A Syntactic Justification for Occam's razor

John Woodward
Computer Science department
University of Nottingham Ningbo China
John.Woodward@nottingham.edu.cn

Andrew Evans
Computer Science department
University of Nottingham Ningbo China
Andy.Evans@nottingham.edu.cn

Paul Dempster
Computer Science department
University of Nottingham Ningbo China
Paul.Dempster@nottingham.edu.cn

ABSTRACT

Informally, Occam's razor states, "*Given two hypotheses which equally agree with the observed data, choose the simpler*", and has become a central guiding heuristic in the empirical sciences and in particular machine learning. We criticize previous arguments for the validity of Occam's razor.

The nature of hypotheses spaces is explored and we observe a correlation between the complexity of a concept yielded by a hypothesis and the frequency with which it is represented when the hypothesis space is uniformly sampled. We argue that there is not a single best hypothesis but a set of hypotheses which give rise to the same predictions (i.e. the hypotheses are semantically equivalent), whereas Occam's razor suggests there is a *single* best hypothesis. We prefer one set of hypotheses over another set because it is the larger set (and therefore the most probable) and the larger set happens to contain the simplest consistent hypothesis. This gives the appearance that simpler hypotheses generalize better. Thus, the contribution of this paper is the justification of Occam's razor by a simple counting argument.

INTRODUCTION

Occam's razor has been adopted by the machine learning community and has been taken to mean *"The simplest explanation is best"* (Cover et. al. [1] page 1). It has been argued that simpler explanations are more probable to give better predictions on unobserved data. Kearns et. al. [2] (page 32) state that Occam's razor has become a central doctrine of scientific methodology.

Why is a simpler hypothesis more likely to generalize to unseen data? While Occam's razor may feel intuitively satisfying, it is worthy of further investigation. We may have two hypotheses that agree with all the observations but we cannot dismiss either as being 'wrong' if they agree with the observed data. Both hypotheses are equally good accounts of the data.

Webb [6] states "*Several attempts have been made to provide theoretical support for the principle of Occam's razor in the machine learning context. However, these amount to no more than proofs that there are few simple hypotheses and that the probability that one of any such small selection of hypotheses will fit the data is low*".

Based on Langdon's work on hypothesis spaces [3], we make the observation that there is a correlation between the complexity of a function and the number of instances of programs representing the given function. Therefore, we make the central assumption that simple functions are represented in more ways than complex functions, and therefore occur with higher frequency.

A set of hypotheses that make the same predictions are equally valid accounts of what is observed. We base our justification of Occam's razor on a uniform distribution over the set of hypotheses. In doing so, we are constructing an argument based on syntax, rather than an argument based on semantics. We believe this to be a plausible approach as it is the underlying syntax which is responsible for the semantics we observe.

The intuition behind our argument is as follows. There are more hypotheses that correspond to the simplest hypothesis (i.e. produce the same function), and fewer hypotheses that correspond to more complex hypotheses. Therefore we prefer the set corresponding to the simplest hypothesis purely because this set is larger.

There are Fewer Simpler Hypotheses

Mitchell [4] (page 65), Russell et. al. [5] (page 535) and Cover et. al. [1] (page 161) argue there are fewer simpler hypotheses than more complex ones so is less likely a simpler hypothesis coincidentally fits the data. While this is true, it then raises the question that, while there are also many complex hypotheses that fit the data but will fail to generalize, there are also many complex hypotheses that do fit subsequent data. It still does not explain why we should choose a shorter hypothesis in preference to another one. We should not arbitrarily discard more complex hypotheses which are consistent with the data.

Formally Stating Occam's Razor

We can state preference as a probability where $p(f)$ is the probability of the function f is correct. Thus a preference for one function over another can be stated as

$$p(f_1) > p(f_2)$$

We denote the complexity of a function as $c(f)$, and can then say that one function is more complex than another as the statement

$$c(f_1) < c(f_2)$$

i.e. f_1 is less complex than f_2. We can combine these two statements to arrive at a formal statement of Occam's razor;

$$p(f_1) > p(f_2) \leftrightarrow c(f_1) < c(f_2)$$

In words, f_1 is preferred (i.e. more probable) to f_2, if and only if f_1 is simpler (i.e. less complex) than f_2.

DEFINITIONS

A program is an algorithm which implements a function. A function is a mapping from one set (the domain) to another set (the range). Programs correspond to syntactic objects and functions correspond to semantic objects. A hypothesis corresponds to a program and predictions correspond to a function (i.e. the semantic interpretation of the syntactic object). Primitive functions are a set of atomic functions used to construct programs.

The size of a program is defined as the total number of bits it contains. The complexity of a function is the size of the smallest description (i.e. program) which expresses it, given the primitive functions we have at our disposal. A function, a semantic entity, is associated with a complexity while a program, a syntactic entity, is associated with a size (i.e. we talk about the size of a syntax parse tree and the complexity of the function that the parse tree expresses).

