Enhancing the Human Experience of Learning with Technology:
New challenges for research into digital, open, distance & networked education
European Distance and E-Learning Network (EDEN) Proceedings
2020 Research Workshop | Lisbon, 21-23 October, 2020
ISSN 2707-2819
doi: 10.38069/edenconf-2020-rw0015

THE USE OF MACHINE LEARNING IN EDUCATIONAL DATASETS

David Menoyo-Ros, Antonio Garcia-Cabot, Eva Garcia-Lopez, Adrian Dominguez,
University of Alcala, Spain

Abstract

The use of machine learning in educational datasets can be very important in predicting and detecting different behaviours related to education, so knowing the most useful techniques that help to effectively analyse these datasets will be very beneficial for any type of machine learning that needs to be performed. This paper explains the typical and common workflow for machine learning with educational datasets (interpreting the term *machine learning* as the algorithms that learn from data automatically), especially focusing on the pre-processing of these datasets, a process that takes place before machine learning algorithms are performed.

Keywords: Machine learning, data science, datasets, education, workflow, pre-processing, outlier detection, imputation of missing values, data transformation, dimensionality reduction.

Introduction

The use of machine learning extends to almost any field one can imagine, such as cybersecurity, computer vision applications, medical analysis, economic predictions, sentiment classification, or even the discovery of hidden patterns in student data. In short, the applicability of machine learning is so high that it is understandable that several techniques and methods were created to approach with guarantees of success any machine learning project.

Therefore, knowing the most basic aspects of different machine learning techniques will be very helpful to accurately and correctly decide the most appropriate workflow for every specific problem, but first, it is crucial to explain some basic foundations on machine learning and the techniques around it.

Definition of machine learning and importance of data science

Data science and machine learning are among those fields that have been generating the most impact in recent years, mainly due to their usefulness in practically any sector. As a result, the vast amount of research being carried out today produces an enormous amount of techniques, concepts, methodologies, good practices and, ultimately, computational tools that, when put together, could overwhelm any engineer. In addition, it is very likely that the concepts of machine learning, deep learning, data science, big data, etc. seem equivalent, and although the differences between any of the previous terms are not entirely clear, there seems to be a certain generalized approach based on the Venn diagram developed by Drew Conway (2011) (Figure 1).

As it can be seen from Figure 1, machine learning uses hacking skills and statistics and mathematics as the only components. In addition, it is interesting to highlight the role of data science, since it brings together all the existing concepts and skills, and therefore, many of the techniques that help improve machine learning performance come from data science rather than machine learning itself, a field that could concentrate machine learning algorithms exclusively.

Leaving Drew Conway's diagram aside, machine learning could be defined in different, more formal ways such as:

> *"Machine Learning is the study of computer algorithms that improve automatically through experience" (Mitchell, 1997)*
>
> *"A machine learning algorithm is an algorithm that is able to learn from data" (Goodfellow et al., 2016)*

Thus, the emphasis placed on performing some intelligent learning behaviour is more formally appreciated, and not so much on optimization and filtering techniques, more typical of data science.

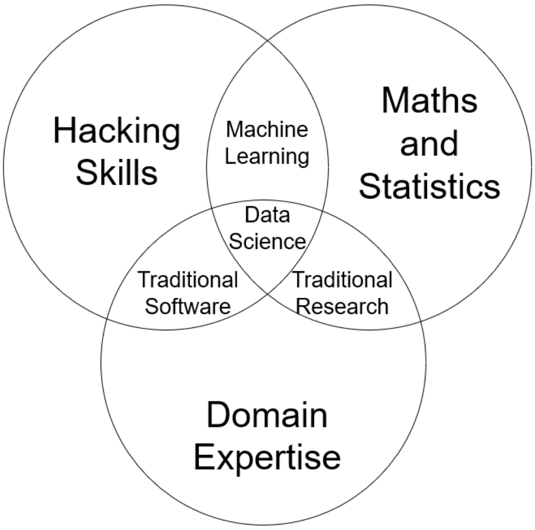

Figure 1. Venn diagram of Drew Conway

Importance of machine learning in education

Until not many years ago, major education-related studies tried to maximize student performance, classify students, decrease absenteeism in university students, etc. with classic strategies of motivational assessment (Ryan & Deci, 2000), gamification (Domínguez et al., 2013) and socialization by developing gamified platforms with which students would find motivation and enthusiasm for learning, by developing attractive practical activities, or even by implementing platforms with social media elements (De-Marcos et al., 2016b; 2016a; 2014) that would positively impact the academic environment.

