ISSN: 2581-4311
DOI: doi.org/10.30967/IJCRSET/Pau-Figuera/194

URF PUBLISHERS
connect with research world

International Journal of Current Research in Science, Engineering & Technology

https://urfpublishers.com/journal/ijcrset

Vol: 8 & Iss: 3

Perspective

The Long Road for Explainability in Machine Learning: A Probabilistic and Algebraic Insight

Pau Figuera*

University of Deusto, Bilbao, Spain

Citation: Figuera P. The Long Road for Explainability in Machine Learning: A Probabilistic and Algebraic Insight. *Int J Cur Res Sci Eng Tech* 2025; 8(3), 366-372. DOI: doi.org/10.30967/IJCRSET/Pau-Figuera/194

Received: 19 July, 2025; **Accepted:** 24 July, 2025; **Published:** 27 July, 2025

***Corresponding author:** Pau Figuera, University of Deusto, Bilbao, Spain

ABSTRACT

In this manuscript, we describe several methods to achieve interpretability in machine learning. Explainability implies the use of well-sound algebraic and probabilistic tools. These depend on the algorithms used in the training phase. It also translates into the graphical capabilities of each method. We present methods based on algebra and probability. Both approaches converge with the well-known nonnegative matrix factorization techniques after transforming observations into probabilities. Those techniques enable the introduction of a probability-based measure that generalizes and unifies multiple methods. We introduce the description of the methods from a historical perspective.

Keywords: Explainability; Artificial intelligence, Machine learning

Introduction

AI (Artificial Intelligence) has been an increasing field in recent years. It uses the methods of ML (Machine Learning). This last discipline is the intersection of computer engineering and statistics[1]. The objective is to execute the tasks as humans do. The methods require learning a task with a metric for its evaluation. The acceptance of these methods by the scientific community for data compression, summarization, classification or prediction has given rise to the appearance of many competing methods. Those methods involve several families of approaches and they do not offer the same results in many cases.

In the basic sciences, the results obtained with mathematical tools are highly reliable. In the case of ML, the discussion becomes complex. The main idea of ML is to select a model (a function model), determine its parameters and evaluate the quality of the conclusions. This last step is the validation phase. The parameters are selected from examples and they constitute the training phase. Depending on whether the examples are previously classified (or labeled) or not, the methods are supervised or unsupervised. The training phase is sensitive to extreme values (outliers), zeros (stability), overlapping variables and variation. A problem shared by many methods is that they may favor the researcher's requirements and abilities[2].

The most classic interpretable methods are related to statistical regression and they are well-sound. Another widespread family of methods is clustering. Clustering methods involve grouping similar observations into homogeneous groups, which constitutes a partition of the set of observations. Then, the estimation of the observation's similarity is done by introducing a metric based on distance. There are also probabilistic methods (or fuzzy methods, although there is a difference between them). These methods do not create a partition in which each observation uniquely belongs to a single group, but they assign a certain degree of probability to each group. The SVM (Support Vector Machines) relies on the idea of kernelization. They use the matrix of scalar products to assess the similarity between observations. This

property is called spatial separability. In practice, separability requires transforming the observation space (input space) into another space, not necessarily of the same dimension (feature space). Another family of methods, constituting a field and gaining popularity, is NN (neural networks). These methods are inspired by the brains of biological organisms and introduce the concept of neurons as connected nodes that simulate brain function. Formally, the connection of nodes grouped into layers constitutes a computational graph. The electrical transmission of signals circulating in biological brains is the activation function. The training phase is weighing the neuron connections. Despite their heuristic justification, some authors claim that they allow for obtaining any mathematical function[3].

As a result, for each family of methods, there is a diaspora of algorithms. Many of them are difficult to understand. In this context, explainability gains interest. Recently, this topic has experienced growing interest. This issue, contextualized in NNs, is presented in[4]. A more general treatment that offers a discussion of the problem of the concept of explainability is[5]. Explainability needs formalisms. In the presence of uncertainty, it is achieved by using analytical tools, algebra, probability theory and statistics. Nonnegative-entry matrices with an appropriate normalization condition support the algebraic structures of probabilistic methods. This algebra unifies algebraic and probabilistic statements, leading to multivariate density estimations.

