Teaching Machine Learning with Applied Interdisciplinary Real World Projects

Gulustan Dogan [1]

Abstract

One of the challenges in teaching machine learning to computer science students consists in finding the right hands-on projects that students can work on. Most of the popular Machine Learning projects such as Titanic Survival Prediction and Housing Prices Prediction have many solutions available online. Students working on these problems do not get challenged enough to improve their experimental skills. Moreover, they might follow the methodology in the existing solutions which can discourage them from designing a novel ML solution. Real-world applied projects without online solutions in which students can use their creative problem-solving skills are needed to teach ML courses effectively. Otherwise, students fall into an overfitting problem in which they become ML coders reusing existing codes without ever writing their code. Consequently, we present an approach for creating course project material to achieve this goal. This approach is supported as an applied learning curriculum design methodology with an internal grant at our institution. It is validated experimentally for two semesters in a course taught to both graduate and undergraduate computer science students.

1. Motivation

There is a growing interest in machine learning from many different domains. Besides computer scientists, scholars, from different disciplines look for methods to use ML to solve their domain-specific problems. For instance, Physics problems that would only be solved with extensive simulations can be solved with an ML approach consisting of a couple of lines of Python code (Acar et al., 2022).

There have been papers in previous versions of this workshop that also propose courses for ML across disciplines (Guhr et al., 2022). In this work, the scholars have taught ML to physical science students (Acquaviva, 2021). They also emphasize that the existing datasets are very ”curated” limiting the students learning opportunity to clean and prepare a dataset. To overcome this problem and create real-world likeliness in the projects they work on real-world datasets that come with research papers. Muller et al. have a workshop paper on teaching ML to molecular biologists (Müller et al., 2021). They have stated that the choice of datasets relevant to the domain is important. Hagen et al. also use real-world interdisciplinary data from companies not to use artificial scenarios for teaching ML (Hagen et al., 2020). Rather than introducing students to machine learning with toy datasets such as the Iris or MNIST dataset, Kazmi also designed their course using domain-specific real-world examples (Kazmi, 2021).

There is a vast amount of free sources in ML such as blogs, tutorials, online classes, and code snippets (ml). Even a beginner in ML can implement an image recognition model in a couple of hours by following an online tutorial and reusing existing code. We believe that this can motivate students by preventing them to struggle too much and make them like the field. However, although this is a great advancement, it can become a roadblock for students to gain a deeper understanding of ML systems by building/designing their solutions.

We observed in our classes that often students reuse existing code and upload Kaggle solutions as their assignments without really understanding the code. Also if they would be asked to design an ML solution for a different domain, they would generally be clueless.

To overcome this problem, an applied learning teaching methodology was proposed as part of the Applied Learning and Community Engagement Pedagogy Initiative Awards which is an internal grant in our institution. The proposal got accepted and the methodology offered in the proposal has been implemented in ML classes for two semesters. This paper will present the methodology designed and will share the insights gained over two semesters.

[*] Equal contribution [1] Department of Computer Science, University of North Carolina Wilmington, Wilmington, USA. Correspondence to: Gulustan Dogan <dogang@uncw.edu>.

Proceedings of the 3^{rd} Teaching in Machine Learning Workshop, PMLR, 2022.

2. Interdisciplinary Machine Learning Intelligence Initiative

In the Fall 2021 and Spring, 2022 semesters Advanced Artificial Intelligence (AI) course is taught according to Applied Learning principles that will be described in this paper. The course is taught to both graduate and undergraduate students.

Specifically, graduate and undergraduate students were organized into teams and assigned real research questions using data sets owned by faculty from the different departments of our institution such as Exercise Science, Health Sciences, Geology, and Marine Sciences. Teams consisted of both graduate and undergraduate students. Special attention was given in forming teams to make sure every team had both graduate and undergraduate members.

Machine Learning is a highly interdisciplinary field. Problems in various domains such as finance, environment, health, nature, art, law, and engineering can be solved with AI. Faculty in different departments have datasets related to the research they are doing. They also have research problems that can easily be solved with Artificial Intelligence, however, they lack the technical CS (Computer Science) knowledge to do this. As a preparation for this course, many non-CS faculty were reached out, and the mentor faculty were very enthusiastic about working with CS students on their own research project.

Faculty from different disciplines made a presentation of their research problem and datasets at the beginning of the course and students chose which faculty they want to work with on a first come first served basis using the Teams interface of the Canvas Learning Management System that our instution uses.

The instructor of the course monitored each team and helped them with the technical (cleaning, coding, testing, deployment) part of the project. Faculty who own the datasets helped guide students throughout the conceptual side of the project, explaining the dataset, the research problem, and domain knowledge. The course has an implicit objective to produce publishable papers from the interdisciplinary project results, this worked as a good strategy to incentive writing skills and connections to research.