In short, the interest in improving and understanding the educational field is of considerable interest in the scientific community, and therefore, it was a matter of time before machine learning emerged as a key and fundamental tool in many of the research that is carried out today.

If one looks at the data (Wallace, 2019; Carter, 2019; ODSC Community, 2019), research on the application of machine learning in education has a major boost in countries like U.S. and China, both investing millions of dollars in research and development. Thus, many of the benefits expected in the near future may be:

- Precise grading: Many education professionals aim to unify the evaluation system and eliminate the bias while teachers evaluate their students. A current example is that of 60,000 Chinese students whose essays were evaluated using AI almost as accurately as a real teacher would do (Wiggers, 2018).
- Predicting career paths: It is estimated that only 10-15% of U.S. students know which path to select after graduating from school, therefore, using machine

learning tools to objectively advise each student based on their skills and interests could greatly help them choose the right path.

- More personalization in classrooms: Adapting the learning methodology is vitally important to maximize the potential of each student, especially when it comes to students with some kind of learning disabilities.
- Preventing student dropout in distance learning (Kotsiantis et al., 2003) by implementing a machine learning algorithm that recognizes students with high dropout probability.

The growing importance of this type of research and development requires an operating methodology that maximizes the chances of success when using any type of educational dataset. In this way, the main objective of this paper is to explain, in a conceptual and simple way, the main steps when tackling any educational dataset through machine learning, making special emphasis on the data science phase in charge of filtering, processing, correcting and maximizing the importance of the dataset to be used in the machine learning algorithm.

Essential steps in manipulating educational datasets

When it comes to pre-processing an educational dataset there are many techniques that can be used, and therefore, the data scientist must choose the most appropriate techniques in every case, since each dataset is different and will require different processing.

Among all the available techniques, we will develop the most relevant ones in the following sub-sections.

Imputation of missing values

Empty values are very common in datasets, and if they are not correctly processed, they may make it difficult to create machine learning models, so addressing the problem of empty values in datasets is vitally important.

Generally, when the dataset has many samples and the number of empty values is very small (less than 5%), we can ignore those samples without too many consequences, but when our dataset is scarce, deleting certain samples can generate a great loss of information for our machine learning algorithms, which could be solved with the use of some missing value imputation technique.

A simple way to impute missing values is to calculate the mean of each attribute from the observed values, and impute the result obtained from the previous operation in each of the empty values. This simple technique is called *simple imputation*, but it has the great disadvantage of distorting other properties of the variables, such as the variance.

In other peculiar cases, such as time series data, the method can be sophisticated by just taking a moving window or margin, and then replacing the missing values with the mean of all the existing values in that window. This imputation technique is known as the *moving average method*.

Another interesting way to carry out the imputation of empty values is to impute impossible values, for example negative values in properties that are always positive, to simply carry out a quick test, leaving the imputation for later with some advanced technique. However, this technique is only used for getting a fast analysis that provides the data scientist with some valuable information, and should not be used for models intended for production.

Although the previous simple techniques can get someone out of trouble, much more sophisticated and elaborated algorithms are usually used, and those sophisticated algorithms use complex techniques such as multiple imputation (Rubin, 1978), bootstrapping, PMM (predictive mean matching), etc.

Many of these advanced algorithms are available in the major programming languages used for machine learning. For example, one of the most powerful algorithms in the R programming language is MICE (van Buuren & Groothuis-Oudshoorn, 2010), and its choice to impute empty values in an educational dataset can be very beneficial.

Outlier detection

Outlier detection is the process of finding data whose values move away from normal ones.

Outliers are generated for a variety of reasons, but it is usually due to sensor errors or human errors and manipulations. However, in some cases it is possible that the outlier comes from totally valid natural reasons, and the data scientist in charge will be responsible for interpreting and reasoning the appearance of such outlier, since it is quite possible that some novel and valuable behaviour has been discovered by chance.

Whatever the reason for the appearance of the outlier, its detection is crucial so as not to worsen the performance of the machine learning process, since the presence of several outliers in the learning could have unfavourable consequences for the generalization and accuracy of the learning performed.