To illustrate this viewpoint, the place of causality and uncertainty in the formation of qualitative models is briefly described (section *Science Models*). The tools for achieving explainability from algebraic-probability theory constitute the section *Formal Tools*. The *Explainable Machine Learning Landscape* section introduces the primary techniques with explainable outcomes before concluding.

Science Models Evolution

The classic methods for modeling problems are differential equations (DE). With the development of physics, every branch has given rise to its specific differential equations. The set of many of these equations for solving particular problems is known as Mathematical Physics. This method is the backbone of the work of scientists in the 18th and 19th centuries and has extended to many other branches of science. The phenomena modeled in this way are both interpretable and explainable in the sense of the introduction.

The quantitative study of other phenomena linked to heredity and evolution is a debt to the great British mathematician and statistician Karl Pearson. Leaving aside his many fundamental contributions, he modeled uncertainty in terms of probabilities to predict the outcomes of future generations, expounding these results with clarity and instructiveness in the ancient book The Grammar of Science[6].

Any scientific discipline falls within the paradigmatic extremes based on its degree of uncertainty. Between these two positions, there is a gradation depending on whether a greater or lesser degree of uncertainty is assumed. The methods that constitute the foundations of AI can be classified in this way, as shown in **(Figure 1)**.

Formal Tools

The formal tools for explainability are those of multivariate statistics. They are matrix algebra, especially the SVD (Singular Value Decomposition) Theorem, probability theory and some basic analytical tools. It is essential to consider that the probabilistic space data transformation enables the use of probability as a measure of similarity. Additionally, it has geometrical significance.

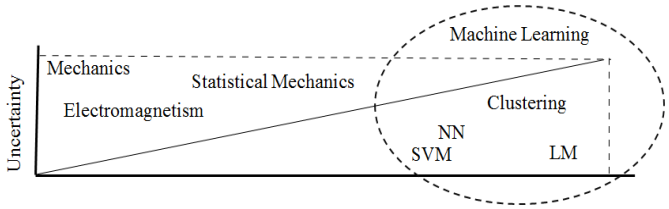

Figure 1: Degree of uncertainty assumed by descriptive models of reality in the sciences. Mechanical models, both classical and relativistic, are modeled by differential equations (in this case, mainly with the help of tensor algebra and differential geometry) and they are deterministic. Electromagnetic theory shares this position, perhaps to a lesser extent, due to the behavior of waves. Statistical mechanics introduces more uncertainty into the underlying hypotheses. The methods used in the learning phase of tasks have a much greater component associated with uncertainty. We omit Quantum Mechanics from the classification due to its complexity and place Optics within electromagnetic theory (they are the propagation of the visible spectrum waves).

Singular value decomposition

One of the most relevant concepts in pure and applied mathematics is the Singular Value Decomposition Theorem. It plays a crucial role in pure and applied mathematics, a field known as Eigenanalysis. From the point of view of applications, it finds a place in almost all branches of Physics, structural engineering and data analysis. For real entries-matrices, it is currently stated as[7]:

$$\mathbf{X} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V} \qquad (1)$$

Where $X \in R^{mxn}$ is the data matrix, $U \in R^{mxk}$, $V \in R^{kxn}$ are orthogonal matrices and

$$\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$$

are the eigenvalues. For k the rank of X, the case k < min (m, n) is the low-rank approximation and k=min (m, n) the full-rank case.

The PCA (Principal Components Analysis) is a direct application of the SVD, after centering and normalizing the columns of X. The trace is the variance of the data, also called inertia. PCA provides graphical representations for the orthogonal projections of observations on the planes formed by the consecutive pairs of columns of the matrix V. In applications, the terms SVD and PCA are sometimes confusing.

