**Homework 1**

**COSC 6342: Machine Learning**

**Submitted by**

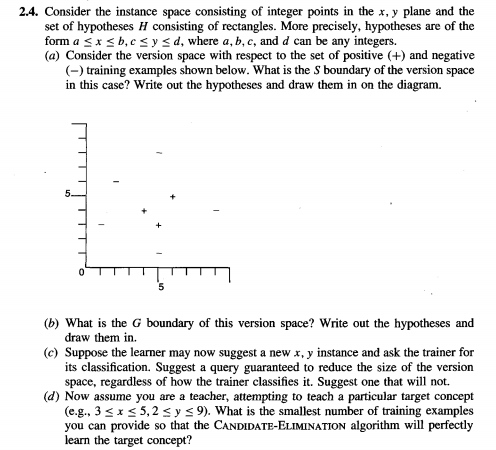
**S M Salah Uddin Kadir (1800503)**

**Rubayat Jinnah ()**

**Concept Learning**

Machine Learning by Tom Mitchell, Chapter 2.

**Question 2.4. (25 points)**



**Solution:**

**(a)**

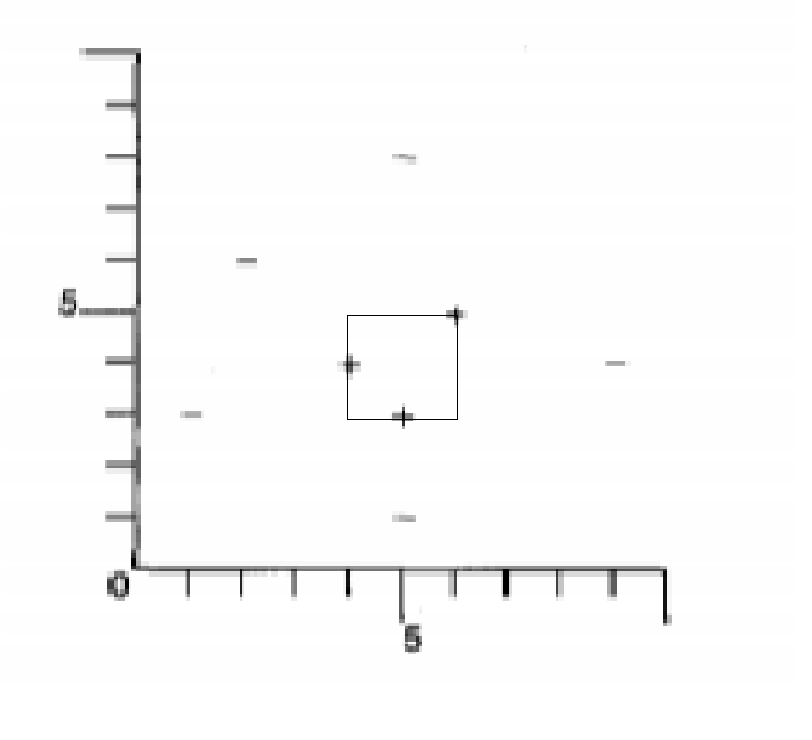


Fig: S boundary of the version space

So, from the graph, we can see that the most specific hypothesis is,

S: {(4 <= x <= 6), (3 <= y <= 5)}

This assumes that a rectangle is at minimum 1 x 1. This is also assuming that generalization or specification is the decreasing or increasing in value of a, b, c, or d.

