Austin John Felices

Professor Adam Bowers

MATH 163

05/15/2020

Mathematical History of Machine Learning: Two Familiar Foundational Concepts

INTRODUCTION:

In the 21st century, our world is dominated by various forms of technology. From autonomous cars to computers that can automatically identify us, these types of technologies utilize a continuously growing field alongside its counterpart Artificial Intelligence or AI. This technological field is known as Machine Learning. Machine Learning is "an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed" [1]. In other words, Machine Learning is about training computer systems to have the ability to access, learn and utilize data [1]. However, what we now know of Machine Learning should not be taken for granted because this field took centuries to develop. In fact, the technological field of Machine Learning did not arise if it were not for three concepts from one significantly contributing field: Mathematics. Machine Learning should not be taken for granted because in the world Mathematics, Bayes' Theorem and Least Squares made significant contributions in spearheading Machine Learning and should all be considered and known as the most essential foundations for such a continuously evolving technological field. Without these two concepts from Mathematics, Machine Learning would not be understood as it is today and knowing these concepts are necessary if one is to be involved within such statically growing field of technology.

SECTION I: BAYES' THEOREM

The first Mathematical concept of Machine Learning needed to be known is Bayes' Theorem. Bayes' Theorem, or Bayes' Formula as its other name, is a probability and statistics theorem discovered by British Mathematician Thomas Bayes based on his 1764 paper *Essay towards solving a problem in the doctrine of chances* [2], studied and brought to light by Bayes' friend Richard Price [2] and later developed and accepted by French Mathematician Pierre-Simon Laplace in 1781 via a memoir [2]. The idea of Bayes' Theorem goes as this: You have two events A and B with respective probabilities. Assuming events A and B are related in some setting, you can find the probability of event A given Event B occurred equaling the product of the probability of A and the probability if B given event A occurred divided by the probability of event B. In mathematical terms, Bayes Theorem is:

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)} \text{ [4]}$$

The event such as probability of A given event b occurred is known as a conditional probability [5]. But Bayes' Theorem can also be expanded in another form that goes as follows below but with probability event B occurring given event A occurred and in the setting of events A and B as shown previously and when other events in a situation need to be considered:

$$P(B|A) = \frac{P(B \cap \mathrm{A})}{P(A)} = \frac{P(B)P(A|B)}{\sum_{i=1}^{n} P(B)P(A|B)} \text{ [5]}$$

To better understand the use of this theorem than how it is presented face value, let us take a modern example in the setting of the court case of O.J. Simpson with events of O.J. being guilty and evidence supporting his guilt. Let us also take the fact that "*only 0.1% of the men who physically abuse their wives actually end up murdering them"* [5]. We let E be the event of

evidence O.J. committed the crime, G be the event O.J. is guilty of the crime, and G^C be the compliment of O.J. being guilty. Using Bayes' Theorem we can have as follows below:

$$P(G|E) = \frac{P(E|G)P(G)}{P(E|G)P(G) + P(E|G^c)P(G^c)} \quad [5]$$

The values for each of these probabilities is unknown but through other tools of mathematical statistics, we can eventually perform the calculations of this probability and find an 81% chance O.J. is guilty of the crime [5].

With our understanding of Bayes' Theorem from the example provided, it is appropriate to ask what the purpose of Bayes' Theorem is. In other words, why did Thomas Bayes, Richard Price and Pierre-Simon Laplace develop such a theorem and what does it serve in the world? According to Price, Bayes' Theorem was created in order to prove God exists [6]. Price conveyed that events considered "miracles" are not possible and claimed that Bayes Theorem proved these types of events wrong. However, when Price studied a million observations on ocean tides and calculated "that there is a 50% chance the true probability of the tide not coming in one day is somewhere between 1 in 600,000 and 1 in 3 million" [6], he contradicted his claim and discovered that a "miracle" chance is not possible to disregard based on "a large number of negative observations" [6]. Price did not mean God does not exist but rather he demonstrated from his study that Bayes' Theorem has not enough evidence to prove God's existence. Thus, Bayes' Theorem's original purpose does not hold merit.

But the theorem should not be ignored for it has become a powerful mathematical tool within probability and statistics through its many uses in applications such as the example previously mentioned.