Probabilistic latent semantic analysis

PLSA (Probabilistic Latent Semantic Analysis) is an unsupervised learning technique developed for information retrieval purposes, also known as PLSI (Probabilistic Latent Semantic Indexing). The classical reference is *Unsupervised Learning by Probabilistic Latent Semantic Analysis* by Hofmann[8]. PLSA is a probabilistic remake of LSA (Latent Semantic Analysis)[9]. This model crosses two categorical variables to obtain a table of co-occurrences. Arranged as a matrix, the SVD space span is a set of multinomial latent

variables. The adjustment of the probabilities is done with the EM (Expectation-Maximization) algorithm. The estimation of the similarity between categorical variables is a distance.

This model was first formulated in the context of IR (Information retrieval): for a set of di (i=1, …, m) documents, containing wj (j=1, …, n) words of a thesaurus, the pairs are frequencies. Then, exist k (k=1, …, k) latent variables (intuitively, the documents' subjects), arranged as the count matrix N (di. wj) leads to the probabilities:

$$P(d_i, w_j) = \frac{N(d_i, w_j)}{\sum_{ij} N(d_i, w_j)} \qquad (2)$$

Using the Bayes rule, this expression admits the factorizations

$$P(d_i, w_j) = P(d_i) \sum_k P(w_j|z_k) P(z_k|d_i) \qquad (3)$$

And

$$P(d_i, w_j) = \sum_k P(z_k) P(d_i|z_k) P(w_j|z_k) \qquad (4)$$

According to Hofmann, Formulas (3) and (4) are the asymmetric and symmetric formulations, providing the generative models shown in **(Figure 2)**. $P(d_i)$ is the probability of a document (also, observation). $P(z_k)$ are probabilities of the k latent variables, modelized as a uniform distribution. P (·|·) are conditional probabilities.

In his original work, Hofmann pointed out the analogy of the PLSA and the SVD (for the symmetric formulation) by assimilating Formula (3) as:

$$\mathbf{U} = P(d_i|z_k) \qquad (5)$$

$$\sum \square = diag P(z_k) \qquad (6)$$

$$\mathbf{V} = P(w_j|z_k) \qquad (7)$$

On the other hand, the conditional probability

$$P(d_i|w_j) = \frac{N(d_i, w_j)}{\sum_i N(d_i, w_j)} \qquad (8)$$

can be written in matrix form as

$$\tilde{\mathbf{Y}} = \mathbf{XD} \; (s.t. \mathbf{D} = diag\left(\sum_i x_{ij}\right) \qquad (9)$$

Where $\mathbf{X} \sim N(d_i, w_j)$ and $\tilde{\mathbf{Y}} \sim P(d_i|w_j)$, indication the tilde symbol that each sum column is one and accomplishing that

$$\mathbf{Y} = \tilde{\mathbf{Y}}\mathbf{D} \; (s.t. \mathbf{D} = diag(n)) \qquad (10)$$

Several researchers relate the PCA to probabilistic techniques in a more precise way, as is done in the PLSA. Klingenberg provides a geometrical interpretation compared with classical PCA[10]. The dimension problem reduction has a probabilistic significance. An attempt to find the optimal dimension of the PLSA is due in[11]. Since PCA solutions are not unique (they depend on the dimension reduction), it is not guaranteed to extract the latent variables involved in a problem.

Nonnegative matrix factorization

Many authors attribute the introduction of the NMF (Non-negative Matrix Factorization) techniques to Paatero[12]. We prefer to do it to Chen[13]. The work of Chen connects with the early 20th-century findings on the properties of singular values. Currently, the standard formulation is[14]:

$$\mathbf{X} = \mathbf{WH} + \mathbf{E} \qquad (11)$$

for a matrix of non-negative inputs **X**, where matrices $\mathbf{W} \in R_+^{mxk}$ and $\mathbf{H} \in R_+^{mxk}$ is the factorization and $\mathbf{E} \in R_+^{mxn}$, the error approximation matrix, also of non-negative entries.

and generalizing the PLSA to every non-negative real entry's matrix.

