCS6375 Don’t care, xCS6350 Taken, xJava Don’t care )

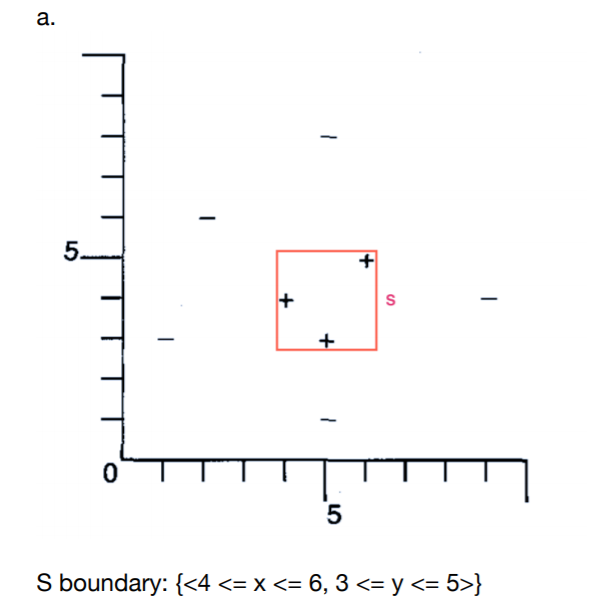
The result states , the set of students who got internship from the above dataset :

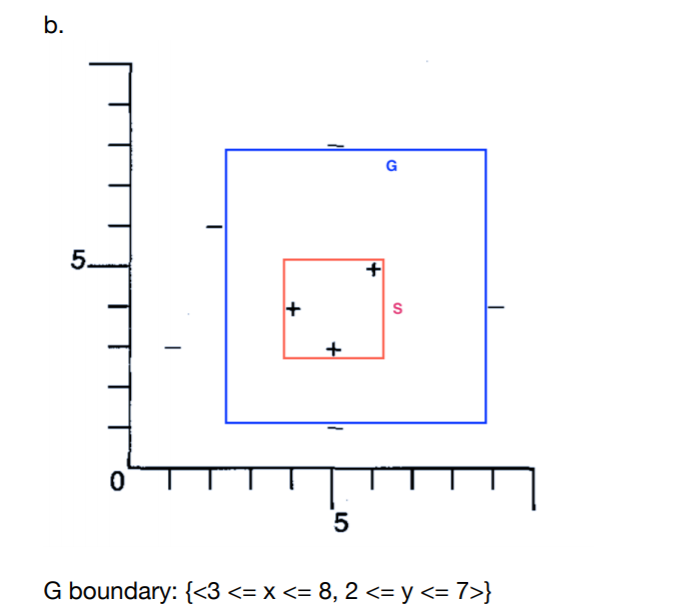
Have a CGPA of above 3.5 , have more than 2 years of experience , doesn’t care about the course CS6375 , has taken the course CS6350 and doesn’t care if he has taken advanced Java Skills .

8. DNF :

(GPA<3.5 ^ Exp >= 3 ) ∨ (GPA>=3.5 ^ Exp >=1 )

9.





1. S ≡ (4 ≤ x ≤ 6) ∧ (3 ≤ y ≤ 5) – the red boundary is the most specific boundary since it includes only positive examples.
2. G ≡ (3 ≤ x ≤ 8) ∧ (2 ≤ y ≤ 7) – the blue boundary is the most general boundary .
3. A query lying between S and G would guarantee to reduce the size of the version space, Eg : x = 5, y = 6 .

A query outside G or inside S would not reduce size of the version space,

Eg : x = 1, y = 1 .

1. Four training samples would be sufficient to achieve a target concept via Candidate-Elimination algorithm.

For Eg : ( 3 <= x <= 5 , 2 <= y <= 9 ) – For this example , the below training samples would suffice to learn the target concept .

+ (3,2)

+ (5,9)

- (2,1)

- (6,10)

10. Trace :

a)

i) <(ug,se, l, hs),(gr, cs, h, hs)> , +

S = <(ug,se, l, hs),(gr, cs, h, hs)>

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

ii) <(ug,se, h, fr),(gr, cs, h, hs> , +

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

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

iii)<(gr,se, l,so),(gr, cs, h,se)> -

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

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

iv)<(ug,se, l, ju),(gr,se, h, ju)>+

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

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

b) The total number of consistent hypothesis at the end of candidate elimination algorithm is **2.**

Only the most general hypothesis G = <(ug,?,?,?),(?,?,?,?)> is consistent with the given example.