Homework 1
COSC 6342: Machine Learning

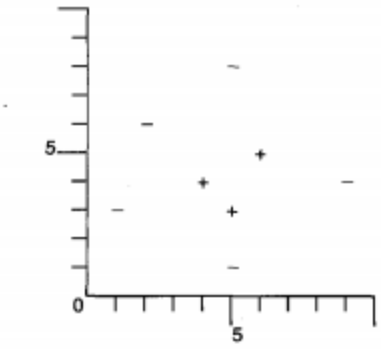
Submitted by
S M Salah Uddin Kadir (1800503)
Rubayat Jinnah (1891217)

Concept Learning

Question 2.4.

2.4. Consider the instance space consisting of integer points in the x, y plane and the set of hypotheses H consisting of rectangles. More precisely, hypotheses are of the form $a \leq x \leq b, c \leq y \leq d$, where a, b, c, and d can be any integers.

(a) Consider the version space with respect to the set of positive (+) and negative (−) training examples shown below. What is the S boundary of the version space in this case? Write out the hypotheses and draw them in on the diagram.



(b) What is the G boundary of this version space? Write out the hypotheses and draw them in.

(c) Suppose the learner may now suggest a new x, y instance and ask the trainer for its classification. Suggest a query guaranteed to reduce the size of the version space, regardless of how the trainer classifies it. Suggest one that will not.

(d) Now assume you are a teacher, attempting to teach a particular target concept (e.g., $3 \leq x \leq 5, 2 \leq y \leq 9$). What is the smallest number of training examples you can provide so that the CANDIDATE-ELIMINATION algorithm will perfectly learn the target concept?

Solution:

(a)

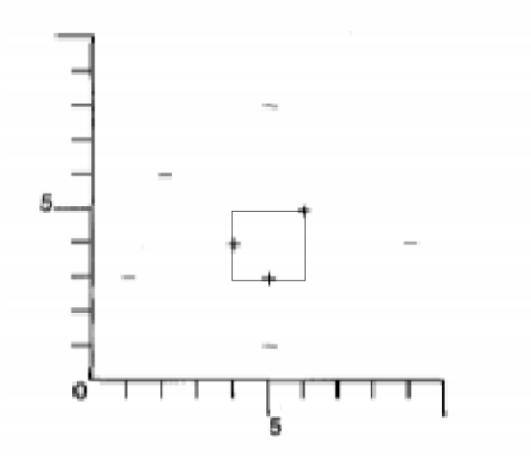


Fig: S boundary of the version space

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)

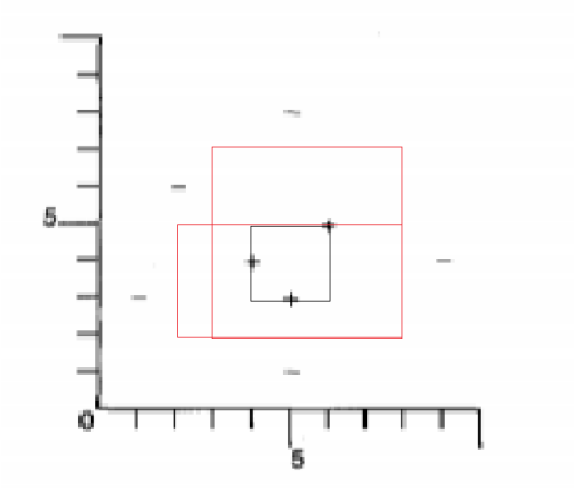


Fig: G boundary of the version space

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)

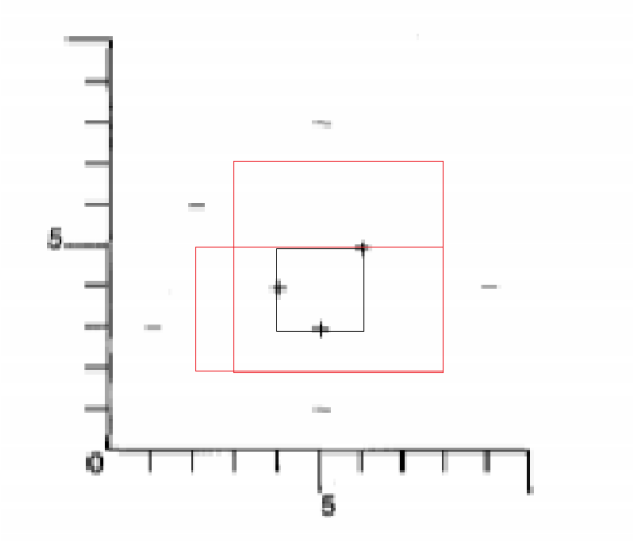


Fig-3

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 negative and positive examples to learn the target concept where G and S will converge. We can use diagonal two points of a rectangle in the x, y plane to provide training examples.

For the given target concept (e.g., 3 <= x <= 5, 2 <= y <= 9), we can use diagonal two points (3, 2) and (5, 9) as positive training example to make the target rectangle. We can also use other two diagonal points (2, 1) and (6, 10) as negative example where general and specific hypothesis will converge. These two points are exactly 1 unit in opposite direction for both x and y axis from the target rectangle.

So, the minimum no. of training example we can provide so that the CANDIDATE-ELIMINATION Algorithm will learn the target concept is (2+2) = 4.