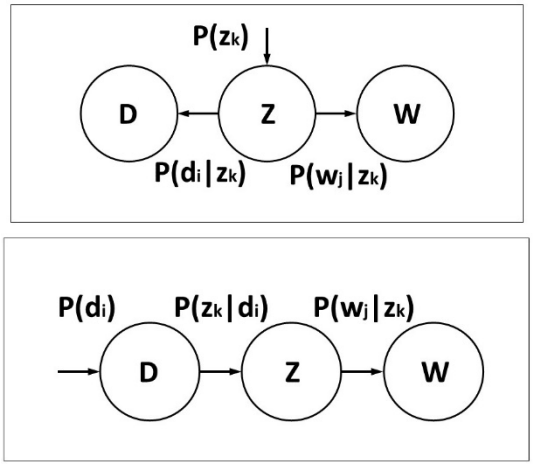

Figure 2: Generative models for the PLSA. Reproduced from[8]: (a) asymmetric formulation. (b) symmetric formulation

There exist three families of algorithms to solve this approximation problem: (i) the gradient methods, (ii) alternating least squares and (iii) the iterative updates methods. From the point of view of explainability, the most interesting are the iterative updates, since they use a loss function related to geometrical or statistical properties. The loss function is a distance or a divergence (distances *d* are maps that satisfy, for vectors **a**, **b** and **c**, the axioms: (i) $d(\mathbf{a},\mathbf{b})=d(\mathbf{b},\mathbf{a})$ (symmetry); (ii) $d(\mathbf{a},\mathbf{b})=0$ if $\mathbf{a}=\mathbf{b}$ (identity); and (iii) $d(\mathbf{a},\mathbf{c})+d(\mathbf{c},\mathbf{b}) \geq d(\mathbf{a},\mathbf{b})$ (triangular inequality). A divergence does not satisfy one of these axioms, usually symmetry and it is more suitable for measuring how densities are similar.

Obtaining matrices **W** and **H** requires selecting the dimension of the factorization space, initializing them **and** an iterative process until a satisfactory fit between **Y** and the product of **W** and **H** is achieved. The dimension of this space is the number of model components.

The NMF applications are numerous, ranging from text mining and text classification to analysis of linguistic knowledge acquisition, labeling, computational biology, clustering algorithms, image classification and speech recognition, among others. This technique has furnished new methods for data analysis for experimental sciences as well as Economics, Social Sciences and a wide range of areas. It is not possible to list it briefly.

Writing conditional probabilities of Formulas (2) and (3) as matrices, a consequence is that any local maximum solution of the PLSA is a solution to the NMF problem[15]. Devarajan[16] unifies various models and provides a rigorous proof of monotonicity for multiplicative updates, generalizing the relationship of

NMF and PLSI within this framework. Studies on the NMF correlation have been conducted by He[17]. Further generalization and establishing more rigorous conditions for the equivalence between PLSA and SVD are studied in[18], stating the equivalence between the SVD and PLSA.

Explainable Machine Learning Landscape

Explainability is achieved by mathematically well-sound methods. Interpretability refers to the ability to associate results with the underlying data. Furthermore, a human operator needs to intuitively understand the results, especially in big data environments and it is achieved through a graphical representation **(Table 1)**.

Table 1: Log-linear models apply the logarithm function for a linear transformation. They are extensively used in binary classification. Multiple regression uses several explanatory variables to introduce a model of the form $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$ where the X_i (i=1,…,p) vectors take their values in the p explanatory variables and the βp are the parameters to be determined. Logistic regression is helpful in binary classification.

Explanatory variable	Response variable			
	Binary	Multinomial	Discrete	Continuous
Binary	Logistic regression.	Log-linear models	Log-linear models.	
Multinomial	Log-linear models.			
Continuous	Logistic regression.			Multiple regression.
Covariates				
Random effects	Mixed models.			

Regression

The main idea is to estimate an expectation function $m(x)=E(Y|X=x)$ for the known values $X=(x1,\ldots, xm)$ or predictors for the random variable Y. It takes the form $Y=\beta_0+\beta_1 X+\epsilon$. A model fits the data for new observations if a loss function of the form f(Y, X) is minimized (usually the Euclidean distance). From the point of view of ML, the learning problem is the estimation of the coefficients βi Regression methods have been successfully used in different fields before the emergence of contexts related to ML. Furthermore, there are many variants depending on the type of data. Table 1 shows the main types of regression. A classical and authoritative introduction to the regression methods is[19].

