Equation discovery as vanishing ideal basis problem

Boštjan Gec[1,2], Ljupčo Todorovski[1,2,3], Sašo Džeroski[1,2]

[1] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
[2] Jozef Stefan International Postgraduate School, Ljubljana, Slovenia
[3] Faculty of Mathematics and Physics, University of Ljubljana, 1000 Ljubljana, Slovenia

We present the machine learning task, central to the field of so called Equation discovery, and show its strong connection to the basic notion of algebraic geometry, namely the correspondence between affine variety and vanishing ideal. More specifically we suggest that the problem of calculating the Groebner basis of vanishing ideal is related and helpful for the task of equation discovery. While algorithms derived from Buchberger's algorithm have been previously utilized to discern systems of ordinary differential equations, it is our understanding that, within the realm of equation discovery, the segment pertaining to algebraic equations (inclusive of the so-called exact equations) has not yet been explored by such algorithms. We present two use cases, to illustrate the potential of this approach. In one, we take a look at the classical problem of equation discovery, where discovery of algebrac equations is the field chemistry or more specifically material science is aimed. In the other, we focus on an application of equation discovery that is closer to the mathematician's point of view to display a potential relevance and use of such approaches in their daily research work.

For this, we take a course into the realm of integer sequences with the goal to discover or reconstruct already known exact (recursive) equations such as the one defined by Fibonacci out of the given sequence terms. As a side note, we point at the potential of the use of equation discovery techniques as well as machine learning in general in the field of mathematics which has recently seen a lot of traction including in the field of theoretical algebraic geometry.

Equation discovery or symbolic regression is a field of machine learning, concerned with the task of computationally discovering the equation that best describes the data. More precisely, the goal is to find a mathematical expression that best fits the data, and is as simple as possible. Details of the task are the following. Usually, the input data is given in the form of a table $A \in \mathbb{R}^{m \times n}$ (also called data set), where each row A_i represents a data point (or example in the data set), and each column A^j holds values of a corresponding feature x_j at a given data point. The goal is to find a mathematical equation

$$E(x_1, ..., x_n) = 0$$

involving some of the features x_j (equivalently its corresponding expression E) of any form that minimizes the two criteria:

- the so called error-of-fit (e.g. root-mean-square-error (RMSE) of the expression against the data set),
- the complexity of the expression (e.g. number of nodes in its abstract syntax tree).

The rationale behind the task definition is the Occam's razor principle (connected also to the minimum description length principle), which states that the simplest explanation is usually the best.

That is, aim is not just to predict the data as precise as possible but also to extract some meaningful knowledge from it. This is also connected to the notions of explainability and causality that are geting more and more attention in the machine learning community lately, where understanding of the underlieing patterns is a priority.

Note, that if the second criterion is ignored, we obtain the classical machine learning task of regression. Although, the way machine learning algorithms usually work is that they assume a certain equation form (e.g. linear, polynomial, neural network) and then optimize the parameters of the equation to minimize the error-of-fit. This makes task of symbolic regression seen more as a search of the form of the equation, rather than the parameters. Its intention is to automate the scientific discovery process, where the scientist would manually search for the form of the equation that best describes the data, such as in days of Kepler, Newton, or Maxwell. In third and fourth paragraph of present abstract we will elaborate on this principle motivation and describe the results of equation discovery field in action in the field of chemistry and mathematics. Below we will present the connection between the equation discovery and the algebraic geometry.

Algebraic geometry is on the other side a branch of mathematics that studies solutions of sets of polynomial equations, known as algebraic varieties. In this field, one of the key concepts is the ideal, which is a special subset of a ring, an algebraic structure that extends the familiar concept of integers.

The variety X may be infinite, representing an infinite set of solutions to a system of polynomial equations. Similarly, an ideal associated with this variety may also be infinite, encompassing an infinite set of polynomials that vanish on X.

However, the beauty of algebraic geometry lies in its ability to simplify these seemingly complex structures. Despite the potential infiniteness of an ideal, it can be represented by a finite set of generators. These generators form a smaller set of polynomials whose ideal is the same as the original, effectively simplifying the infinite ideal into a more manageable, finite form.

