Insights into the design of an introductory course for data science and machine learning for engineering students

Katharina Bata, Angela Schmitz, Andreas Eichler

Insights into the design of an introductory course for data science and machine learning for engineering students

Katharina Bata[1], Angela Schmitz[1] and Andreas Eichler[2]

[1]TH Köln – University of Applied Sciences, Germany; katharina.bata@th-koeln.de; angela.schmitz@th-koeln.de

[2]Universität Kassel, Germany; eichler@mathematik.uni-kassel.de

Due to their interdisciplinary nature, data science methods, such as machine learning, can be taught in many different ways. This paper presents an approach that takes advantage of the close content connection to statistics and of the mathematical structure of data science methods to develop an introductory course for engineering students. Following the research methodology of design research, we discuss the theoretical motivation and methodological implementation of the design principles for the course and show first insights into empirical results from the design cycles.

Keywords: Data Science, Machine Learning, Design Research, SRLE.

Introduction

A large proportion of the methods attributed to data science are (computer-aided) applications of statistics, making data science an essential topic in the statistics education research community (Engel, 2017; Gould, 2017). The possibility of applying data science methods in almost all areas of industry and research has created a need for subject-specific concepts for teaching the methods over the past years (Engel, 2017; Grillenberg & Romeike, 2018).

Due to the interdisciplinarity of data science between mathematics, statistics, computer science, ethics, and the respective reference science, there are many ways to approach the topic. We want to show insights into the design of an introductory course for data science (DS), particularly machine learning (ML), for early mechanical engineering bachelor students in a few lectures focusing on statistics, particularly mathematics. The development of the course follows the methodology of design research (Gravemeijer & Cobb, 2006).

In this paper, we motivate the design principles of the course and present their respective elaboration. For this reason, the central part of this paper is theoretical, followed by examples from the introductory course that illustrate the explicit implementation of the design principles. In the end, we switch to a first empirical evaluation and give a brief insight into the students' views on the developed introductory course.

Theoretical considerations for designing the introductory course

This section gives insights into the current state of research regarding teaching DS and in the methodology of design research, followed by the theoretically motivated design principles.

The current state of research in teaching data science with a focus on machine learning

In a DS study program, the versatility of what can be taught and in which ways it can be taught is wide (Grillenberg & Romeike, 2018). There are different approaches to the concretization of a DS

curriculum for schools (Heinemann et al., 2018), undergraduate programs (De Veaux et al., 2017), and competence models for sub-aspects such as data literacy (Ridsdale et al., 2017) or data management (Grillenberg, 2019). One subfield of DS and part of many DS curricula is data analysis, especially ML (Grillenberg & Romeike, 2018). There are many open questions and few empirical studies about how learning ML occurs under different teaching methods (Steinbach et al., 2020).

Especially for students without a mathematical or computer science background, there are different approaches to how to deal with the more complex mathematical and programming details that seem to be a hurdle for students (Lavesson, 2010). Suppose one additionally considers the easy accessibility of methods nowadays, there is a danger: Using ML without theoretical expertise, for example, on fundamental mathematics and statistics, creates the risk of harmful socio-technical systems (Heuer et al., 2010). To date, there is little consideration of the role of ML in the context of statistical literacy and data literacy (Grant, 2017; Kadijevich & Stephens, 2020; Schüller, 2017). In this context, the distinction between the terms statistical and data literacy is still fluid, with broad similarities, and somewhat arbitrary (Gould, 2017; Schüller, 2017).

Theoretical considerations on design research

The research methodology of design research focuses on the close connection between the systematic design of teaching-learning material and the investigation of learning processes working with this material (Gravemeijer & Cobb, 2006). Especially in the case of little empirically tested teaching-learning material, design research can be used sensibly with the two following goals: To get empirically tested and cyclically improved teaching-learning material and to get research results on the learning processes of the target group when working with the material.

For this purpose, first, a prototype of the teaching-learning material is developed, considering the so-called design principles (see section *The design principles for the introductory course*). The development of the prototype also includes theoretical considerations about the students' learning processes, so-called intended learning trajectories. Subsequently, the prototype is tested with the target group in the so-called design experiments, for example to compare the students' individual learning paths with the intended learning trajectories. By analyzing the design experiments, a local (concerning the target group and the material) teaching-learning theory emerges, which contributes to the further development of the material. A cyclical continuation then provides improved teaching-learning material and a sharpened local teaching-learning theory (Gravemeijer & Cobb, 2006).