Clustering

The introduction of clustering techniques is a debt to James MacQueen in 1967[20], with the classical k-means method. The main idea is to assume that there exists a certain number of groups and each observation can be assigned exclusively to a group, creating a partition. Assigning coordinates to the data points representing each group, an iterative algorithm provides the means and the observations are classified according to the distance to this point. It is an unsupervised method.

The fuzzy clustering methods assume a different degree of association with each cluster and the partition is not disjoint. The emergence of fuzzy methods is due to the ability of PCA to solve classification problems. The development is related to the introduction of probability as a measure in the classification problem[21]. These methods are well-sound. Probability, from a mathematical point of view, is an evaluator of the measure. From the point of view of explainability, it represents a milestone.

The use of NMF techniques on a probabilistic basis is introduced by Ding[22]. Those results require the identification of the space span factorization with clusters, converting the factorization problem into a classification problem, which is currently a classical problem. Furthermore, in this case, the introduction of a Bayes classifier leads to a partitional classification (it is equivalent to k-means).

Example 1a

For a data set containing seven documents {d1, …, d7} containing items {a, b, c, d, e, f}, if the matrix X containing the co-occurrences is:

$$X = \begin{array}{c} \\ d1 \\ d2 \\ d3 \\ d4 \\ d5 \\ d6 \\ d7 \end{array} \begin{array}{c} \begin{array}{cccccc} a & b & c & d & e & f \end{array} \\ \begin{bmatrix} 4 & 3 & 4 & 0 & 0 & 0 \\ 5 & 3 & 3 & 0 & 0 & 0 \\ 4 & 3 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 9 & 0 & 0 \\ 0 & 0 & 0 & 8 & 2 & 0 \\ 0 & 0 & 0 & 0 & 3 & 4 \\ 0 & 0 & 0 & 0 & 4 & 3 \end{bmatrix} \end{array}$$

A simple visual inspection suggests that k = 2 or 3 groups. The transformation with relations (9) and (10) provides the matrix:

$$Y = \begin{bmatrix} 0.06 & 0.05 & 0.06 & 0.00 & 0.00 & 0.00 \\ 0.08 & 0.05 & 0.05 & 0.00 & 0.00 & 0.00 \\ 0.06 & 0.05 & 0.05 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.14 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.12 & 0.03 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.05 & 0.06 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.06 & 0.05 \end{bmatrix}$$

Selecting k=3, the matrix **W** and the qualitative matrix $\mathbf{W_d}$ obtained by replacing the row labels in each column are

$$W = \begin{bmatrix} 0.34 & 0.00 & 0.00 \\ 0.34 & 0.00 & 0.00 \\ 0.31 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.48 \\ 0.00 & 0.02 & 0.52 \\ 0.00 & 0.49 & 0.00 \\ 0.00 & 0.49 & 0.00 \end{bmatrix} W_D = \begin{bmatrix} d1 & - & - \\ d2 & - & - \\ d3 & - & - \\ - & - & d4 \\ - & d5 & d5 \\ - & d6 & - \\ - & d7 & - \end{bmatrix}$$

Applying a Bayes classifier ($w_i=\max(w_{ij})$ s.t. $w_{ij}\epsilon \mathbf{W}$) provides a result analogous to k-means:

$$W_{bayes} = \begin{bmatrix} d1 & - & - \\ d2 & - & - \\ d3 & - & - \\ - & - & d4 \\ - & - & d5 \\ - & d6 & - \\ - & d7 & - \end{bmatrix}$$

The number of groups is a critical parameter. In many cases, this is a subjective question and adjusted a posteriori through a process known as clustering validation. This branch of techniques has experienced rapid development in recent years.