**(b)**

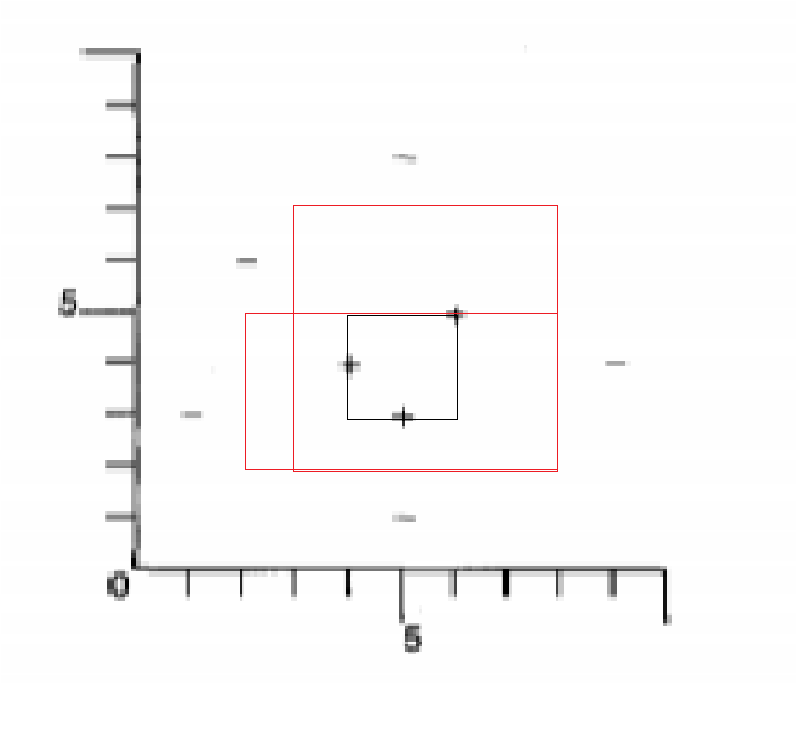


Fig: G boundary of the version space

So, from the graph, we can see the most general hypothesis (indicated by red color),

G: {(3 <= x <= 8), (2 <= y <= 7)}

and,

G: {(2 <= x <= 8), (2 <= y <=5)}

**(C)**

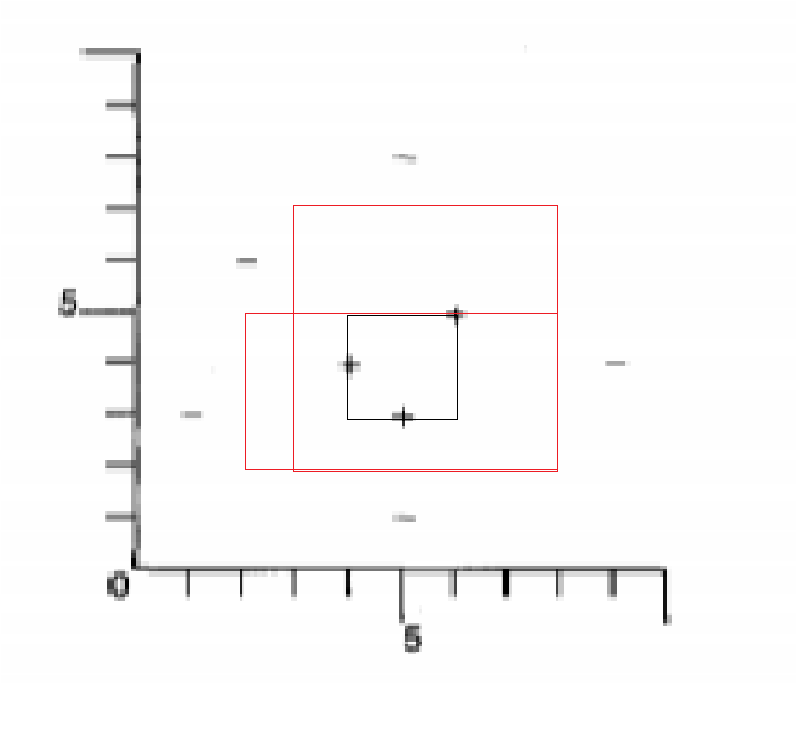


Fig-3

From the figure, we can see that the learner could request (5, 6) for classification. Actually, any point in (4 <= x <= 7, y = 6) or (x = 7, 3 <= y <= 6) will work. This is because the points along these two lines are between the version space bounds identified by S and G. Since S and G should converge upon one hypothesis, one must generalize or specialize, respectively.