The design principles for the introductory course

In this section we motivate the design principles (DP) of the course and explain their methodical implementation. In the following section *Insights into the course,* two examples, *The unit square* and *Reflection tasks,* illustrate how the design principles are incorporated into the design of the course.

The first design principle is *Strong inclusion of statistics and mathematics to approach the DS/ML methods* (DP1). There are two main reasons for this design principle: One is the proximity between DS and statistics, respectively mathematics, in terms of content and the personal interest in this connection. The other is the fact that engineering students are, due to their curriculum, a target group with a comparatively strong mathematical background.

To implement the first design principle, we use the "four-level approach for specifying and structuring mathematical learning content" (Hußmann & Prediger, 2016). The four-level approach illustrates how to proceed methodically when the focus within a design research project is on analyzing the learning content. Using the approach, the prototype of the material emerges by answering a series of systematic questions on three theoretical levels in the sense of a "classic didactical analysis of subject matters" (Hußmann & Prediger, 2016). The first level, the formal level, addresses the logical structure and the formal representation of the (mathematical) learning content (objects and procedures). The second, the semantic level, addresses the sense and meaning of the objects and procedures under study; helpful representations and mental models are identified and linked to each other (see a concrete example in section *The unit square*). On the subsequent concrete level, the last theoretical level, learning situations, and examples for experiencing the concepts and procedures are developed (see a concrete example in section *Reflection tasks*). The fourth level then equals the implementation and evaluation of the design experiments.

The second design principle is *Embedding all methods in the overall context of data analysis* (DP2). It is an idea, which is used in several different contexts while learning methods to handle with data (Heinemann et al., 2018; Wild & Pfannkuch, 1999).

This design principle is implemented by using the "CRoss-Industry Standard Process for Data Mining" (CRISP-DM, Chapman et al., 2000) model to structure the course and some teaching activities. The CRISP-DM model is a process model that describes all essential steps of a DS process in an industrial context, starting with a question and ending with the implementation of the results. It has already been fruitfully used in other projects to structure teaching activities in the context of DS (Heinemann et al., 2018). The CRISP-DM also gives an overview of some core ideas of DS, according to DP3 (*core ideas of DS and ML*, see next paragraph), and we additionally use it in the sense of DP4 (*classroom activities*, see next paragraph) to design tasks that encourage students to discuss core ideas and own proceedings (see a concrete example in section *Reflection tasks*).

The further four design principles (DP3 to DP6) refer to the basic ideas of the "Statistical Reasoning Learning Environment" (SRLE, Garfield & Ben-Zvi, 2008). The SRLE is a well-structured and proven approach to create teaching-learning environments in the context of data. The origins of the SRLE go back to Cobb (1992) and were developed further by different statistics educators within the following decades. Because of the close proximity of DS and statistics in terms of content, it offers to use some ideas of the SRLE as design principles for the introductory course.

The following ideas from the SRLE (Garfield & Ben-Zvi, 2008) are adopted as design principles: *Focusing on the developing core ideas of DS and ML* (DP3, original: "Focus on developing central statistical ideas"), *Using classroom activities to support the development of students' reasoning* (DP4), *Using realistic and motivating data sets* (DP5) and *Integration of appropriate technological tools* (DP6). DP4, DP5 and DP6 are adopted literally from the SRLE.

Insights into the course

To give an overview, we first present the content components of the course. Then we illustrate how the design principles shape the course by giving two examples.

The course components

The selection of the learning content is mainly based on the subjectively set goal of the course to convey the usefulness, practical relevance, and methodology of DS, especially ML, in the engineering sciences. Students shall be enabled to delve deeper into the topic of DS and ML. This goal results in three sessions of approximately 3 hours each:

Session 1 - fundamentals: A first overview of the possibilities to use DS methods in engineering is shown, and the CRISP-DM model is introduced. The handling of data within this setting is discussed and the basic concepts of ML up to classification are introduced.

Session 2 - k-nearest-neighbor classification (kNN): The basic concepts of ML are explored in depth by discussing the kNN as a possible method for classification.