In the case of the previous example, it consists of deciding if k=3 is a good choice. Clustering validation is a crucial step to decide if the parametrization is suitable. There exist several

studies focused on using probability for validation. Perhaps, it can be attributed to Har[23], who introduced an chi-square indicator function for observations based on kernelization and the null hypothesis as a classifier. Smyth[21] used likelihood cross-validation to infer information on the number of model components. Differently, a is used to measure how similar two clusters are[24]. More recent works include Olivares[25], which, in the scope of astronomical observations and under the hypothesis of normality and the existence of a correlation, presents an algorithm in which the posterior of the correlation follows a gamma pdf (probability density function). The work of Hamid[26] presents a purely algebraic approach, in which elements are clustered by co-linearity. A review work, centered on the impact and importance of clustering validation in the context of bioinformatics is provided by Ullmann[27]. The NMF matrix factorization is a powerful tool to solve this problem. To select the number of clusters, a gamma pdf represents the credibility of the number of clusters[28].

Example 1b

The next step is to decide if the selection of k=3 is a good choice. Gamma density modeling has the parameters α= 2 and λ= 0.5 and it is shown in the **(Figure 3)**.

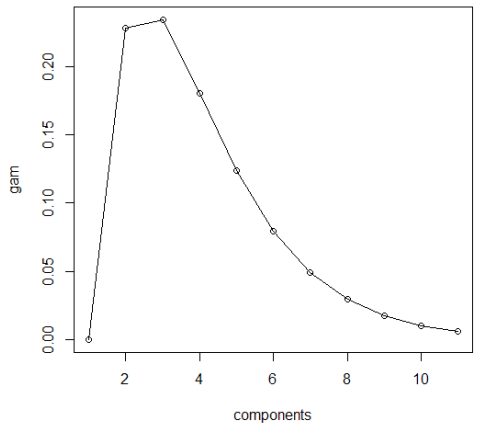

Figure 3: Gamma Model for Clustering Validation. For details on its obtention, we refer the reader to[28]

The NMF applications are numerous, ranging from text mining and text classification to analysis of linguistic knowledge acquisition, labeling, computational biology, clustering algorithms, image classification and speech recognition, among others. This technique has furnished new methods for data analysis in experimental sciences as well as Economics, Social Sciences and a wide range of areas. It is not possible to list it briefly.

Writing conditional probabilities of Formulas (2) and (3) as matrices, a consequence is that any local maximum solution of the PLSA is a solution to the NMF problem[15]. Devarajan[16] unifies various models and provides a rigorous proof of monotonicity for multiplicative updates, generalizing the relationship of NMF and PLSI within this framework. Studies on the NMF correlation have been conducted by He[17]. Further generalization and establishing more rigorous conditions for the equivalence between PLSA and SVD are studied in[18], stating the equivalence between the SVD and PLSA.

Support vector machines

The significance of the kernel is different depending on the mathematical branch. In our context, we refer to the dot product matrices (or Gramm matrices). In this context, kernel methods are a family of supervised tools used for classification.

A widespread use of the kernel matrices is the SVM (Support Vector Machine). A label identifies the instances or observations. The set of similar observations is a region of the space, creating a separation line[29]. This concept gives rise to a classification rule. **(Figure 4)** illustrates this idea.

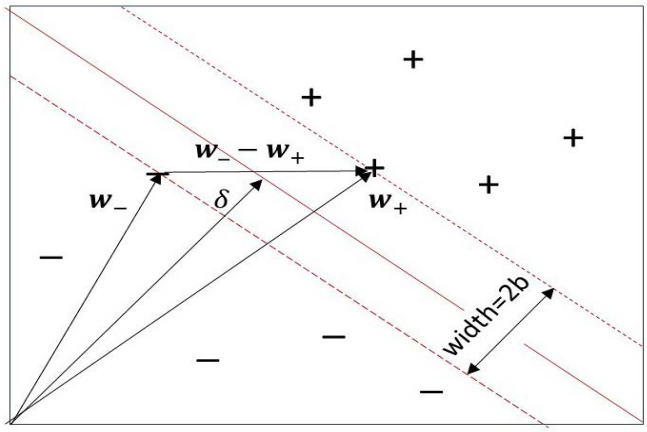

Figure 4: Accuracy of the SVM classification. Because the dot product is a projection, the vectors located in the area above the separation line (observations identified with +) have a dot product greater than 1. The opposite occurs for the vectors in the zone of the points identified with - (their dot product is less than 1). The zone w- - w+=2b is the street, representing the misclassification rate

There exist many types of kernels, but the Fisher kernel, introduced by Jaakkola[30], deserves special attention. It provides a metric for a probabilistic model and a consistent estimator (the density converges in probability to the value of the parameter that generates the distribution) for the posterior (density of the data)[31].