This process of simplification is a fundamental aspect of algebraic geometry, allowing mathematicians to study and understand the intricate relationships between ideals and varieties, and explore the rich and fascinating landscape of algebraic structures.

To be a bit more precise while keeping all the necessary mathematical details ([**?**]) aside, we are going to define the vanishing ideal $I(X)$ associated to the subset X in the so called affine space $K^n = K \times \cdots \times K$ where K is an algebraically closed field, e.g. $\mathbb{C}$. Vanishing ideal is defined as the set of all (multivariate) polynomials $f \in \mathbb{K}[X_1, ..., X_n]$ that equal zero (i.e. vanish) in every point $(x_1, ..., x_n)$ of the given $\mathrm{X} \subset K^n$. I.e.,

$$I(X) = \{f \in \mathbb{K}[X_1, ..., X_n] \mid f(x_1, ..., x_n) = 0 \text{ for all } (x_1, ..., x_n) \in X\} \ .$$

As a side note, the set X is called the variety of the ideal $I(X)$ if it includes all zeros of common zeros of all elements of $I(X)$. The connection between the equation discovery and the algebraic geometry is the following. As is evident from the definition above, for each element of the ideal $I(X)$ the equation

$$f(x_1, ..., x_n) = 0$$

holds for all points in X. Therefore, if set X is considered as the input data set, and if only the polynomial expressions are considered, the problem of equation discovery can then be clearly seen as a problem of finding the simplest element of the corresponding vanishing ideal. Although equation discovery is not limited only to polynomials, nothing prevents us from employing algebraic geometry tools for our task if no additional background knowledge about the data is given.

The process of simplifying the vanishing ideal as mentioned above, amounts to finding a finite set of generators that form a basis for the ideal and hold certain conditions. Example of this kind of basis is the so called Groebner basis. Its additional conditions defined by the so called monomial order in some sense require lowest degrees of the polynomials in the basis. Calculation of Groebner basis is a well known problem in algebraic geometry, and is solved by the so called Buchberger's algorithm. Since the Groebner basis is unique (up to a choice of a monomial order), it is a natural choice for the solution to the problem of equation discovery.

Gröbner basis computation can be seen as a multivariate, non-linear generalization of both Euclid's algorithm for computing polynomial greatest common divisors, and Gaussian elimination for linear systems. It is a fundamental tool in computer algebra, and has applications in many areas of mathematics and computer science, including algebraic geometry, cryptography, and automated theorem proving.

Because of its exact nature, Buchberger's algorithm is not appropriate for the numerical data, where any noise in the data can lead to overfitting of meaningless high degree terms apearing in the polynomials to the data. Recognizing the potential power to extract geometric information from data, there is a growing body of research focused on developing approximate versions of the algorithm to handle numerical data. One of the pioneering ideas in this direction, was to look for polynomials f that "almost" vanish on the given data, i.e. $f(x_1, ..., x_n) < \varepsilon$ for a chosen threshold ε. The resulting approach ([**?**]) of so called *approximate vanishing ideal* (AVI) has seen variety of applications in the field of machine learning, one example being feature construction akin to principal component analysis, called *vanishing components analysis* ([**?**]) published at ICML which is one of the three primary high impact machine learning conferences.

In rich repertoaire of improvements and modifications of the avi algorithm including the so called monomial agnostic variational ideal (MAVI) ([**?**]), their applications were utilized also in the modeling of differential equations as well as in the system identification task which can be seen as a equation discovry task where differential equations are the target equations. In papers ... they used one of the

vanishing ideal approaches to discover differential equations but to best of our knowledge, MAVI was not applied for finding algebraic (non-differential) equations.

