Artificial Intelligence in Analytical Spectroscopy, Part I: Basic Concepts and Discussion

February 1, 2023
Jerome Workman, Jr.
Howard Mark

Publication *Article*

Spectroscopy
February 2023
Volume **38** Issue **02**
Pages: 13–22
https://doi.org/10.56530/spectroscopy.og4284z8

Columns | Column: Chemometrics in Spectroscopy

Artificial intelligence (AI), and its subfield machine learning (ML), are major buzzwords in today's technology world. In a two-part series, let's begin to take a look under the hood and behind the scenes to see what AI is and its applicability to analytical chemistry and spectroscopy for future discussion and elaboration. What are the benefits and limitations, as well as the praises and detractions, of AI? How does AI relate to chemometrics?

As with any modeling technique, it might serve us well to remember Robert A. Heinlein's reference: "There ain't no such thing as a free lunch (TANSTAAFL)" (1), and supercalibratestatisticexpedientalgorithmadnauseous, a play-on word inspired by the wisdom of Mary Poppins (2), reminding us about the hype of new algorithms. The term was coined to mock the then-current practice of trying all-possible combinations of available algorithms and test statistics to automate the process of developing calibration models for near-infrared (NIR) analysis. It took a long time, but eventually, the NIR community recognized the futility of that methodology. Remember, if you are going to make mistakes, automation allows you to make them more quickly and repetitively. One must be cautioned when using powerful and

predictive model based on the information they are presented. As Mark Twain once said, "There are three kinds of lies: lies, damn lies, and statistics" (3–5). And as Rasmus Bro, a leader in the field of chemometrics, has said:

"As most people are aware, there is currently a hype on machine learning and artificial intelligence (AI), and that is all fine and good. Part of the hype is that these methods are sometimes oversold, and so then people are relieved that now they don't have to think about things, saying, 'I don't have to think about experimental design, sampling error, analytical quality, etc.' And when that happens, those projects typically fail, because these new and excellent methods don't actually replace the responsibility for you to actually know what you are doing" (6).

With all these precautionary tales in mind, we now move ahead with our initial overview discussion of AI and chemometrics.

Basically, AI and machine learning (ML), as they are used today, consist of applying signal processing and multiple layer neural networks to model a function or set of functions that has been loosely associated with the intelligent activity of living organisms. The same challenges for building successful models may be present in all modeling activities, including chemometrics; these may include sparse or limited data, noisy data, unknown limits to the potential data and predicted values, a lack of basic understanding of the nature of the data, and so forth. In chemometrics, it is generally accepted that a chemist or chemical engineer is deriving (or at least interpreting) a quantitative, qualitative, or classification model for a chemical system, and therefore has domain knowledge of the nature of the data and the expected model prediction results. *Domain knowledge*, or fundamental knowledge of the system one is working with, is essential to avoid major errors, using computer modeling. That is why, if one is using AI to build a self-driving automobile, it would be very handy to have expert drivers involved during the modeling process. Without domain knowledge, one will undoubtedly see major errors in relying on models generated without such knowledge used as input information. These caveats also apply to modeling weather (where one asteroid or erupting volcano changes everything), the future economy (with the potential for wars and famines), and other unbounded or somewhat unknown domains.

For an "official" AI response, we asked the question, "How are machine learning and chemometrics related?" to the program Chat.OpenAI.com (7), an AI program in development. The AI answer is transcribed below. Note that as the AI program "learns," its responses will change over time.

"Machine learning and chemometrics are both fields that involve the use of statistical and computational methods to analyze and interpret data. However, they are applied in different contexts

and for different purposes. Machine learning is a subfield of artificial intelligence that focuses on the development of algorithms and models that can learn from and make predictions or decisions based on data. These algorithms and models are trained on a set of labeled data, which consists of input data and corresponding output data, and are then used to make predictions or decisions about new, unseen data. Machine learning methods are widely used in a variety of applications, including image and speech recognition, natural language processing, and predictive modeling.

Chemometrics, on the other hand, is the application of statistical and mathematical methods to the analysis of chemical data. It is used to extract meaningful information and knowledge from chemical data sets, such as those obtained from spectroscopic or chromatographic measurements. Chemometrics methods are used to analyze and interpret chemical data in order to gain insights into the underlying chemical processes and to make predictions about the properties and behavior of chemical substances.

Both machine learning and chemometrics involve the use of statistical and computational techniques to analyze and interpret data. However, while machine learning is generally applied to a wide range of data types and applications, chemometrics is specifically focused on the analysis of chemical data."

What is Intelligence?