Question 2.5. (25 points)

2.5. Consider the following sequence of positive and negative training examples describing the concept "pairs of people who live in the same house." Each training example describes an *ordered* pair of people, with each person described by their sex, hair

color (black, brown, or blonde), height (tall, medium, or short), and nationality (US, French, German, Irish, Indian, Japanese, or Portuguese).

$$+ \quad \langle \langle male\ brown\ tall\ US \rangle \langle female\ black\ short\ US \rangle \rangle$$

$$+ \quad \langle \langle male\ brown\ short\ French \rangle \langle female\ black\ short\ US \rangle \rangle$$

$$- \quad \langle \langle female\ brown\ tall\ German \rangle \langle female\ black\ short\ Indian \rangle \rangle$$

$$+ \quad \langle \langle male\ brown\ tall\ Irish \rangle \langle female\ brown\ short\ Irish \rangle \rangle$$

Consider a hypothesis space defined over these instances, in which each hypothesis is represented by a pair of 4-tuples, and where each attribute constraint may be a specific value, "?," or "Ø," just as in the *EnjoySport* hypothesis representation. For example, the hypothesis

$$\langle \langle male\ ?\ tall\ ? \rangle \langle female\ ?\ ?\ Japanese \rangle \rangle$$

represents the set of all pairs of people where the first is a tall male (of any nationality and hair color), and the second is a Japanese female (of any hair color and height).

(*a*) Provide a hand trace of the CANDIDATE-ELIMINATION algorithm learning from the above training examples and hypothesis language. In particular, show the specific and general boundaries of the version space after it has processed the first training example, then the second training example, etc.

(*b*) How many distinct hypotheses from the given hypothesis space are consistent with the following single positive training example?

$$+ \quad \langle \langle male\ black\ short\ Portuguese \rangle \langle female\ blonde\ tall\ Indian \rangle \rangle$$

(*c*) Assume the learner has encountered only the positive example from part (*b*), and that it is now allowed to query the trainer by generating any instance and asking the trainer to classify it. Give a specific sequence of queries that assures the learner will converge to the single correct hypothesis, whatever it may be (assuming that the target concept is describable within the given hypothesis language). Give the shortest sequence of queries you can find. How does the length of this sequence relate to your answer to question (*b*)?

(*d*) Note that this hypothesis language cannot express all concepts that can be defined over the instances (i.e., we can define sets of positive and negative examples for which there is no corresponding describable hypothesis). If we were to enrich the language so that it *could* express all concepts that can be defined over the instance language, then how would your answer to (*c*) change?

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 hypothesis is, ((male, brown, ?, ?) (female, black, short, ?))

And

The general hypothesis 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 $\log_2 256 = 8$ 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,

((female, 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 $\log_2 2^N = N$ 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).

12. Let $\omega_{max}(\mathbf{x})$ be the state of nature for which $P(\omega_{max}|\mathbf{x}) \geq P(\omega_i|\mathbf{x})$ for all i, $i = 1, ..., c$.

(a) Show that $P(\omega_{max}|\mathbf{x}) \geq 1/c$.

(b) Show that for the minimum-error-rate decision rule the average probability of error is given by

$$P(\text{error}) = 1 - \int P(\omega_{max}|\mathbf{x})p(\mathbf{x})\, d\mathbf{x}.$$

Solution:

(a)

We know, summation of all probability is 1,

$\sum_{i=1}^{c} P(\omega_i \mid X) = 1$

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

$P(\omega_i \mid X) = P(\omega_j \mid X)$

Then we can also write,

$P(\omega_i \mid X) = P(\omega_j \mid X) = 1/c.$

So the maximum probability will be also 1/c if all the probability is equal,

$P(\omega_{max} \mid X) = 1/c$

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

$P(\omega_{max} \mid X) > 1/c.$

So, applying both cases, we can say that

$P(\omega_{max} \mid X) >= 1/c.$

(b)

We know that,

According to the minimum-error-rate decision rule, we will accept the maximum probability as a correct probability. We also know that the summation of all probability is 1.

So,

the average probability of error = 1 – probability of correct (accepted probability).

or,

$P(\text{error}) = 1 - \int P(\omega_{max} \mid X)\, p(x)\, dx$

Question 13.

13. In many pattern classification problems one has the option either to assign the pattern to one of c classes, or to *reject* it as being unrecognizable. If the cost for rejects is not too high, rejection may be a desirable action. Let

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \qquad i, j = 1, ..., c \\ \lambda_r & i = c+1 \\ \lambda_s & \text{otherwise,} \end{cases}$$

where λ_r is the loss incurred for choosing the $(c+1)$th action, rejection, and λ_s is the loss incurred for making a substitution error. Show that the minimum risk is obtained if we decide ω_i if $P(\omega_i|\mathbf{x}) \geq P(\omega_j|\mathbf{x})$ for all j and if $P(\omega_i|\mathbf{x}) \geq 1 - \lambda_r/\lambda_s$, and reject otherwise. What happens if $\lambda_r = 0$? What happens if $\lambda_r > \lambda_s$?

Solution:

For i = 1,...,c,

$R(\alpha_i \mid x) = \sum_{j=1}^{c} \lambda(\alpha i|\omega j)P(\omega j\,|x)$

$= \lambda_s \sum_{j=1, j\neq i}^{c} P(\omega j|x)$

$= \lambda_s\,[1 - P(\omega i \mid x)]$.

For i = c + 1,

$R(\alpha_{c+1} \mid x) = \lambda_r$

Therefore, the minimum risk is achieved if we decide,

$R(\alpha_i \mid x) \leq R(\alpha_{c+1} \mid x)$

Or,

$P(\omega_i \mid x) \geq 1 - \lambda_r / \lambda_s$, and reject otherwise.

So if, $\lambda_r = 0$, we always reject.

and, $\lambda_r > \lambda_s$, we will never reject