One of the advantages of using MAVI approach over other state-of-the-art symbolic regression methods is not only the ability to capture geometry of the data in the algebraic geometry sense, but also the ability to discover equations in implicit form. To the best of our knowledge there was developed only one tool for discovering equations that is able to tackle implicit equations and that is the SINDy-PI ([**?**]). Method SINDy-PI aims to discover differential equations, but it can be used also for the implicit algebraic equations. The reason behind rareness of the tools that can discover implicit equations is that the symbolic regression task is usually defined in explicit form described in equation **??** in the footnote **??**. If typical symbolic regression tool would be used with the implicit equation being assumed, the tool would automatically find the obvious trivial solution $\mathrm{E} = 0$.

Although the task of equation discovery originally concerns numerical measurement data of some physical phenomena, it can be applied also to noise-free, i.e. exact data such as that describing mathematical objects. In this case, approximative version or modification of algorithm is redundant. Therefore, for exact data, the original Buchberger algorithm should actually perform better than its approximative versions. Nonetheless, we chose integer sequences' data as our testbed for trying out MAVI and run some preliminary computational experiments on it. This is due our previous experience with this specific data set.

We used the data that we have obtained in our previous studies from The online encyclopedia of integer sequences (OEIS) [**?**]. First 200 available terms of selected sequences were stored into a `csv` file and put on our publicly accessible repository [**?**]. Here we chose the set of 164 so called "core" sequences denoted by OEIS maintainers as the most important and representative ones. We loosely compare mavi performance against experiments from our previous studies where we used Diofantos and SINDy methods instead. We emphasize the unfair comparison to the benefit of the MAVI method due to the comparative analysis methodology that we followed for the previous studies. That is, in those experiments we did not change some of the parameters (e.g. the number of sequence elements considered) of the MAVI parameters. Although on the other side, comparison is unfair also to the MAVI method due to our unfamiliarity with the method. For example, we did not tune the parameters such as scale factor α

The results of the preliminary study are promising as we can see in table **??**.

Table 1. Preliminary comparison of the performance of MAVI against *Diofantos* and SINDy on the *core* data set. Results are grouped by categories. Note, that the comparison is unfair, as discussed in the paper.

Seq. category	#Sequences	MAVI*	*Diofantos*	SINDy
Trivial	4	4	4	4
Simple	66	25	32	31
Other	94	0	0	0
Σ	164	29	36	35

We can see that MAVI performs poorly against other methods, but we are confident that with the proper tuning of the parameters and especially with the use of exact version of the algorithm, we can achieve comparable results if not even better.

The value of development of exact symbolic regression methods lies also in the potential it can bring to the field of mathematics. Machine learning in general has seen a lot of traction in the field of mathematics including the field of theoretical algebraic geometry.

In [**?**], for example, researchers pioneered the use of multi-layer perceptron (a small artificial neural network) to classify whether a Calabi-Yau variety is elliptically fibered or not. Similarly, in their subsequent paper [**?**], they successfully tackled regression task of predicting Hodge numbers associated with Calabi-Yau varieties with high precision. In particular, the authors intriguingly suggested the existence of a formula for the hodge number $h^{3,1}$, inspired by the success of their machine-learning approach. Founding such a formula seems like a task taylored for the exact equation discovery methods. In addition to the aforementioned studies, numerous recent successful attempts to utilize machine learning in making novel discoveries in the field of mathematics include, but are not limited to, studies [**?**], [**?**], [**?**], and [**?**], among others.

Apart from computational experiments on integer sequences data, we plan for the poster also to conduct computational experiments in the field of material science. There we are going to apply the MAVI method aiming at the discovery of the algebraic equations from chemical data to display the potential of the method in the field of material science. Such discovery could make a breakthrough in the mode of operation of a specific modern mesurement device and make it more widely applicable.
[4]

[**?**]
[**?**]

[4] Trivial solution $E := 0$ obviously satisfies the equality E = 0 with the minimal complexity. But since it does not give us any new information, we are not interested in it. Actually, usually equation of the form

$$x_j = E(x_1, ..., x_n) \tag{1}$$

$x_j = E(x_1, ..., x_n)$ is considered, where x_j is any of the features and is excluded from the variables appearing on the right hand side. This is equivalent to finding the equation of the form $E = 0$ with additional constrain of x_j combining linearly in E.