By selecting (5, 4), (4, 5), (5, 5), (6, 3), (6, 4), or (9, 9) reducing the space should be avoided. Since these points are already included by S, there should be no change in the space. This is a result of the bias imposed by the hypothesis representation.

So, any point within the G boundary and outside the S boundary would reduce the Version Space, and anything outside the G boundary or within the S boundary would not reduce the Version Space.

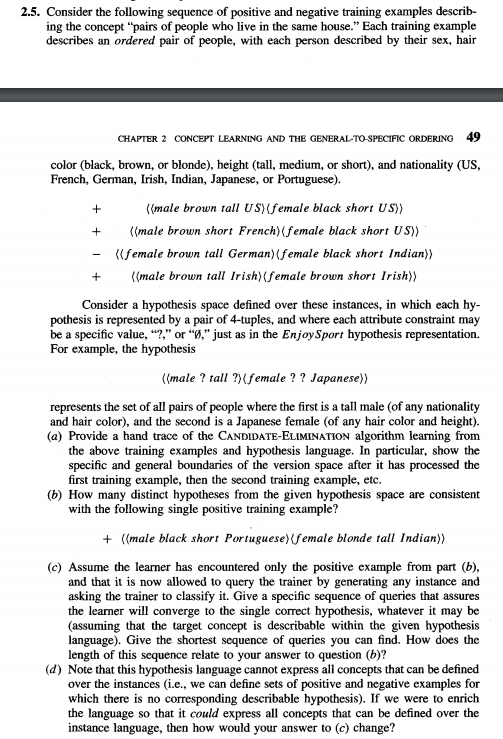
**(d)**

We need 8 points. 4 positive points delineating the target rectangle, and 4 negative points

delineating a rectangle exactly 1 unit outside the target rectangle.

For example, if the rectangle is 3 <= x <= 5, 2 <= y <= 9. The 4 positive points could be (3,2), (3,9), (5,2), (5,9). And 4 negative points could be (2,1), (2,10), (6,1), (6,10).

**Question 2.5. (25 points)**



**Solution:**

**(a)**

We have,

+ ((male brown tall US) (female black short US))

+ ((male brown short French) (female black short US))

- ((female brown tall German) (female black short Indian))

+ ((male brown tall Irish) (female brown short Irish))

Let the initial state,

S: ((0, 0, 0, 0) (0, 0, 0, 0))

G: ((?, ?, ?, ?) (?, ?, ?, ?))

Adding the first pair which is positive,

S: ((male, brown, tall, US) (female, black, short, US))

G: ((?, ?, ?, ?) (?, ?, ?, ?))

Adding the second pair which is positive,

S: ((male, brown, ?, ?) (female, black, short, US))

G: ((?, ?, ?, ?) (?, ?, ?, ?))

Adding the third pair which is negative,

S: ((male, brown, ?, ?) (female, black, short, US))

G: ((male, ?, ?, ?) (?, ?, ?, ?) ), ( (?, ?, ?, ?) (?, ?, ?, US))

Adding the fourth pair which is positive,

S: ((male, brown, ?, ?) (female, black, short, ?))

G: ((male, ?, ?, ?) (?, ?, ?, ?))

So, after adding all the training examples,

The specific boundary is, ((male, brown, ?, ?) (female, black, short, ?))

And

The general boundary is, ((male, ?, ?, ?) (?, ?, ?, ?))

**(b)**

We have a single positive training example,

+ ((male, black, short, Portuguese) (female, blonde, tall, Indian))

There are 8 attributes in the given hypothesis. Each attribute can have either the specified value or “?”. So, the total number of consistent hypothesis is,

2^8 = 256

**(c)**

The given positive instance is,

((male, black, short, Portuguese) (female blonde tall Indian)).

