Investigation into the inner workings of Concept Learning and Decision Trees

Adriaan Louw (53031377)

April 2, 2018

Contents

1 Abstract **1**

2 Method **1**

3 Results **1**
3.1 Concept Learning 1
3.1.1 What is Concept Learning? 1
3.1.2 General-to-specific ordering of hypotheses 2
3.1.3 The FIND-S algorithm 2
3.2 Decision Trees 3

4 Conclusion **3**

1 Abstract

2 Method

3 Results

3.1 Concept Learning

3.1.1 What is Concept Learning?

We can describe Concept Learning in terms of trying to teach a machine to classify animals as either dogs or not.

We assuming we classify leaves by the following parameters

- Is the animal large or small? (Size)
- Is it brown or black? (Colour)
- Is the tail long or short? (Tail length)
- Does it have 2 or 4 legs?
- Does it bark?

We can then describe animals according to these parameters $\{size, colour, taillength, nooflegs, doesitbark$ This is called our hypothesis space.

Then a particular animal can be described as $\{small, brown, short, 4, no\}$ for example. For our dataset we have a selection of animals and whether that animal is a dog or not.

The concept that we wish the machine to learn is which values of each parameter defines the Concept of a do. In our case that would be $\{?, ?, ?, 4, yes\}$ where ? means any value.

Formally we can say:

- We have X which is all the instances in the hypothesis space
- We also have $c(x)$ which is the function that returns whether an animal is a dog or not
- Training data $D = \{\langle x, c(x)\rangle : x \in X, c(x) \in \{0, 1\}\}$

(Chandola, 2018)

3.1.2 General-to-specific ordering of hypotheses

From the above example, the most general hypothesis would be $\{?, ?, ?, ?, ?\}$. This means any animal would satisfy the hypothesis. The most specific hypothesis is $\{\emptyset, \emptyset, \emptyset, \emptyset, \emptyset\}$.

We can sort hypotheses based on whether they are more general or specific than other hypotheses.

Consider the following hypotheses: $h_1 = \{?, ?, ?, ?, yes\}$ and $h_2 = \{?, brown, ?, 4, yes\}$. h_1 is more general than h_2. In other words it less specific on what the parameters have to be to satisfy the hypothesis.

(Riedmiller, 2009)

3.1.3 The FIND-S algorithm

The find S-Algorithm is to find the "maximally specific hypothesis". This means it is the most specific hypothesis that can be found that represents the concept.

The algorithm works as follows. We start with a empty hypothesis $h = \emptyset$. Then we iterate through our data set (see Equation 1). if the

Assuming we have a the following dataset :

$$\begin{aligned} D &= \sum_{i=1}^{n} \{x_i, c(x_i)\} \\ &= \{\{\{small, brown, short, 4, no\}, 0\}, \\ &\quad \{\{small, brown, long, 4, yes\}, 1\}, \\ &\quad \{\{large, black, short, 2, no\}, 0\}, \\ &\quad \{\{large, black, long, 4, yes\}, 1\}\} \end{aligned} \tag{1}$$

(Chandola, 2018)
(Cardie, 2012)

3.2 Decision Trees

4 Conclusion

References

Cardie, C. (2012). Inductive Learning or Concept Learning. Retrieved from https://www.cs.cornell.edu/courses/cs4740/2012sp/lectures/ml-basics-ai-lecture-4pp.pdf

Chandola, V. (2018). *Introduction to machine learning.* Retrieved from https://www.cse.buffalo.edu// chandola/teaching/machinelearningdocs/parta2-handout.pdf

Riedmiller, M. (2009). *Concept Learning.* Retrieved from http://ml.informatik.uni-freiburg.de/former/_media/documents/teaching/ss09/ml/version.pdf