A reformulation of the work of Jaakkola is Hofmann[32]. Based on PLSA, assumes multinomial distributions. Chappelier relaxes this hypothesis, postulating only iid (independent and identically distributed) distributions[33].

More recently, NMF techniques have allowed to obtain kernels[34,35]: under suitable normalization conditions, the obtained matrices are stochastic, allowing us to relate NMF to the Fisher kernel. In this case, the parameters are the product matrices of Formula (11). This statement, under Gaussian assumptions, leads to more understandable and stable classifications[36].

Neuronal networks

NN (Neural Networks) are techniques that aim to utilize biological learning mechanisms. The elements for simulating the behavior of neurons are nodes, which exhibit a relationship between their inputs and outputs called an activation function. Nodes are grouped according to graphs, giving rise to different families of NNs. The training phase involves assigning weights to the connections between nodes.

The choice of the activation function, the number of nodes and their connections is a subject that some authors consider more of an art than well-defined criteria. On the other hand, NN allows for any mathematical function[2].

Among the typologies of NNs, RBMs (Restricted Boltzmann Machines) have their basis in the geometry of information, making them explainable. They are probabilistic models of binary states of nodes and represent latent variables. Furthermore, activation functions can be related to the partition function (in physics, the function that associates particles with the energy states[37,38]).

Conclusion

Explainability in ML is a topic attracting the attention of researchers recently. The computational complexity, in terms of the methods involved, has obscured many of the techniques as well as the results obtained. It can be explained by the predominance of applications and the need for results in research, rather than theoretical and fundamental issues and justified for commercial needs. ML provides powerful tools, giving rise to fields of application that are, in turn, the subject of active research areas: recommender systems, information retrieval, visual and auditory recognition systems and a long list of other fields that are difficult to enumerate. However, these fields reveal several commonalities between the techniques. The solidity of their foundations allows for explainability.

Furthermore, the path we have explained is not exclusive. We have chosen works that use a probabilistic transformation and then apply statistical and/or algebraic techniques due to the more robust properties of the (non-orthogonal) projection of the data into the probability space (the almost sure convergence or probability convergence, at the limit is a Cauchy sequence and therefore also converges in the ordinary sense). We believe this is a path worth pursuing and that it offers significant research opportunities by providing a framework that connects with Information Geometry.

The path outlined is not unique. Some authors focus exclusively on a purely algebraic treatment, relying solely on the properties of the eigenvalues[37]. Other approaches from a strictly probabilistic point of view[38], with implications for the work of[39,40]. The NMF with suitable normalization conditions does not make any hypothesis on the nature of the parameters, with sufficient not minimal parameters W referred to the base H. It provides an algebraic-probabilistic explainable environment and allows the construction of supervised and unsupervised methods for current ML paradigms.