Session 3 - model quality of classification models: Model properties (variance and bias), as well as different performance measures (accuracy, precision, recall), are discussed to be able to evaluate and compare classification models and to select the model parameters for a specific question.

The following example *The unit square* shows the use of the "four-level approach" on the first two levels, and thus gives insights into the implementation of DP1. The next example *Reflection task* shows the use of CRISP-DM as the elaboration of DP2 and some synergies with the design principles adopted from the SRLE (DP4 to DP6).

The unit square – An example for the analysis of the learning content on the first two levels

In Session 3, *model quality of classification models*, different model characteristics and performance measures are discussed with the students. When creating a classification model, the available data set consists of examples with the characteristics of the independent variables (called features) and a dependent variable (called a label). The total data set is first divided into training data and test data. The training data is used to build the model, and the test data is then used to check how well the model can predict the correct label. Concerning a binary classification model, the testing phase is usually represented using a confusion matrix as in Figure 1. Here, the number of correctly classified examples (true positive and true negative) separated by class is on the main diagonal, and the number of incorrectly classified examples (false positive and false negative) is on the opposite diagonal.

		Actual Class	
		Class 1	Class 2
Predicted Class	Class 1	true positive	false positive
	Class 2	false negative	true negative

Figure 1: Example of a confusion-matrix

All performance measures and performance criteria are derived from the values in the confusion matrix. A content analysis of the learning object, as it has been done on the formal level of the four-level approach, reveals that all performance measures can be represented by a probability space, which explains the relations of the values among each other:

Each example classified by the model[1] can be represented as $\omega^i = (\omega_1^i, \omega_2^i), i = 1, \dots n,$ where $\omega_1^i, \omega_2^i \in \{1,2\}$ with ω_1^i representing the actual class of the example and ω_2^i representing the new

[1] We continue to consider a binary classification problem here, a transfer to higher dimensionality does not pose a problem.

classified class. n is the number of examples. This gives $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ with $\Omega = \{\omega^i : i = 1, \dots, n\}$ and $\mathbb{P}$ as the normalized count measure, a probability space that represents the testing phase.

Going on in the formal level, all addressed performance measures and their interrelationships among each other can be represented based on this probability space (for example, certain performance measures are equivalent to conditional probabilities). On the subsequent semantic level, the following questions arise concerning the prerequisites of the target group: How can the performance measures and their interrelationships be communicated without addressing this probability space and, for example, conditional probabilities?

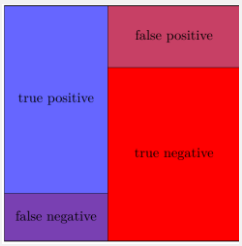

Figure 2: Example of a unit square

The unit square (see Figure 2) is considered to be a proven means to visualize proportions and probabilities up to conditional probability (Böcherer-Linder et al., 2017). The analysis in the two levels reveals that the unit square provides a way of representing the testing phase. The unit square in addition can visualize the performance measures to show them as formulas based on the values of the confusion matrix. Thus, by linking the content to mathematics, the unit square is included as a visual representation in the course.

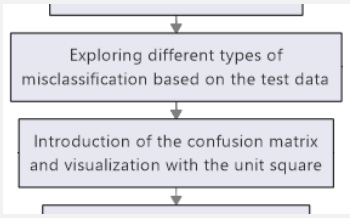

Figure 3: Excerpt of an intended learning trajectory

Besides the elaboration of different representatives of the performance measures (in this example, the unit square, the formula, and the values calculated by hand or with Python), the core ideas of Session 3 (such as the distinction between types of misclassification, which can also be visualized in the unit square) emerge from analyzing the learning content up to the semantic level. The analysis of core ideas, connections, and representatives results in a theoretical sequence in which they can be worked out: the intended learning trajectory (see Figure 3).