The detection of outliers firstly goes on to distinguish two criteria: depending on the context and according to the dimensionality:

- Depending on the context:
  - Global outlier: The outlier differs from all other normal values clearly.

   - Contextual outlier: The outlier differs from all other normal values but within a specific situation. For example, 35 degrees Celsius in winter in Moscow (Russia) is not normal at all, but in summer it would be a perfectly possible temperature; so in winter it would be considered as a contextual outlier, while in summer it would be a common value.
   - Collective outlier: In this case the outlier appears grouped with other outliers of similar value, and therefore can be camouflaged more easily, but its creation mechanisms are still different from those of the normal data and therefore the person in charge of performing this technique will need to have sufficient ability to detect them.
- According to the dimensionality:
   - Univariables: Outliers are detected in each independent variable, taking into account the data of each variable exclusively.
   - Multivariables: The outliers are detected jointly. For example, a pair of variables mark and attendance of student that is atypical, and although mark alone and attendance alone may not be atypical, the detection of the outlier is defined jointly by both variables, and not individually. For instance, a student with the highest mark and the lowest attendance is very strange.

With the above essential concepts in mind, it only remains to use the outlier detection algorithms, which could be classified as follows:

Table 1: Types of outlier detection algorithms

Type	Specific algorithms
According to statistical techniques	Boxplot, standard deviation, standard error
Depending on proximity	KNN (K-Nearest Neighbours)
Depending on density	DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Once the outliers are detected with any of the aforementioned techniques, we can apply three possible solutions:

10. Discard outliers directly, removing them from the dataset.
11. Change the value of the outlier to another value that is not considered an outlier, imputing the most extreme value possible without being considered as an outlier.
12. Delete the outliers and carry out some algorithm to impute empty values.

Data transformation

Data transformation seeks to apply some mathematical function or operation to the original data to obtain new data that fulfils some new characteristic. Generally, machine learning algorithms require some transformation in the data that they are going to process, since if some initial conditions are not met, it is very likely that they will not learn correctly or their performance will be decayed.

We could highlight 4 main types of data transformation:

Table 2: Typical algorithms for detection of outliers

Type	Note
Obvious transformations by domain	It consists of keeping variables or data in a variable in the same unit or scale.
Normalization transformations	It consists of getting the same range of values in all the variables. For example: All variables having minimum value of 0 and maximum value of 1. This type of transformation is extremely important, as the vast majority of machine learning algorithms do not work well if their variables have different ranges.
Standardization transformations	It consists of homogenizing some property on all attributes of the dataset. For example: All variables having mean 0 and standard deviation 1.
Distribution transformations	It consists of transforming an original distribution into another more convenient distribution for machine learning, generally the normal distribution. For this type of transformation a function is used with the main requirement that it has an inverse.

All the aforementioned transformations are important, but the one that can give the most complications is the *distribution transformation*, since there are three issues to keep in mind (McDonald, 2014):

13. The transform function must be invertible in order to retrieve the original value, if needed. For example, the inverse of $\log_2 x$ would be x^2.
14. Not all functions support any input value. For example the transformation with the function $f(x) = \sqrt{x}$ does not admit negative values or zero, therefore it can only be used in variables with all their positive values, unless some previous transformation is carried out to the same variable that guarantees that all the values are positive, such as adding to all the values a constant that makes all values positive.
15. Transformations can be chained, but it transformations made should be justified to the scientific community.

Some of the most well-known transformation functions are:

Table 3: Most Commonly Used Distribution Transformation Functions

Type	Important formulas	Reverse formula	Comments
Logarithmic	$\log_{10} x, \log_e x$	$10^x, e^x$	Used in natural and biological behaviours
Square root	$\sqrt{x}$	x^2	Used for counting
Arcsine	$\arcsin(\sqrt{x})$	$\sin^2(x)$	Used for proportions

Dimensionality reduction

Dimensionality reduction tries to decrease the number of variables required for machine learning, so that the time required for learning can be reduced, as well as computing resources optimized.