References

1. Mitchell T. The discipline of machine learning. Carnegie Mellon University, School of Computer Sci 2006;9.
2. Aggarwal CC, Chandan KR. Data clustering: Algorithms and applications. Chapman & Hall. CRC Data mining and Knowledge Discovery Series 2014.
3. Aggarwal CC. Neural Networks and Deep Learning. Springer 2023.
4. Heuillet A, Couthouis F, Díaz-Rodríguez N. Explainability in deep reinforcement learning. Knowledge-Based Systems 2021;214:106685.
5. Hickling T, Zenati A, Aouf N, Spencer P. Explainability in deep reinforcement learning: A review into current methods and applications. ACM Computing Surveys 2023;56(5):1-35.
6. Pearson K. The Grammar of Science. Dover Publications 1892.
7. Zhang XD. Matrix Analysis and Applications. Cambridge University Press 2017.
8. Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 2001.
9. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J American society for Information Sci 1990;41(6).
10. Klingenberg, B, Curry J, Dougherty A. Non-negative matrix factorization: Ill-posedness and a geometric algorithm. Pattern Recognition 2009;42(5):918-928.
11. Donghwan K. kfda: Kernel Fisher Discriminant Analysis. R package version 2017.
12. Paatero P, Tapper U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 1994;5(2):111-126.
13. Chen JC. The nonnegative rank factorizations of nonnegative matrices. Linear algebra and its applications 1984;62:207-217.
14. Cichocki A, Zdunek, Phan, AH, Amari S. Non-negative Matrix and Tensor Factorizations. Applications to Exploratory Multi-way Data Analysis and Blind Source Separation. John Willey and Sons Ltd 2009.
15. Gaussier E, Goutte C. Relation between PLSA and NMF and implications. Proceedings 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'05).
16. Devarajan K, Wang G. Ebrahimi N. A unified statistical approach to non-negative matrix factorization and probabilistic latent semantic indexing. Machine Learning 2015.
17. He Y, Wang C, Jiang C. Correlated matrix factorization for recommendation with implicit feedback. IEEE Transactions on Knowledge, Data Eng 2018;31(3):451-464.
18. Figuera P, Bringas PG. On the Probabilistic Latent Semantic Analysis Generalization as the Singular Value Decomposition Probabilistic Image. J Stat Theory App 2020;19:286-296.
19. McCullagh P, Nelder JA. Generalized Linear Models. 2nd Edition, Chapman and Hall 1989.
20. MacQueen J. Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability 1967:281-297.
21. Smyth P. Model selection for probabilistic clustering using cross-validated likelihood. Statistics and computing 2000;10(1):63-72.
22. Ding C, Li T, Peng W. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. Computational Statistics, Data Analysis 2008;52:3913-3927.
23. Har-Even M, Brailovsky V. Probabilistic validation approach for clustering. Pattern Recognition Letters 1995;16(11):1189-1196.
24. Pallis G, Angelis L, Vakali A, Pokorny J. A probabilistic validation algorithm for web users' clusters. 2004 IEEE Int Conf Systems, Man and Cybernetics 2004;5:4129-4134.
25. Olivares J, Sarro LM, Bouy H, et al. Kalkayotl: A cluster distance inference code. Astronomy and Astrophysics 2020;664(A7).
26. Hamid U. Clustering, multicollinearity and singular vectors. Computational Statistics, Data Analysis 2022;173.
27. Ullmann T, Hennig C, Boulesteix AL. Validation of cluster analysis results on validation data: A systematic framework. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2022:1444.
28. Figuera P, Cuzzocrea A, Bringas PG. Probability Density Function for Clustering Validation. Hybrid Artificial Intelligent Systems 2023:133-144.
29. Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995;20.

30. Jaakkola T, Diekhans M, Haussler D. A discriminative framework for detecting remote protein homologies. J Computational biology 2000;7(1-2):95-114.

31. Tsuda K, Akaho S, Kawanabe M, Müller KR. Neural computation 2004;16(1):115-137.

32. Hofmann T. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. Advances in neural information processing systems 2000:914-920.

33. Chappelier JC, Eckard E. Plsi: The true fisher kernel and beyond. Joint European Conf Machine Learning and Knowledge Discovery in Databases 2009:195-210.

34. Zhang D, Zhou ZH, Chen S. Non-negative matrix factorization on kernels. Pacific Rim Int Conf Artificial Intelligence 2006:404-412.

35. Lee H, Cichocki A, Seungjin C. Kernel nonnegative matrix factorization for spectral EEG feature extraction. Neurocomputing 2009;72(13-15):3182-3190.

36. Salazar D, Rios J, Aceros S, López-Vargas O, Valencia C. Kernel Joint Non-Negative Matrix Factorization for Genomic Data. IEEE Access 2021;9:101863-101875.

37. Hopfield JJ. Neural Networks and Physical Systems with Emergent Collective Compuational Abilities. National Academy of the Sciences of the USA 1982;79(8):2554-2558.

38. https://www.netflixprize.com/community/topic_1537.html

39. Jialu L, Han J. Spectral clustering. Data clustering. Chapman and Hall/CRC 2018:177-200.

40. Valiant LG. A Theory of the Learnable. Communications of the ACM 1984;27(11):1134-1142.