As with many technical matters, it is useful to begin by discussing the various concepts involved with the subject of interest, and compose a list of (and a set of definitions for) those terms. If one is to develop AI technology and ML algorithms, it is helpful to know what *intelligence* and *learning* are. These concepts seem self-evident or intuitive to sentient beings; however, clear definitions of terms are helpful for classification and understanding. There have been many attempts to precisely define *intelligence* and *learning*. Let's see what we can find.

A definition of *intelligence* from Merriam-Webster includes the following:

1) "...(T)he ability to learn or understand or to deal with new or trying situations: Reason; also: the skilled use of reason; and 2): the ability to apply knowledge to manipulate one's environment or to think abstractly as measured by objective criteria (such as tests)" (8).

The definition of intelligence from Oxford Learner's Dictionaries is

"...(T)he ability to learn, understand, and think in a logical way about things; the ability to do this well...." (9).

What is AI?

Definitions of AI

A definition of AI from the Oxford English Dictionary is as follows:

"(T)he capacity of computers or other machines to exhibit or simulate intelligent behavior; the field of study con- cerned with this. Abbreviated AI" (10).

The often-referred to Wikipedia source has a more detailed definition:

"AI is intelligence demonstrated by machines, as opposed to the natural intelligence displayed by animals including humans...."

Wikipedia notes that older definitions of AI referred to machines that mimic human cognitive skills, such as learning and problem solving. A more modern definition if AI uses phrases such as "machines that act rationally," which somewhat muddies the waters in terms of a precise definition (11). We might summarize the various ideas about a definition of AI, for our purposes, as:

"(A)n automated technology using computers and algorithms that is capable of simulating learning, thought, creativity, rationality, and analysis; and that is able to perform routine or complex tasks that have been historically accomplished by humans."

Machine Learning as a Subfield of AI

ML may then be defined as a subfield of AI where computers (machines) are programmed in such a way that they are able to self-iterate and optimize problem solving for modeling data in an automated way, using computational algorithms specifically designed for such purposes. Let us be frank here: the perceived intelligence or learning is a computer simulation using automation and optimization routines of computer algorithms —no more and no less. Once programmed with appropriate algorithms, the computer (machine) is capable of simulating problem solving and learning functions that are normally attributable to humans.

What is the Foundation of AI?

The foundation of learning AI is to understand the variety of ML (deep learning) algorithms, and then learn how to apply these ML algorithms for various applications. Next, one may design a user interface to interact with the AI programming. The main subfields involved include using the various models for defining an AI problem, learning how to design AI

programs, deciding how humans will interact with the AI programs, and learning how to modify and teach an AI program. One may add here that it is also important that experts evaluate or check the output of an AI program to test its legitimacy.

Algorithms Used for AI

It may surprise those analytical chemists completely unfamiliar with AI that many of the chemometrics (multivariate) algorithms they are familiar with are also classified as AI algorithms—these are algorithms that are routinely taught and used by AI programmers. We refer our readers to the following articles for a head start on this discussion (12–14). This previous series of articles provides definitions and a set of references for signal preprocessing, component analysis, quantitative (calibration) methods, qualitative (classification) methods, and programming platforms.

Table I shows a comparison of the chemometrics algorithms and the AI and ML algorithms as derived by the authors. Table II is a comparison table generated by an AI program (Chat.OpenAI.com) that compares AI and ML algorithms with chemometrics algorithms using its own AI terminology.

Figure 1 shows a basic flow chart of commonality between AI approaches and chemometric modeling. Note that any modeling system is only capable of accurately modeling the data it has been shown or data that follow a logical framework, such as game theory. Within this limited framework, modeling techniques are able to interpolate within its known boundaries and generally produce an accurate result, or at least, a result with a bounded estimate of error. If the modeling system has not seen the data, or if there is not a rational or known framework for potential results, then the results are merely an automated or computerized guess.

FIGURE 1: Flow chart of commonality between model optimization in AI approaches and chemometric modeling. A finished model must include all possible data variables to be a closed system (interpolation) or be forced to predict outside the modeling data space with potential for greater error (failure to predict accurately) if in an open system (extrapolation). The red rectangle illustrates the "never-ending" process of collecting more data when modeling applications try to predict "open systems." The iterative process of modeling, testing, and adding additional data is illustrated. Note that the algorithms and modeling steps may also be varied with more modeling iterations.

The concept of an open or closed model is helpful at this point, because we may imagine what could happen when we model a constrained data set and when the nature of the entire potential data set is not well understood. As Figure 2 illustrates, the limitation of models using powerful fitting algorithms is the "completeness" of the data used and the assumptions made about the nature of the data. Modeling "uncharted" data in science can be, and often is, problematic and unpredictable. Model prediction outside of a closed system is based on hypothesized assumptions about the nature of the data. If the nature of the data and the assumptions about the data are unknown or not clearly understood, then the ability of any modeling system to predict a result accurately is also unknown. Therefore, caution when using extrapolation in a modeling system is wise.