There are two approaches to reduce the dimensionality of a dataset:

- The first approach is based on selecting and removing attributes based on their importance level, but without changing the original values of the selected variables. Some simple techniques are:
  - Low variance: Variables with low variance are removed, as they barely provide information to the dataset.
  - High correlation: One of the two variables that are highly correlated is deleted, as the information they transmit is redundant.
- The second approach is based on constructing new variables from the original ones, so that an attempt is made to maximize the information from the first new variable to the last new variable, always reflecting as much information as possible in the first variables, so that fortunately, it takes far fewer new variables to provide nearly the same information as with all the original variables together.
  - PCA (Principal Component Analysis): It is the main algorithm of this approach, and one of the most used in any machine learning operation. As a concrete example, if we start with an original dataset containing 100 variables, the objective will be to construct 100 new variables formed from the 100 original variables but maximizing the information, so that with only the first 60 new variables, the 99% of the variation of the original dataset of 100 variables is included, and therefore, with these 60 new variables we can perform machine learning faster and with less computational costs than if we had to use the original 100 variables.

Conclusions

Throughout this paper, the most important and widespread techniques that help filter, improve and optimize datasets have been developed and explained in a conceptual way, and therefore, many tools have been offered to improve machine learning on gamification datasets.

Finally, it is important to note that the order in which the techniques were explained is totally arbitrary, and the data scientist will be responsible for choosing the techniques to apply and in what order, by taking into account the characteristics of the educational dataset to be processed.

References

van Buuren, S. & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1-68.

Carter, S. (2019, July 18). Roles & Responsibilities of Artificial Intelligence in Education. ODSC [Blog post]. Retrieved September 05, 2020, from https://opendatascience.com/roles-responsibilities-of-artificial-intelligence-in-education/

Conway, D. (2011). Data Science in the US Intelligence Community. *IQT Quarterly*, *2*(4), 24-27.

De-Marcos, L., Domínguez, A., Saenz-de-Navarrete, J., & Pagés, C. (2014). An empirical study comparing gamification and social networking on e-learning. *Computers & education*, *75*, 82-91.

De-Marcos, L., Garcia-Lopez, E., & Garcia-Cabot, A. (2016a). On the effectiveness of game-like and social approaches in learning: Comparing educational gaming, gamification & social networking. *Computers & Education*, *95*, 99-113.

De-Marcos, L., García-López, E., García-Cabot, A., Medina-Merodio, J. A., Domínguez, A., MartínezHerráiz, J. J., & Diez-Folledo, T. (2016b). Social network analysis of a gamified e-learning course: Smallworld phenomenon and network metrics as predictors of academic performance. *Computers in Human Behavior*, *60*, 312-321.

Domínguez, A., Saenz-De-Navarrete, J., De-Marcos, L., FernáNdez-Sanz, L., PagéS, C., & MartínezHerráIz, J. J. (2013). Gamifying learning experiences: Practical implications and outcomes. *Computers & education*, *63*, 380-392.

McDonald, J. H. (2014). Handbook of Biological Statistics. Retrieved September 05, 2020, from http://www.biostathandbook.com/transformation.html

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1). Cambridge: MIT press.

Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2003, September). Preventing student dropout in distance learning using machine learning techniques. *Proceedings of the International conference on knowledge-based and intelligent information and engineering systems,* 267-274. Berlin, Heidelberg: Springer.

Mitchell, T. M. (1997). *Machine learning.* Burr Ridge, IL: McGraw Hill.

ODSC Community. (2019, December 12). Machine Learning for Education: Benefits and Obstacles to Consider in 2020. ODSC [Blog post].Retrieved September 05, 2020, from https://opendatascience.com/machine-learning-for-education-benefits-and-obstacles-to-consider-in-2020/

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. *Proceedings of the survey research methods section of the American Statistical Association,* 20-34. American Statistical Association.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54-67.

Wallace, E. (2019, July 16). Greatest Hurdles in AI Proliferation in Education. ODSC [Blog post]. Retrieved October 1, 2020, from https://opendatascience.com/greatest-hurdles-in-ai-proliferation-in-education/

Wiggers, K. (2018, May 28). Chinese schools are testing AI that grades papers almost as well as teachers. VentureBeat – The Machine, Making sense of AI [Blog post]. Retrieved September 05, 2020, from https://venturebeat.com/2018/05/28/chinese-schools-are-testing-ai-that-grades-papers-almost-as-well-as-teachers/

Acknowledgments

This work is funded under the “GAMILEARN” project (funded by Government of Comunidad de Madrid (Grant CM/JIN/2019-037)). Authors also want to acknowledge support from the University of Alcala.