BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY
COMPUTING FACULTY
SCHOOL OF RESEARCH AND POSTGRADUATE STUDIES

MASTER OF SCIENCE DEGREE IN SOFTWARE ENGINEERING
MACHINE LEARNING

Individual Assignment-I

By:
Sisay Negash (BDU1500286)

Submitted to:
Gebeyehu B. (Dr. of Eng) Associate Professor

May 2, 2023
Bahir Dar, Ethiopia

Based on the table 2.1 answer the following questions.

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cloud	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

1. Give the sequence of S and G boundary sets computed by the CANDIDATE-ELIMINATION algorithms if it is given the sequence of training examples from the above table. Although the final version space will be the same regardless of the sequence of examples (why?), the set S and G computed at intermediate stages will depend on this sequence. Can you come up with ideas for ordering the training examples to minimize the sum of the sizes of these intermediate S and G sets for the H used in the EnjoySport example?

Answer

To answer the first part of the question, the sequence of s and g boundary sets computed by the candidate-elimination algorithm given the table in above is as follows:
Assuming that we are using the version space algorithm, the sequence of s and g boundary sets computed by the candidate-elimination algorithm given the sequence of training examples from the above table would be as follows:

Initially, the version space consists of the most general hypothesis G, which includes all possible instances of the input attributes, and the most specific hypothesis S, which includes only the null hypothesis. At each step, the algorithm updates the boundaries based on the training examples:

1. After processing the **first example** (Sunny, Warm, Normal, Strong, Warm, Same, Yes), the hypothesis **G** is narrowed to: (?, Sunny, ?, ?, ?, ?, ?) and the hypothesis **S** remains the same: (Sunny, Warm, Normal, Strong, Warm, Same, Yes).
2. After processing the **second example** (Sunny, Warm, High, Strong, Warm, Same, Yes), the hypothesis **G** is narrowed to: (?, Sunny, Warm, ?, ?, ?, ?) and the hypothesis **S** remains the same: (Sunny, Warm, Normal, Strong, Warm, Same, Yes).
3. After processing the **third example** (Rainy, Cold, High, Strong, Warm, Change, No), the hypothesis **G** is narrowed to: (?, Sunny, Warm, ?, ?, ?, ?) and the hypothesis **S** is updated to: (?, ?, ?, ?, ?, ?, No).
4. After processing the **fourth example** (Sunny, Warm, High, Strong, Cool, Change, Yes), the hypothesis **G** is narrowed to: (?, Sunny, Warm, ?, ?, ?, ?) and the hypothesis **S** is updated to: (Sunny, Warm, ?, ?, ?, ?, ?).
5. After processing the **fifth example** (Overcast, Warm, High, Strong, Cool, Change, Yes), the hypothesis **G** is narrowed to: (?, Warm, ?, ?, ?, ?, ?) and the hypothesis **S** is updated to: (Sunny, Warm, ?, ?, ?, ?, ?).

6. After processing the **sixth example** (Rainy, Cold, Normal, Strong, Cool, Change, No), the hypothesis **G** is narrowed to: (?, Warm, ?, ?, ?, ?, ?) and the hypothesis **S** is updated to: (?, Warm, ?, ?, ?, ?, No).

Finally, the algorithm outputs the version space, which is the set of all hypotheses that are consistent with the training examples: [(Sunny, Warm, ?, ?, ?, ?, ?), (?, Warm, ?, ?, ?, ?, No)]

Yes, the final version space will be the same regardless of the sequence of examples because the set of hypotheses always reduces as we add more examples until it reaches the final version space, which is the intersection of the consistent hypotheses.

To minimize the sum of the sizes of intermediate s and g sets for the h used in the enjoysport example, we can follow these ideas for ordering the training examples:

1. We can start with the most general hypothesis and gradually specialize it by adding examples. This way, we can minimize the size of the g set as we already have a general hypothesis that covers most examples.

