Experiential learning in Data Science: from the dataset repository to the platform of experiences

Emilio Serrano a,1, Martin Molina a, Daniel Manrique a and Luis Baumela a

aDepartment of Artificial Intelligence, Universidad Politécnica de Madrid, Spain

Abstract. Data science is a revolution that is already changing the way we do business, healthcare, politics, education and innovation. There is a great variety of online courses, masters, degrees, and modules that address the teaching of this interdisciplinary field, where there is a growing demand of professionals. However, data science pedagogy has repeated a number of patterns that can be detrimental to the student. This position paper describes an ongoing educational innovation project for the study of methods, experiences, and tools for experiential learning in data science. In this approach, the student learns through reflection on doing instead of being a recipient of already made content.

Keywords. Active learning, experiential learning, project based learning, data science, deep learning.

1. Introduction

Data Science (DS) is an interdisciplinary field devoted to extract knowledge from data. This discipline is particularly complex with Big Data: large volumes of data that hinder standard Computer Science technologies from storing, processing, and analyzing these vast amounts of information. DS is a revolution that is already changing the way we address business, health, politics, education and innovation [14].

The great diversity of applications and the growing demand of experts in the DS field has made courses, books and manuals in DS proliferate. The standard pedagogical method that we can appreciate in these courses consists of four steps:

1. The explanation of the different machine learning branches (supervised, unsupervised, and by reinforcement).
2. The detail of some learning paradigms under some of these branches; such as decision trees or artificial neural networks.
3. The illustration of these paradigms using toy datasets such as Weather or Iris [21].
4. Assignments with a straightforward application of the ideas previously exposed using some DS framework such as Weka [9] or Caret [10].

1ORCID ID: 0000-0001-7587-0703. E-mail: emilioserra@fi.upm.es.
The existence of different dataset repositories [5,7] on which to build knowledge offers a privileged breeding ground for designing a DS course as a series of experiences in real world problems. Few fields allow students to put themselves in the shoes of profiles as diverse and interesting as: economists, business managers, physicians, biologists, or website administrators. Similarly, few disciplines can offer rewards as attractive to the students as the three million dollar prize for winning a contest in predicting the patients who will be admitted to a US hospital the following year [3]; or the one million dollar prize with which the company Netflix awarded the best predictor of movie ratings [8].

In this position paper, we present an ongoing educational innovation project to develop methods, experiences, and tools for the **experiential learning** (EL) in DS. After introducing the concept of EL in section 2, some shortcomings detected by the authors in the standard pedagogical method for DS are presented in section 3. Then, a framework for EL in DS is presented in section 4. Section 5 offers specific experiences in **Deep Learning** (DL), a popular sub discipline of DS [11]. The requisites of an EL platform are described in section 6. Finally section 7 concludes and gives future works.

2. What is experiential learning?

Aristotle wrote in the Nichomachean Ethics “for the things we have to learn before we can do them, we learn by doing them. Men become builders by building”. Although this ancient quote is commonly used to explain the concept of EL, it can also be misleading. EL is more than just getting learners to “do something”. As Qualters and Wehlburg explain [17]: “unless experiences outside the classroom are brought into the classroom and integrated with the goals and objectives of the discipline theory, students will continue to have amazing outside experiences but will not readily connect them to their in-class learning”.

Therefore, even when the term “experiential learning” is sometimes used to define any training that is interactive, with minimal lecture (and slides) [20], the students reflecting on their product is a fundamental part of EL. Without a careful curriculum involving **structured, reflective** skill building, students may never learn what we hope outside the four walls of the classroom [17].

As the “Association for Experiential Education” [1] claims, to ensure that EL is effective, the learner has to be actively engaged in posing questions, investigating, experimenting, being curious, solving problems, assuming responsibility, being creative, and constructing meaning. The educator and learner may experience success, failure, adventure, risk-taking and uncertainty, because the outcomes of experience cannot totally be predicted. Therefore, EL is an approach that encourages collective and critical reflection as well as individual learning [18]. In summary, acquiring skills requires more than “monkey see, monkey do” [20].