Reflection task – An example for considerations on the third level

The work areas of the CRISP-DM are learning objectives (see Session 1). The CRISP-DM is also used to structure the course (see section *The design principles for the introductory course*). The structure is implemented, among others, by giving students a task at the end of each session to reflect on the learnings in the framing of the CRISP-DM. For example, at the end of the third session, when students have to design a model and use the model for a decision afterwards (Bata et al., in press), this task reads:

Reflect on your decision in the Jupyter Notebook together in groups in the context of your notes from the last lecture[2] *and the CRISP-DM. If you find it useful, complete your answers again.*

This task demonstrates the incorporation and the interrelation of some design principles while formulating explicit learning opportunities on the third level of the "four-level approach": Students work with a data set regarding the quality of steel (*realistic data sets*, see DP5), the Jupyter Notebooks are used as technical support throughout the task (*appropriate technological tools*, see DP6). The

[2] This refers to the notes of the reflection task of the past lecture (Session 2).

open assignment encourages students, who are working together in groups at this point, to discuss their results and to defend them argumentatively using the CRISP-DM (*overall context*, see DP2, *classroom activities*, see DP4).

First empirical insights

The design cycles and data collection

The introductory course has been conducted in two design cycles in different settings so far. In the first cycle, seven students participated in a laboratory setting (in groups of two and three, accompanied by the lecturer during the processing of the tasks). In the second cycle, 39 students participated in a course setting. The third session was additionally conducted with 4x2 students in a laboratory setting. All sessions took place via an online conference tool. Each session was video-recorded and transcribed. The group work was additionally documented using written products.

In addition, data were collected using a one-minute paper in each session. Students were asked five questions per session, each to be answered at one point after the session within a given time (usually one minute) and without looking into the learning material. The five questions were intended, among others, to help gather information about the learning environment. The evaluation gives a first insight into the students' views, from which we present first results.

Results

The question of the one-minute paper "How relevant do you think the content of the past lecture is to your studies and future career, and why?" was evaluated using points to characterize the students' answers: 0 points means no relevance, 1 point means medium relevance and 2 points means high relevance. In addition, the reasons were collected and grouped into content-related groups. Mean values between 1.81 and 1.92 across the cycles indicate that most students perceive the learning content as very relevant. However, the reasons for their ratings varied: Only about 10 percent of the students justify the relevance with concrete content like "validation of ML models"; instead, general facts are used as reasons. For example, students mention the presence and relevance of DS and ML in engineering or everyday life or Python as an essential competence for jobs and studies.

Two questions of each one-minute paper focused on the content goals of the particular session, for example: "For which data sets is the performance measure accuracy not recommended?" To evaluate the questions, 1 point (answered completely correctly), 3 points (answered partially correctly), or 5 points (answered incorrectly or not answered at all) were assigned to each response. The scores give an overview regarding the students' learning results concerning the questions. The questions were largely answered in a meaningful way in terms of content, the mean values of the answer points per question ranged from 1.95 to 2.63 across both cycles.

Discussion

This paper gives insights into the design principles and development of a short introductory course for DS and ML for engineering students. Especially the first design principle, implemented by the approach of specifying and structuring the learning content focusing on its statistical and mathematical aspects, opens a way to analyze ML methods, which have so far rarely been investigated

from the perspective of the classic didactical analysis of subject matters. For example, the connection to the unit square has two potentials: On the one hand, learning methods with threshold parameters, which are discussed in every advanced ML course, can be transferred to the representation of the performance measures with an animated unit square. This visualization can show the direct influence of the threshold parameter on the performance measures. On the other hand, the very visual representation of the unit square can be used when students' backgrounds are not as mathematical as in the case of mechanical engineers.

The first analysis of the one-minute-paper questions shows a pleasing result, as the planned contents seem to reach the students and seem relevant to them. Nevertheless, the question arises about how the design principles, and the resulting developed or chosen representations, visualizations, and instructional activities contribute to the students' learning processes. The overall design study aims to explore students' individual learning paths through a qualitative analysis of the resulting video material to address this question. From this analysis, results are expected on whether and how the statistical and mathematical details are learned by students (which is unanswered by now) and used when applying the methods (first results see Bata et al., in press). Based on these findings, the role of statistics and mathematics in ML, specifically in the context of data and statistical literacy, can be addressed in greater depth.

References

Bata, K., Eichler, A., & Schmitz A. (in press). How engineering students argue in an introductory course in data science. *Proceedings of the IASE 2021 Satellite Conference on Statistics Education in the Era of Data Science.*

Böcherer-Linder, K., Eichler, A., & Leuders, T. (2017). Anteile und Wahrscheinlichkeiten darstellen - das Einheitsquadrat als Visualisierung nach dem Spiralprinzip. *MU – Der Mathematikunterricht*, *63*(6), 11–18.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. CRISP-DM consortium. https://the-modeling-agency.com/crisp-dm.pdf

Cobb, G. W. (1992). Teaching statistics. In L. Steen (Ed.). *Heeding the call for change: Suggestions for curricular action* (pp. 3–43). The Mathematical Association of America.