FIGURE 2: Illustration of closed and open model data showing commonality between AI approaches and chemometric modeling. Data shown only in three dimensions (D1, D2, D3). A finished model must include all possible data variables to be closed (interpolation) or forced to predict outside the modeling data space with potential for greater error if open (extrapolation). Note that a model based on

fitting to the black dots would not necessarily accurately predict either the green dots or red dots. The limitation of models using powerful fitting algorithms is the "completeness" of the data used and the assumptions made about the nature of the data. Modeling "uncharted" data in science can, and often is, problematic and unpredictable. Model prediction outside of a closed system is based on hypothesized assumptions about the nature of the data.

One can imagine that game playing is a good example of what a modeling system or AI could easily master, because the universe of possible game moves is well-known and constrained to a finite set of options. With the high speed of computation possible in today's computers, a modeling system is able to compute all possible combinations of moves and describe the best option (or options) available based on probability, separate decision algorithms, or a competitor's responses to its moves. It would seem that a well-written AI or computer system would be the master of game playing and associated tasks. A next level of complexity may arise in self-driving automobiles, but this activity is still fairly well-defined under normal circumstances. That being said, there are occasionally extreme circumstances, such as weather, unusual sky or light reflections, sudden objects appearing on the roadway, multiple car collisions, and so forth, that would be extremely difficult to model and negotiate.

AI computers would also seem to be able to master inventory, logistics, traffic control, basic war games, and prediction of basic human behaviors—all with relatively constrained circumstances. Even a form of pseudo-creativity would seem to be relatively straightforward with access to the Internet and computerized data processing. For new inventions, a computer system is capable of reviewing the prior patent art and literature, evaluating the known physical laws and basic science regarding the subject, and then "suggesting" potentially novel ideas not yet tried. However, issues of compassion, love, sacrifice, selflessness, and generosity would be extremely difficult to model; these are also factors that weigh in heavily for positive human decisions.

So let us get back to the task at hand, which is to discuss AI for applications in analytical spectroscopy. Potential uses would be for instrument alignment (instrument calibration) and diagnostics, instrument measurement conditions and parameter optimization, experimental design, signal preprocessing, quantitative and qualitative calibration, library

searching, and automated imaging and image enhancement functions.

Steps Used for AI and Chemometrics

There are several steps in developing both AI and chemometric models that are basic and somewhat unchangeable. These would include the following steps, which can be accomplished either manually or in an automated fashion:

1) Define the problem that needs to be solved.

2) Define the type and amount of data required to train the model.

3) Collect the data.

4) Clean or preprocess the data.

5) Select the modeling algorithm(s).

6) Train the algorithm.

7) Test or validate the model.

8) Repeat the process to optimize the model .

How to Get Started in Learning AI

There are three main skill sets that are essential in learning how to work with AI. They are as follows:

1) Learn how to operate basic software tools (such as Python), language, and syntax, and the types of algorithms needed.

2) Learn the basics of machine learning algorithms: regression, k-nearest neighbors (KNN), support vector machine (SVM), artificial neural networks (ANNs), reinforcement learning methods, computer vision tools, and so forth.

3) Learn how to specifically solve the problems you are interested in by researching other examples and those researchers already problem solving in your area of interest.

Conclusion

In Part I of this two-part set of introductory articles on AI, we have attempted to give a very concise and top-level overview of AI, a technology topic that is now taking the world by storm. AI discussions are now ubiquitous across media, politics, business, academia, and especially in science and engineering. There is so much more to be said, but we have here brought an introduction to a topic we may continue to explore through additional articles and social media, such as in the "Analytically Speaking" podcast series, Episode 9, featuring Rasmus Bro as the guest speaker (6). In addition, in Part II of this written series, we plan on exploring actual applications of AI and ML to

different vibrational spectroscopy methods, such as Raman, Fourier transform infrared (FT-IR), NIR, and UV–visible (UV-vis) spectroscopic techniques.

References

(1) Heinlein, R. *The Moon is a Harsh Mistress*; G.P. Putnam's Sons, 1966; pp. 164.

(2) Mark, H. *Spectroscopic Calibration* (John Wiley & Sons, New York, NY, 1991), pp. ix.

(3) Twain, M. *Mark Twain's Autobiography, Volume I* (University of California Press, Berkeley, CA, 2010), p. 228. (Attributed to Twain dictating this to his secretary in Florence, April, 1904, in "Notes on "Innocents Abroad"). http://gutenberg.net.au/ebooks02/0200551h.html (accessed 2023-01-03).