3. Shortcomings in the standard pedagogical method in Data Science

The four-step standard pedagogical method in DS is applied in popular online courses such as “Machine Learning” at the Stanford University [6] or the “Data Science Specialization” at the Johns Hopkins University [2]. The authors have also employed this
hegemonic method in Master courses at Technical University of Madrid, such as “Data Mining”. The following limitations of this method have been observed by the authors:

1. The student has difficulties in selecting relevant information about how learning paradigms work and the information they offer. They often consider them as black boxes where the model built has no relevance and only quality metrics are studied.

2. As a result, the student usually obviates the details and data of the concrete problem. As Witten [21] declares, nothing replaces a good understanding of the data.

3. Also corollary of the first point, the student does not perceive the iterative nature of DS. The construction of different prediction models provides valuable information about the data that must be fed back into subsequent iterations to achieve more valuable results.

4. Creativity in solving problems is considerably restricted because DS is perceived as the application of well-known solutions to well-known problems. Nevertheless, a fundamental component of the data scientist work is the research on new methods and their extension or variation for new and challenging problems.

EL naturally mitigates these tendencies when learning DS because it focuses on problems to be solved instead of on specific methods. In addition, starting with realistic experiences gives students more experience about real-world problems. More importantly, creativity and divergent thinking are encouraged when searching for different solutions to a concrete experience.

4. A framework for experiential learning in Data Science

Much of the development of EL theory in the past 30 years has gained relevance by the work of David A. Kolb [13], where he synthesizes the principles of this learning approach. Many publications, journal articles, and research studies have explored the explanatory power and usefulness of this theory in various disciplines and professional fields [15].

Kolb’s experiential learning style theory is typically represented by a four stage learning cycle. McLeod [16] summarizes these stages in the graph shown in figure 1. According to Kolb [13], effective learning involves progressing through this cycle: having a concrete experience; observation of and reflection on that experience; the formation of abstract concepts (analysis) and generalizations (conclusions); and, testing them by active experimentation, resulting in new experiences (iterations in the cycle). Therefore, every new attempt to address a problem is informed by a cyclical pattern of previous experience, reflection and observation, conceptualization, and experimentation.

In the scope of our educational innovation project, Kolb’s learning cycle is revisited and instantiated for the specific field of DS as a framework to provide learners with significant experiences. Although there are recent and relevant works proposing this kind of experiences for information technologies [19], this research project focuses on EL for DS. Practical examples of using this theoretical framework in the DS field are detailed below for each of its four stages:

1. Concrete Experience. A new experience or situation is encountered, or a reinterpretation of an existing experience is offered. At the beginning of the course and
to justify the machine learning approach, an experience in which it is difficult or impossible to establish algorithms that step by step solve a problem can be facilitated to the learners; for instance, the recognition of faces in images. As the course progresses, the experiences can challenge what has already been learned about DS.

More examples are: the use of unsupervised learning methods to improve the efficiency in training supervised models fed by Big Data; specific methods for artificial vision that consider the variations of position and light in different images of the same object; or, how to process natural language as input when analyses have been conducted only on dense representations.

2. **Reflective Observation**. In this phase, the students need time to detect inconsistencies between experience and understanding before giving possible answers to the problem. Jacobson and Ruddy [12] propose five questions that can be instantiated to the concrete experiences to start the reflection or group discussion: Did you notice...?; Why did that happen?; Does that happen in life?; Why does that happen?; and, finally, How can you use that?.

For example: Did you notice that Facebook recognizes the faces in the pictures so you can label them?; or, what happened when you try to train a model with millions of instances as training data?.

3. **Abstract Conceptualization**. Reflection brings new ideas to address the problems faced in the experience and observation, or a modification of an existing abstract concept.

Modifications of already studied methods can be requested to the students as assignments, at least on an abstract level if programming these ideas is too demanding.
4. **Active Experimentation.** The learners apply the new concepts to the world, i.e. the data studied, to see what results. If implementing the new ideas is feasible and there are clear quality metrics, this experimentation can be *gamified* in the context of a contest among the students.