De Veaux, R.D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, Li, Cheng, L. Z., Francis, A., Gould, R., Kim, A. Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R., Sondjaja, M., … Ye, P. (2017). Curriculum Guidelines for Undergraduate Programs in Data Science. *Annual Review of Statistics and Its Application. 4*(1), 15–30. https://doi.org/10.1146/annurev-statistics-060116-053930

Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal, 16*(1), 44-49.

Garfield, J. B., & Ben-Zvi, D. (2008). *Developing Students Statistical Reasoning. Connecting Research and Teaching Practice.* Springer. http://dx.doi.org/10.1007/978-1-4020-8383-9

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal*, *16*(1), 22–25. https://doi.org/10.52041/serj.v16i1.209

Grant, R. (2017). Statistical literacy in the data science workplace. *Statistics Education Research Journal – Special Issue: Statistical Literacy. 16*(1), 17–21.

Gravemeijer, K. P. E., & Cobb, P. (2006). Design research from a learning design perspective. In Van den Akker, J., Gravemeijer, K., McKenney, S., & Nieveen, N. (Eds.). *Educational Design Research* (pp. 45-85). Taylor and Francis Ltd.

Grillenberger, A. (2019). *Von Datenmanagement zu Data Literacy: Informatikdidaktische Aufarbeitung des Gegenstandsbereichs Daten für den allgemeinbildenden Schulunterricht* [Doctoral dissertation, Freie Universität Berlin]. Refubium - Repositorium der Freien Universität Berlin. https://refubium.fu-berlin.de/bitstream/handle/fub188/24160/Grillenberger_Dissertation.pdf

Grillenberger, A., & Romeike, R. (2018). Ermittlung der informatischen Inhalte durch Analyse von Studienangeboten. *Commentarii informaticae didacticae*, *10*(1), 119–134.

Heinemann, B., Opel, S., Budde, L., Schulte, C., Frischemeier, D., Biehler, R., Podworny, S., & Wassong, T. (2018). Drafting a Data Science Curriculum for Secondary Schools. *Proceedings of the 18th Koli Calling International Conference on Computing Education Research – Koli Calling '18, 17*, 1–5. http://doi.org/10.1145/3279720.3279737

Heuer H., Jarke J., & Breiter A. (2021). Machine learning in tutorials – Universal applicability, underinformed application, and other misconceptions. *Big Data & Society*, *8*(1). https://doi.org/10.1177%2F20539517211017593

Hußmann, S., & Prediger, S. (2016). Specifying and Structuring Mathematical Topics. *Journal für Mathematik-Didaktik*, *37*(1), 33–67. https://doi.org/10.1007/s13138-016-0102-8

Kadijevich, D. M. & Stephens, M. (2020). Modern statistical literacy, data science, dashboards, and automated analytics and its applications. *The teaching of mathematics*, *23*(1), 71–80.

Lavesson, N. (2010). Learning Machine Learning: A Case Study. *IEEE Transactions on Education*, *53*(4), 672–676. https://doi.org/10.1109/TE.2009.2038992

Ridsdale, C., Rothwell, J., Smit, M., Ali-Hassan, H., Bliemel, M., Irvine, D., Kelley, D., Matwin, S., & Wuetherick, B. (2015). *Strategies and Best Practices for Data Literacy Education: Knowledge Synthesis Report*. Dalhousie University. http://dx.doi.org/10.13140/RG.2.1.1922.5044

Steinbach, P., Seibold, H., & Guhr, O. (2020). Teaching Machine Learning in 2020. *Proceedings of Machine Learning Research, 141*, 1–6.

Sulmont, E., Patitsas, E., & Cooperstock, J. E. (2019). What Is Hard about Teaching Machine Learning to Non-Majors? Insights from Classifying Instructors' Learning Goals. *ACM Transactions on Computing Education*, *19*(4), 1–16. http://dx.doi.org/10.1145/3336124

Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223-248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x