For computer scientists, these problems from interdisciplinary fields are a good opportunity to test our algorithms using real-world data, however, we lack domain knowledge. On the other hand, students often complain about not being able to apply the theoretical knowledge they learn, and they have difficulty connecting theory with practice. Also as mentioned using case studies with toy datasets often fail to teach critical thinking skills to students. These projects opened incredible applied opportunities for them as it is fascinating to see how CS knowledge can solve real-world problems. Some of the graduate students decided to pursue these topics as their thesis topics. This collaboration between fields, faculty, and students helped in creating an interdisciplinary research culture in our institution, which is a strategic goal.

3. Student Activities

Student teams were assigned a faculty mentor of their choice who owns the data and the research. Their faculty mentor guided them throughout the project by explaining to them the dataset, the research problem, and domain knowledge. Their first step was data cleaning and preparation. Data collected from a real-life event is likely to have missing fields, faulty values, and unreadable content. After cleaning the data, they selected various machine learning, and deep learning methods to apply to data to answer the research questions. Based on the nature of the data and the type of the problem, they had to implement different methods. They were introduced to different ML methods during the lecture part of the course. During class time, instruction was delivered via participatory live coding. We typed and explained code in real time and students followed along. Students implemented the appropriate classification or prediction methods to their project problem based on the problem they are working on. They were encouraged to implement many different ML algorithms and share the results as a table in their research paper. After deploying the model, the model was tested with test data, and based on the model's performance needs for tuning, optimizations were considered. After the final model implementation, they wrote a paper on the dataset, the methods, the results, and the planned future work and submitted their paper to a conference.

4. Sample Projects

The courses were taught synchronously online. Canvas is used as a platform to share the material. For each project, a Google Drive folder was used to keep track of the team project's files. A research paper summarizing the project outcomes was written in an online, cloud-based, collaborative Latex platform called Overleaf. Codes that students have written are kept on Google Drive as cloud accessible, shared Google Colab files. Google Colab was chosen over Gitlab or Github. As the project spanned only a semester, version control was not needed.

Some of the sample projects of the course can be listed as follows. A project on understanding factors on the divers' balance following their dives, using force plate data was carried out with a mentor from the Exercise Science faculty. With a mentor from the Department of Earth and Ocean Sciences a project on population vulnerability vegetation classifications based on the data image from satellite drones was done. One of the projects done with the Department of Physics Physical Oceanography is on applying object detection, neural network models to identify the damaged houses, using the aerial imagery taken after Hurricane Florence. An interdisciplinary Earth and Ocean Sciences project on developing a model which improves the accuracy of remotely sensed coastal water quality, using satellite data and in the situ water quality data was done. With the School of Nursing, a project to find out depression patterns in the data of the parents of children with special health care needs has been carried out. Papers that were written throughout this course are listed as follows: Development and Comparison of Machine and Deep Learning Models for the Prediction of Land Degradation,Rainfall Forecasting With Deep Learning, Ocean Wave Fun Factor Prediction with Artificial Intelligence, Predicting Water Quality Estimates using Satellite Images in Coastal and Estuarine Environments (Hogan et al., 2022), Short Term Prediction of Rainfall in Columbia using a Generative Modeling Approach, Building Damage Assessment From Remote Sensor Imagery Using Classification Model, Defining the Movement of Free-Ranging Adult California Sea Lions, Classification of Activities and Falls within a Multimodal Dataset (Kurpiewski et al., 2021), Using AI to Predict Caregiver Ability to Self-Manage Chronic Illness When Caring For Children With Special Health Care Needs, Using Artificial Intelligence to Predict Patient Electronic Health Record Access Points, Using Artificial Intelligence to Detect Falls.

All of the projects had a conference paper as a deliverable. Since it was emphasized that their grade is determined based on the final paper, all the teams produced a research paper and submitted it to the conference selected by the instructor successfully. A prestigious AI conference with a low acceptance rate was chosen as the venue to get good feedback about their paper. The goal was to encourage students to learn from the rejection email and reviews. After getting reviews, even after the course is over, students edited their papers based on the feedback and submitted them to some other conferences. We make sure each paper is published so that students can list the publication in their CVs.

5. Learning Outcomes of the Course

The goal of the course was to implement an AI application for a real-world problem. The learning outcomes were directly related to the goal of the course. Students demonstrated proficiency in developing hands-on, real-world AI applications in Python programming language. Their code had to compile, run and produce output. By working in teams, they developed project management, communication, and problem-solving skills.

Students learned to understand a real-world problem, communicate technical requirements with someone from a different discipline, understand the requirements, analyze the requirements, design a solution, implement a solution, test, debug and re-design if needed. They also developed teamworking skills. Besides coding, they were asked to document their findings as a paper and do a poster presentation in the research showcases organized on campus. They developed technical writing skills while writing the research paper as an academic conference paper. The paper followed the writing guidelines of the conference. Research papers were submitted to an International Conference On Machine Learning And Applications. They also did presentations at the end of the semester. Faculty mentors were invited to discuss general research questions and interact with students to give feedback about their papers and presentations. After the projects were presented, students edited their papers based on the suggestions they received. After writing the final version, they submitted their research paper to the conference. This also served as an encouraging factor in the class as students were motivated by getting acceptance from the conference.