(4) The University of York, *Lies, Damned Lies, and Statistics*. https://www.york.ac.uk/depts/maths/histstat/lies.htm (accessed 2023-01-03).

(5) Velleman, P.F. Truth, Damn Truth, and Statistics. *J. Educ. Stat.*, **2008**, *16* (2). https://doi.org/10.1080/10691898.2008.11889565 (accessed 2023-01-03).

(6) Spectroscopy, Analytically Speaking Podcast, Episode 9, Feb. 1, 2023. https://www.spectroscopyonline.com/analytically-speaking-podcast (accessed 2023-01-03).

(7) OpenAI, Build Next-Gen Apps With OpenAI's Powerful Models. https://openai.com/api/ (accessed 2023-01-03).

(8) Merriam-Webster Dictionary, Intelligence. https://www.merriam-webster.com/dictionary/intelligence#:~:text=1%20%3A%20the%20ability%20to%20learn,noun (accessed 2023-01-03).

(9) Oxford Learners Dictionaries, Intelligence. https://www.oxfordlearnersdictionaries.com/us/definition/english/intelligence (accessed 2023-01-03).

(10) Oxford English Dictionary, Artificial intelligence, n. https://www.oed.com/viewdictionaryentry/Entry/271625 (accessed 2023-01-03).

(11) Wikipedia, Artificial Intelligence. https://en.wikipedia.org/wiki/Artificial_intelligence (accessed 2023-01-03).

(12) Workman, Jr, J.; Mark, H. A Survey of Chemometric Methods Used in Spectroscopy, *Spectroscopy* **2020**, *35* (8), 9–14.

(13) Workman, Jr, J.; Mark, H. Survey of Key Descriptive References for Chemometric Methods Used for Spectroscopy: Part I, *Spectroscopy* **2021**, *36* (6), 15–19.

(14) Workman, Jr, J.; Mark, H. Survey of Key Descriptive References for Chemometric Methods Used for Spectroscopy: Part II, *Spectroscopy* **2021**, *36* (10), 16–19.

Jerome Workman, Jr. *serves on the Editorial Advisory Board of Spectroscopy and is the Senior Technical Editor for LCGC and Spectroscopy. He is also a Certified Core Adjunct Professor at U.S. National University in La Jolla, California. He was formerly the Executive Vice President of Research and Engineering for Unity Scientific and Process Sensors Corporation.*

Howard Mark *serves on the Editorial Advisory Board of Spectroscopy, and runs a consulting service, Mark Electronics, in Suffern, New York. Direct correspondence to:* *SpectroscopyEdit@mmhgroup.com*

Corrections from October 2022

There were some errors in the October 2022 Chemometrics in Spectroscopy column that evaded detection until after the column was published (1). Therefore, we hereby publish these corrections:

- In Table I, the binary value for 8_{10} was given as 10000; this number contains one more zero than the correct amount. The correct expression for 8_{10} in the binary number system is 1000_2.
- In the first text column on page 20, under "Errors in Representation," there were several instances of inserting the incorrect subscript to define the number system the specified number belongs to. For example, in the passage".

. . $2^{-1}_2 = 0.5_{10}$ is already too large and $2^{-2}_2 = 0.25$ is too small...” it is clear that the digit 2 cannot belong to a binary number, since a binary number can only contain the digits 0 and 1, as we stated several times. The correct statement is “. . . $2^{-1}_{10} = 0.5_{10}$ is already too large and $2^{-2}_{10} = 0.25$ is too small.”

The same error occurred several times over the subsequent six lines. Instead of belaboring each occurrence of the error, here we simply present the corrected passage, starting with “for example”:

“For example, $0.011_2 = 2^{-2}_{10} + 2^{-3}_{10} = 0.25_{10} + 0.125_{10} = 0.375_{10}$, which is again too large. We note that $0.0101_2 = 0.3125_{10}$, which is still too large, while $0.01001_2 = 0.28125_{10}$, is again too small. Finally, we are closing in on 0.3_{10}, but clearly it is not that simple.”

- In Table III, the last two entries under the column heading “Truncated Decimal Number” simply repeat the second-to-last entry instead of reflecting the continuing truncation of pi. The correct table entries are 3.141000000000000_{10} and 3.140000000000000_{10}, respectively.

We apologize for these errors.

Reference

(1) Mark, H.; Workman, Jr, J. Decimal Versus Binary Representation of Numbers in Computers. *Spectroscopy* **2022**, *37* (10), 13–20. DOI: https://doi.org/10.56530/spectroscopy.mm1179p4