This is a great advantage of the DS field when studied under EL. In this vein, Kaggle [4] facilitates academic machine learning competitions where a percentage of the data is retained for testing. This assures an objective ranking of the contestants and discourages them from using overfitted models. Additionally, a reflection on the results and the data they are based on is a must under the EL paradigm. If the implementation of the new concepts is too complex or too time consuming, a discussion of the solution design with the rest of students can be undertaken. Some material addressing the same problem can also be provided to go deeper into the subject in the following iterations.

5. **Experiences in Deep Learning**

The framework presented in section 4 is being currently used by the authors to create experiences for the “Deep Learning” course of the “Master in Data Science (EIT Digital Master School)” at the Technical University of Madrid².

Studying the students’ profile is the key to design experiences which are challenging but not frustrating. In our case, this is eminently technical. Hence, they tend to focus on programming; rather than other major topics of DS, such as statistics or reflection on a particular application domain. Therefore, they tolerate programming assignments easily, but they may find difficult to express ideas and solutions in a more abstract manner. Considering this, after the course introduction, three major units with three experiential approaches are proposed:

1. **Artificial Neural Networks (ANN).** This unit follows the classical flow in a DS course. Firstly, ANN architectures, which are the cornerstone of the Deep Learning field, are explained for both the supervised and unsupervised learning. Secondly, practical advice in solving problems with these networks are described along with two DS frameworks (Weka and H2O.ai³). Finally, a programming assignment is given to apply the explained ideas. More specifically, the practical experience proposes to train and test ANN architectures for breast cancer prognosis to predict whether a breast lesion is malignant or benign giving some attribute values as input ⁴.

2. **Computer vision.** This unit deals with the computer vision problem, i.e. how computers can get high-level understanding from images or videos. After introducing the topic, a contest is proposed to use the methods learned in the first unit to a computer vision problem: predicting the object depicted in an image. For this purpose, the CIFAR-10 dataset is employed⁵: 60000 32x32 colour images in 10 classes, with 6000 images per class. The experience allows a reflec-
tive observation of the low accuracy achieved (all students obtained under 57%), and an abstract conceptualization of some of the challenges of computer vision. Then, Convolutional Neural Networks (ConvNets) are explained along with a DS framework to implement them (Caffe\(^6\)). This allows students to retake the contest and observe the improvement achieved by the new ideas introduced in the course: around 80% accuracy in some cases. Finally, the same approach is followed for a transfer learning problem, similar classification with a small dataset, proposing a third contest / experience.

The key in the experiential approach of this second unit is that students get their prior knowledge challenged by new problems. Learners have time to try known methods to new situations and to reflect on the results. Moreover, the contests act as game-based approach for the EL as proposed by Shiralkar [19]. Students are not required to research on new methods for the new experiences proposed.

3. Applications. This unit starts with an experience in a new application field of Deep Learning: natural language processing (NLP), i.e. how computers can interact with human (or natural) languages. The learners choose a workgroup of up to five members and face a realistic experience: they are members of a research team that wants to apply for a contract offered by a well-known newspaper. They are asked to design a solution to predict the relevance of an article headline regarding its body from a dataset with three attributes: headline, body, and class (relevant / irrelevant). Students have two hours available in a laboratory with computer equipment to carry out an investigation into the problem, and present a manuscript with the solution design. The manuscript should include: a flowchart with the main tasks involved, the DS frameworks that will be used and why, and the bibliographic references. Two short papers discussing NLP methods will be provided at the end of the class to reflect on possible changes in the proposed solution. In the next class, the workgroups will have the opportunity to present their solution to be discussed with the other groups. Finally, a brief lecture on NLP concepts and how to apply them to the study case is thought.

As in the experiences of the previous unit, a challenging problem considering the prior knowledge is proposed. The text requires a very elaborate preprocessing and word embeddings methods to feed ANNs. A key aspect of this experience is that, instead of a concrete implementation, the design of a solution is requested. This is more demanding for the students’ profile in this master. Furthermore, the experience gives freedom in the investigation and proposal of solutions to a problem instead of stating a non debatable approach.