Thus, it is appropriate now to address how Bayes' Theorem plays a modern role in Machine Learning. An appropriate example to demonstrate Bayes' Theorem in Machine Learning is through the Machine Learning topic of classification. Classification is a "predictive modeling problem that involves assigning a label to a given input data sample" [7]. We should think of the basic mathematical thought that X implies Y as another way of explaining classification. How Bayes' Theorem can play a role in classification is all categories or "classes" [7] can follow the theorem as follows:

$$P(class|data) = \frac{\big(P(data|class) * P(class)\big)}{P(data)}$$ [7]

P(class|data) is as read as class probability given certain data. Using this formula originating from Bayes' Theorem we are able to calculate the probability of each class in a situation and whichever class has the highest probability is selected and input data is assigned there [7]. However, the performance of Bayes' Theorem within classification in practice is tedious and finding P(class|data) is not doable unless a number of class combination values from the data is large and evident so we can estimate the probability distribution from these combination values [7]. In addition, Bayes' Theorem becomes "intractable" [7] as the more variables are added into a situation with classification. Thus, other techniques beyond the scope of Bayes' Theorem are applied in classification problems. However, the idea and intuition behind Bayes' Theorem still stands and, therefore, should not be disregarded for its idea here of finding the probability of a class given data is foundational in classification. The significance of Bayes' Theorem should not be ignored and at least knowing the idea mathematically can support us in understanding how classification in Machine Learning works overall.

SECTION II: LEAST SQUARES

The second Mathematical concept of Machine Learning needed to be known is Least Squares. Least Squares is a mathematical concept of both Linear Algebra and Statistics discovered by Mathematician Adrien-Marie Legendre and, sadly too credited in discovery, Carl Friedrich Gauss [8]. According to Gauss, translated through Stewart, the idea behind Least Squares is that "if several quantities depending on the same unknown have been determined by inexact observations, we can recover the unknown either from one of the observations or from any of an infinite number of combinations of the observations" with subjected error that will be "less error in some combinations than in others" [8, p. 31]. What Gauss is conveying is that if we are attempting to predict the respective values of a number of variables that depend on the same unknown variable, then we can find the unknown variable through one of the number of those dependent variables or through a combination of a number of those dependent variables with error that is inevitable yet less with some certain combination of a number of those dependent variables. To be even simpler, we should think backwards with our experience of X and Y where instead we think if we have the values of Y found already, then we can find X with minimal error as long as we can perform this through a number of combinations of Y values. This is indeed technical for us to grasp by Gauss's words alone and even as we attempt to interpret his words so the best way to show what Gauss is conveying is through one example from Linear Algebra since Gauss's least squares is a broad concept within mathematics [8].

An example demonstrating Gauss's explanation of least squares is through the Linear Algebra concept of least squares regression. The idea of this concept is we are given a set of points on a graph and we have to create the equation of the line from this points. The steps to

create the line equation are not too technical but let us assume with five points we created the following line equation:

$$y = 1.518x + 0.305 \; where \; y \; is \; a \; predicted \; y - value$$

with the graph

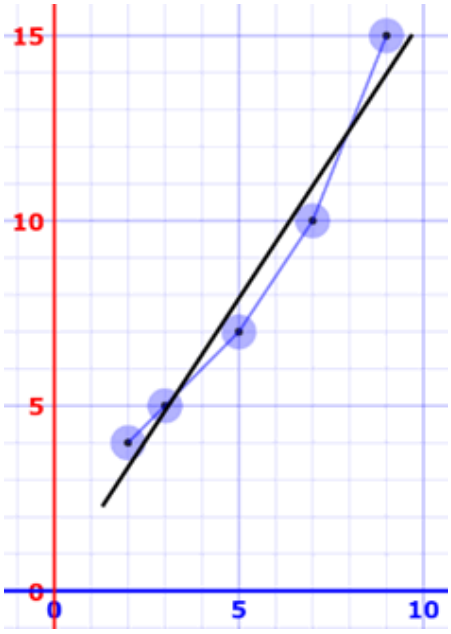

Fig. 1 "Least Squares Regression" of the graph $y = 1.518x + 0.305$ [9]

The points on the graph shown are the exact coordinates and if we plug in an x value into the line equation, for instance x = 3, we get 4.86 as the y predicted output but the actual point at x = 3 is 5. This just means our estimation of x =3 through the line equation was close with error, the difference between the exact point and the point on the line, being small. So in other words, based on what Gauss conveyed on least squares, we see in this example how we were able to predict a y value that was close to the actual y value by way of creating a line equation created from five points and that our y is a predict y in the line equation. Overall, this example should assist us in now grasping what it was Gauss conveyed about least squares.

With our understanding our least squares, we ask now what the purpose of least squares? In other words, why did both Legendre and Gauss create this mathematical concept? Well, they created this mathematical concept to initially study comets and their orbits around our planet. They wanted to understand the behavior of comets based on measurements, which were not exact, of these comets' past locations [10]. Today, modern astronomers use the same tool of least squares to perform calculations on current orbits of comets [9, p. 189]. In totality, the study of the stars is what drove both Legendre and Gauss to create least squares, thus, implying the first application of least squares in the world.