We know from question (b) that there are 256 consistent hypotheses. For each query, the hypothesis space can be reduced by half. So, we need at most log2 256 = 8 queries to reach the final target hypothesis.

One approach is simply to propose queries with all attributes having the same value as the single

positive original example except for one attribute. That would guarantee convergence to a solution.

As an example,

((male, black, short, Portuguese) (female, blonde, tall, Indian))

((male, brown, short, Portuguese) female, blonde, tall, Indian))

((male, black, tall, Portuguese) (female, blonde, tall, Indian))

((male, black, short, French) (female, blonde, tall, Indian))

((male, black, short, Portuguese) (male, blonde, tall, Indian))

((male, black, short, Portuguese) (female, brown, tall, Indian))

((male, black, short, Portuguese) (female, blonde, short, Indian))

((male, black, short, Portuguese) (female, blonde, tall, US))

This is consistent with the 2^8 calculation for the number of total hypotheses consistent with the original training example, or any positive training example for that matter. While G was not examined closely, since the attributes were generalized to "?" and converged with G or took on the value from the first training example and the inconsistent hypotheses in G were removed and replaced by those more specific. This eventually leads to S and G converging.

**(d)**

Instead of checking just 2 possible values for attributes in the hypotheses, we would have to check for every combination of values over all the possible values in the instance space.

In this example,

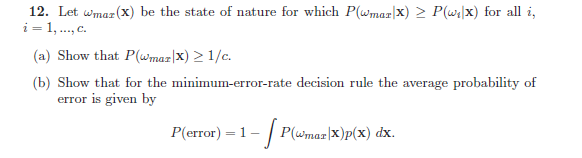
the size of the feature space, N = 2\*3\*3\*7\*2\*3\*3\*7 = 15,876.

So, the number of possible hypotheses is, 2^N, and we would need at most log22^N = N queries to find the target concept.

So, we need as many queries as instances there are in the entire feature space to be sure to find the target concept.

**Probabilistic Learning**

**Question 12 (a) and (b) only. (25 points)**



**Solution:**

**(a)**

We know, summation of all probability is 1,

= 1

Now, if the distribution of all probability is equal that,

P(Wi | X) = P(Wj | X)

then we can write,

P(Wi | X) = P(Wj | X) = 1/c.

So the maximum probability will be also 1/c that,

P(Wmax | X) = 1/c

Now, if any probability is less than 1/c then some others probability will be increased to make it 1. In that case, our maximum probability will be,

P(Wmax | X) > 1/c.

So, applying both cases, we can say that

P(Wmax | X) >= 1/c.

**(b)**

We know,

P(error) = ∫ P(error, x) dx = ∫ P(error | x) p(x) dx

and if for every x we minimize the error then P(error | x) is as small as it can be.

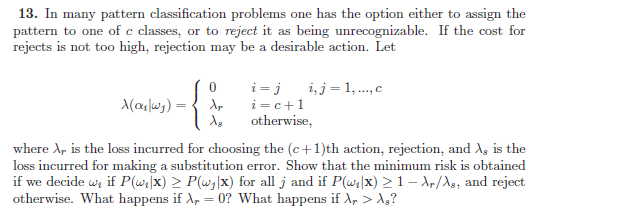
We also know that,

Probability of error = 1 – Probability of correct.

So,

P(error) = 1 - ∫ P(Wmax | X) p(x) dx

**Question 13. (25 points)**



**Solution:**

For i = 1,...,c,

R(αi | x) =

= λ­­­s

= λ­s [1 − P(ωi | x)] .

For i = c + 1,

R(αc+1 | x) = λr

Therefore, the minimum risk is achieved

If, we decide ωi if R(αi | x) ≤ R(αc+1|x),

i.e., P(ωi | x) ≥ 1 – λr / λs , and reject otherwise.

If, λr = 0, we always reject.

If, λr > λs, we will never reject