2. We can prioritize examples that are more informative and can help eliminate a larger number of hypotheses from the version space. For example, if an example distinguishes between two hypotheses, we can prioritize it over an example that only eliminates one hypothesis.

3. We can also prioritize examples that are ambiguous and can help us disambiguate between multiple hypotheses. This way, we can eliminate more hypotheses with a single example and minimize the size of the g set.

Overall, the key is to choose examples that are most informative and can help eliminate more hypotheses from the version space, starting from the most general hypothesis and gradually specializing it by adding more examples.

Note: -Let us consider the following information to answer question number two.

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).

Training Examples

1	+	< < male brown tall US > < female black short US > >
2	+	< < male brown short French > < female black short US > >
3	-	< < female brown tall German > < female black short Indian > >
4	+	< < male brown tall Irish > < female brown short Irish > >

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
< < male ? tall ? > < female ? ? Japanese > > 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).

2. 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.

Answer:

Initial State: -
S0: { < < Ø Ø Ø Ø > < Ø Ø Ø Ø > > }
G0:{ < < ? ? ? ? > < ? ? ? ? > > }
Training Example #1
S1:{ < < male brown tall US > < female black short US > >}
G1:{ < < ? ? ? ? > < ? ? ? ? > >}
Training Example #2
S2:{ < < male brown ? ? > < female black short US > >}
G2:{ < < ? ? ? ? > < ? ? ? ? > >}
Training Example #3
S3:{ < < male brown ? ? > < female black short US > > }
G3:{ < < male ? ? ? > < ? ? ? ? > > < < ? ? ? ? > < ? ? ? US > >}
Training Example #4
S4:{ < < male brown ? ? > < female ? short ? > >}
G4:{ < < male ? ? ? > < ? ? ? ? > > }

A. How many distinct hypotheses from the given hypothesis space are consistent with the following single positive example?

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

   Answer:

Each hypothesis consistent with the above example can have either the specified value seen above or "?" for each attribute. That makes 2 per attribute. There are 8 attributes ,2^8=256.

B. Assume the learner has encountered only positive examples from part (a), 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 (a)?

   Answer:

The shortest sequence of queries that assures the learner will converge to the single correct hypothesis consists of 7 queries:

1. < < male black tall us > < female brown short japanese > > (since this is the only possible remaining hypothesis with a male black short person from Portugal)
2. < < male brown tall us > < female ? ? japanese > > (since we know the female person is Japanese, but we don't know her hair color or height)

3. < < male brown tall irish > < female ? ? japanese > > (since we know the male person is tall and brown-haired, and the only remaining option is Irish)
4. < < male brown tall irish > < female ? tall japanese > > (since we know the female person is tall and Japanese, but we don't know her hair color)
5. < < male brown tall irish > < female ? ? japanese > > (since we don't know the hair color of the female person, but we know she is tall and Japanese)
6. < < male brown tall irish > < female brown ? japanese > > (since we know the female person is Japanese and the only remaining option for her hair color is brown)
7. < < male brown tall irish > < female brown short japanese > > (since the only remaining option for the height of the female person is short)

The length of this sequence of queries is equal to the number of attributes in a hypothesis (8), which is also the same as the answer to part (a) since we are assuming the learner has only encountered positive examples. This means that the number of distinct hypotheses consistent with a single positive example is equal to 2 raised to the power of the number of attributes (2^8).

C. 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 languages so that it could express all concepts that can be defined over the instances of languages, then how would your answer to (b) change?

Answer:

The answer to part (b) would change if we were to enrich the language so that it could express all concepts that can be defined over the instances of languages because there would be more hypotheses to take into account, potentially lengthening the sequence of queries required for convergence. In other words, we would have to verify every conceivable combination of values across all possible values in the instance space, rather than merely evaluating the first two possible values for attributes in the hypotheses. There would undoubtedly be a lot more labor involved.