Thus, it is now appropriate to address how least squares today are involved in Machine Learning. In Machine Learning, least squares can be used in the same way as the example mentioned previously with least squares regression. In Machine Learning, this is called linear regression. Linear Regression in machine learning means just as we were predicting y from x through the line equation as we did in the previous example, linear regression in machine learning can perform the same task. In other words, linear regression can also predict y from x. One example that can demonstrate this is having data where we want to predict weight from height. We would create a linear equation based on this where y is the predicted weight and x is the height. Then, in machine learning context we prepare our data by first following some rules of thumb such as removing noise, or data that is random; linear assumption, where we assume the input and output variables follow a linear pattern [11]; collinearity removal, where we basically avoid the over-fitting our data with linear regression by removing the most correlated input variables [11]; Gaussian Distributions, where we have our input and output variables follow a Gaussian Distribution, or a bell-shaped type curve, so that we make more reliable predictions [11]; and rescaling of inputs, where if we rescale our input variables under

"standardization or normalization" our linear regression will predictions that are more reliable [11]. Once we follow these rules of thumb, we can then perform the same procedure as we did in the previous example but by analyzing how close the predictive outputs of our inputted data are with the actual outputs of the data. In this case, we can analyze how far off the predictive weight is compared to the actual data with the actual heights. Thus, least squares demonstrates today how we can use predictions to makes decisions for possible future outcomes. However, least squares should not be underestimated because this area of linear regression in machine learning is just one example of this broad mathematical concept in modern application [8]. The broad mathematics of least squares should not be overshadowed because knowing what least squares is gives us an idea of how its application in machine learning works not just in the area of linear regression but also other branches of machine learning. Without least squares, we would not know how it is possible to make decisions based on predictive models created under machine learning.

CONCLUSION:

The mathematics underlying Machine Learning have demonstrated their essence. Without the use of Bayes' Theorem and Least Squares, Machine Learning would not be as robust and powerful for our current timeline. Mathematics was the foundation of Machine Learning and must not be taken for granted or ignored. Machine Learning is not what it is now if it were not for the mathematicians who studied and discovered such tools that would hold much more significance and merit far past their time. Their works for such a technological field not known in their time have started and paved the way centuries ago and should be honored. If they were alive today, they would hold nothing but gratitude that their works emancipated the future of the

world. Therefore, as Machine Learning continues to evolve in the years to come, the foundations that built this technological field should always be taught and not overshadowed. Without mathematics and these two particular concepts, Machine Learning would not have the life that start centuries ago, active and alive as of today and continuing its existence into tomorrow.

Works Cited

[1] “What Is Machine Learning? A Definition.” *What Is Machine Learning? A Definition*, 6 May 2020, expertsystem.com/machine-learning-definition/. Accessed 15 May 2020.

[2] O'Connor, J J, and E F Robertson. “Thomas Bayes.” *Thomas Bayes (1702 - 1761)*, (2004), mathshistory.st-andrews.ac.uk/Biographies/Bayes.html. Accessed 15 May 2020.

[3] O'Connor, J J, and E F Robertson. “Pierre-Simon Laplace.” *Pierre-Simon Laplace (1749 - 1827)*, (2004), mathshistory.st-andrews.ac.uk/Biographies/Laplace.html. Accessed 15 May 2020.

[4] “Proof of Bayes Theorem.” *Proof of Bayes Theorem*, www.hep.upenn.edu/~johnda/Papers/Bayes.pdf. Accessed 15 May 2020

[5] Zheng, Tianyi. “Bayes’ Formula.” *Bayes’ Formula,* pi.math.cornell.edu/~mec/2008-2009/TianyiZheng/Bayes.html. Accessed 15 May 2020.

[6] Kopf, Dan. “The Most Important Formula in Data Science Was First Used to Prove the Existence of God.” *Quartz*, Quartz, 3 July 2018, qz.com/1315731/the-most-important-formula-in-data-science-was-first-used-to-prove-the-existence-of-god/. Accessed 15 May 2020.

[7] Brownlee, Jason. “A Gentle Introduction to Bayes Theorem for Machine Learning.” *Machine Learning Mastery*, 3 Dec. 2019, machinelearningmastery.com/bayes-theorem-for-machine-learning/. Accessed 15 May 2020.

[8] Gauss, Carl Friedrich. “Theoria Combinationis Observationum Erroribus Minimus Obnoxiae; Pars Prior, Pars Posterior, Supplementum.” Translated from Latin by G. W. Stewart. Society for Industrial and Applied Mathematics (1995).

[9] *Least Squares Regression*, www.mathsisfun.com/data/least-squares-regression.html. Accessed 15 May 2020.

[10] “The Discovery of Statistical Regression.” *Priceonomics*, econ.ucsb.edu/~doug/240a/The Discovery of Statistical Regression.htm. Accessed 15 May 2020.

[11] Brownlee, Jason. “Linear Regression for Machine Learning.” *Machine Learning Mastery*, 12 Aug. 2019, machinelearningmastery.com/linear-regression-for-machine-learning/. Accessed 15 May 2020.