A hypothesis space consists of all the instances of a given type of representation. As a definition of hypothesis space, we take that given in Mitchell [4] (chapter 1): "*A hypothesis space is the set of all instances of some representation (with some size limit imposed)".* Concept space is the set of concepts represented by the hypotheses in the hypothesis space. For example, a hypothesis space of Java programs maps to a concept space of functions, and this mapping is done via the Java Virtual Machine i.e. $I(p)=f$, where p is a Java program, f is a function and I is the Java Virtual Machine which interprets the program p as the function f. It follows that, $I(p_i)=I(p_j)$ implies the programs p_i and p_j compute the same function and are *semantically equivalent*, while $p_i \neq p_j$ implies the programs are *syntactically different*. A hypothesis space is a set of syntactic entities (e.g. programs) and a concept space is a set of semantic entities (e.g. functions).

We define an equivalence class over a hypothesis space. All hypotheses mapping to the same concept belong to the same equivalence class. This divides the hypothesis space into sets of semantically equivalent hypotheses. For example the Java/C programs {x=x+1;},{x=1+x;},{x++;}, and {++x;} all belong to the same class which could be called *increment*.

During the process of induction, we are eliminating functions not consistent with the data, and therefore discarding the equivalence classes (i.e. sets of programs) which these functions correspond to. This leaves us with a set of equivalence classes, all of which are consistent with the data. We are going to argue that we choose the largest equivalence class of this set, simply because it is the largest, and therefore the most probable.

PROOF OF OCCAM'S RAZOR

We begin by defining our notation. P is the hypothesis space. $|P|$ is the size of the space. F is the concept space. $|F|$ is the size of the space. If two programs p_i and p_j map to the same function ($I(p_i)=f=I(p_j)$), they belong to the same equivalence class (i.e. $p_i \in [p_j] \leftrightarrow I(p_i)=I(p_j)$). The notation $[p_i]$ denotes the equivalence class which contains the program p_i (i.e. given $I(p_i)=I(p_j)$,then $[p_i]=[p_j]$). The size of equivalence class $[p_i]$ is $|[p_i]|$.

We make two assumptions. The first assumption is that we uniformly sample the hypothesis space, and therefore the probability of sampling a given program is $1/|P|$. The second assumption is that there are fewer hypotheses that represent complex functions: $|[p_1]|>|[p_2]| \leftrightarrow c(f_1)<c(f_2)$, where $I(p_1)=f_1$ and $I(p_2)=f_2$. Note that $|[p_1]|/|P| = p(f_1)$.

We begin the proof by a statement of the second assumption;

$$|[p_1]|>|[p_2]| \leftrightarrow c(f_1)< c(f_2)$$

Dividing the left hand side by $/P/$,

$$|[p_1]|/|P|>|[p_2]|/|P| \leftrightarrow c(f_1)< c(f_2)$$

As $|[p_1]|/|P| = p(I(p_1)) = p(f_1)$, we can rewrite this as

$$p(f_1)>p(f_2) \leftrightarrow c(f_1)< c(f_2)$$

This is a mathematical statement of Occam's razor. In other words, we take the probability distribution $p(f)$ to be the frequency with which the functions are represented in the hypothesis space (i.e. the size of the equivalence class). The assumption is $|[p_1]| >|[p_2]| \leftrightarrow c(f_1) < c(f_2)$, where $p(I(p_1))=p(f_1)$, which we have not proved but is reasonable based on empirical work (Langdon [3]). We have not explicitly assumed more complex functions are less likely, but rather, the size of equivalence classes are larger if they contain simpler programs.

DISCUSSION

Occam's razor states that a simpler hypothesis is more likely to generalize to unseen data than a more complex hypothesis. We agree that this appears to be true; however we argue that the underlying reason is that the function corresponding to the simpler description, consistent with the observed data, is represented more frequently in the hypothesis space and this is the reason it should be chosen. We restate Occam's razor:

"The most probable explanation is the one that is most frequently represented in the hypothesis space".

Why are some functions represented more frequently than others in a given hypothesis space? There are a number of reasons for this. The primitive functions may contain functions which are: commutative, associative, distributive or invertible. If the set of primitive functions contain functions which have any of these properties, then there will in general be more than one way that a given function can be represented.

PHYSICAL NATURE OF COMPUTATION

Occam's Razor should say something about the physical world, if it is to be applied to induction problems arising from the real world. The reasons why some functions are more frequent than other is that some functions may be commutative, associative, distributive or invertible. These properties are present in many physical laws we see, for example in conservation laws, it is only the total amount of the quantity which is of interest, not the order in which it arrived (i.e. the summation of energy is symmetric). Many physical systems display the other properties too (e.g. electrical circuts).

REFERENCES

[1] Thomas M. Cover and Joy A. Thomas. Elements of information theory. John Wiley and Sons 1991

[2] Michael J. Kearns and Umesh V. Vazirani. An introduction to computational learning theory. MIT Press, 1994.

[3] William B. Langdon. Scaling of program fitness spaces. Evolutionary Computation, 7(4):399-428,1999.

[4] Tom M. Mitchell. Machine Learning. McGraw-Hill 1997.

[5] S. Russell and P. Norvig. Artificial Intelligence: A modern approach. Prentice Hall, 1995.

[6] G. I. Webb. Generality is more significant than complexity: Toward an alternative to occam's razor. In 7th Australian Joint Conference on Artificial Intelligence – Artificial Intelligence: Sowing the Seeds for the Future, 60-67, Singapore, 1994, World Scientific