The experiences described here allow professors (or “facilitators” in EL terminology) to provide tools for students or “learners” to build their own knowledge, i.e. learning how to learn. Students’ responsibility of obtaining new knowledge is incremental throughout the experiences, requiring to research on a solution for a new problem in the final unit. After finishing the Deep Learning course, students will be surveyed to evaluate different aspects of these experiences.

\(^6\)Caffe website: http://caffe.berkeleyvision.org/
Datasets repositories are a widely used resource in the DS education. Some examples are The UC Irvine Machine Learning Repository [7] and the Kaggle dataset list [5].

UCI datasets are complemented with: descriptive information of the predictive variables, the default task (classification, regression, clustering, recommendation, relational learning, etcetera), and relevant papers which employ the data.

Kaggle, with a significant qualitative leap, offers an immense database of datasets. Moreover, a selection of “featured datasets” particularly well-formatted and documented is selected by the Kaggle team. The website also offers basic tutorials, discussion forums for each dataset, a listing of recent activity, and “Kernels” that users can include to published datasets. Kaggle Kernels is a cloud computing environment that enables reproducible and collaborative analysis. Kernels supports scripts in R and Python, Jupyter Notebooks, and RMarkdown reports. Kaggle is also popular for its active and very lucrative competitions than can be proposed as experiences.

Kaggle is a great approximation to the platform of experiences and it has the potential of engaging students with opportunities to learn through doing. Unfortunately, it repeats the harmful patterns outlined in this paper. This is, the DS is seen as a black box: get the best accuracy you can, nothing else matters. Again, the reflection on the results is essential in EL, and getting insights into the data is fundamental in DS. Under these considerations, some requirements for a platform of experiential learning to offer structure and reflective skill building are:

1. A decision support system for the creation of experiences rather than mere datasets. These include not only the type of problem to be solved (classification, regression)... but the context in which this problem arises for the specific data that permits to elaborate a realistic and engaging experience.
2. The use of this system, not only by instructors, but also by students to propose the experiences that motivate them.
3. A reservation system to allow different workgroups or students to deal with different experiences.
4. A rating system for experiences to allow educators to assess their difficulty, and to articulate a course as a series of experiences.
5. A description of the type of knowledge assumed (ideally indexing previous experiences), and how the new experience challenges that knowledge by posing new problems and situations.
6. A private space for the instructors to consult previous and ongoing results. This prevents students from consulting the answers omitting the reflection and conceptualization phases.
7. The possibility of releasing extra material during a experience that, without giving a solution to the problems, can guide the research and reflection.
8. An anonymous peer evaluation system to facilitate reflection on experiences of other students.
9. Linking or access to environments for reproducing and repeating the experiments such as Jupyter Notebooks.
7. Conclusions and future works

This paper has presented an ongoing educational research project in the use of experiential learning (EL) methods for Data Science (DS). Although there are recent and relevant works proposing this kind of experiences for information technologies [19], this project seems to be the first research work on EL for DS.

This paper lists a number of shortcomings repeated in the hegemonic pedagogical line of prestigious DS courses. Among others, a lack of reflection and critical thinking in the solutions given to programming assignments. To cope with this situation, a framework for EL in DS is proposed based on Kolb’s experiential learning style theory. In the scope of a Deep Learning course, several concrete experiences are detailed increasing students’ degree of freedom to propose new solutions. Finally, some popular resources in DS teaching are revised and a number of requisites are proposed to move from the dataset repository to a platform of experiences.

Our main future works include the survey on the acceptance and quality of experiences proposed in Deep Learning, studying new theoretical frameworks for DS, and the exploration (or implementation) of software tools for the proposed platform of experiences.

Acknowledgments

This research work is supported by Universidad Politécnica de Madrid under the educational innovation project “Métodos, experiencias y herramientas para el aprendizaje experiencial de la Ciencia de Datos”; and by the Spanish Ministry of Economy, Industry and Competitiveness under the R&D project Datos 4.0: Retos y soluciones (TIN2016-78011-C4-4-R, AEI/FEDER, UE).

References