6. Evaluation of the Course

The course evaluation was based on homework and the course project. As the course project required a huge amount of time, not much homework was given. Course Project which is essentially developing an AI solution and writing a paper on it determined students' course grades. If the team had a working ML code and a paper at the end of the term, the students passed with an A.

Overall students did not feel grade pressure during the course as we emphasized that as long as they put the required effort, they will be getting a good grade. This emphasis was needed for this course as students were feeling overwhelmed by the complexity of the project and not having grade pressure helped them reduce their stress and feel more motivated.

7. Conclusion and Lessons Learned

To provide opportunities for students and mentors to engage in reflection, they were asked to write a report using the Gibbs Reflection Framework (Gibbs, 1988). The Gibbs Reflective Cycle has six distinctive stages, namely Description, Feelings, Evaluation, Analysis, Conclusion, and Action plan. This framework was chosen because these six stages help go from a description of the experiences to conclusions and considerations for plans. Gibbs framework worksheet was provided to students, and they filled in the worksheet. Two reflections were made, one during the middle of the semester and one at the end of the semester.

In the reflection documents, most of the students mentioned that they enjoyed working on a real-world problem with a faculty mentor. One of the steps that they had the most difficulty with was data preparation. As real-world data has many missing fields, faulty information, and misspelled text from text boxes. Once they cleaned their data, they tried to apply the ML methods they learned in class to their problem. However, their data was complicated and the basic information they learned in class was generally not sufficient. For instance, during the class, they were introduced to prediction using the housing dataset. They tried to reuse the house prediction code to solve their real-world problem and ran into problems. They needed help with their code as it was not easy for them to implement AI algorithms. In a semester, learning the subject, doing the implementation, and writing the paper was too much work for them. Especially undergraduate students complained about the workload. Writing the paper was easier for them once they have the code since they were provided with sample papers and overleaf templates.

One issue that was addressed by students was the difficulty in framing the research problem that was presented by the faculty as an ML problem. As the mentor faculty did not have an AI background, they would most of the time be interested in statistical analysis. Students had to present to them how they could benefit from a prediction. Also, it was challenging for students to decide what to predict when the mentor faculty did not have clear directions on which data field to predict.

The following concluding points can be made based on the discussions with the mentor faculty and students and their reflection reports. An ML class that will use real-world problems as a case study should carefully select the interdisciplinary projects. Prior meetings with the mentor faculty should be made to make sure their data and research problem is suitable for the class. Also in these meetings, the course instructor should help the faculty mentor frame their research problem as an ML problem. So when the students start their project, they will have a clear, well-planned research question. Outside of the lecture time, students need additional help from the instructor to look closely at their code and solve the obstacles they are facing. If lecture time is used for solving the project problems, students do not get the required time to learn basic ML concepts. On the other hand, traditional ML assignments such as submitting to the Kaggle Titanic competition should be part of the curriculum so that students can gain confidence working on easier problems. They can also compare the real-world problem with the popular ML problems and understand the difference.

Overall the idea of using interdisciplinary real-world ML problems as a term project in teaching an ML course has received a lot of good feedback. It has also helped the institution foster an interdisciplinary environment. Students could get away from the comfortable learning environment introduced in online ML resources such as blogs, tutorials, and experience designing an ML solution for a real-world project.

References

Intro to Machine Learning ML Zero to Hero. `https://laurencemoroney.com/teaching.html`. Accessed: 2022-06-19.

Acar, Z., Nguyen, P., and Lau, K. C. Machine-learning model prediction of ionic liquids melting points. *Applied Sciences*, 12(5):2408, 2022.

Acquaviva, V. Teaching machine learning for the physical sciences:a summary of lessons learned and challenges. In *ECMLPKDD 2021 Workshop TeachML*, 2021. URL `https://openreview.net/forum?id=z-N7JMHjLO`.

Gibbs, G. Learning by doing: A guide to teaching and learning methods. *Further Education Unit*, 1988.

Guhr, O., Kinnaird, K. M., and Steinbach, P. Teaching ml in 2021-an overview and introduction. In *Proceedings of the Second Teaching Machine Learning and Artificial Intelligence Workshop*, pp. 1–4. PMLR, 2022.

Hagen, T., Lauer, T., Sänger, V., Dorer, K., and Trahasch, S. Teaching practical machine learning concepts to professionals and students: An integrated and interdisciplinary qualification project. In *2020 IEEE Frontiers in Education Conference (FIE)*, pp. 1–8, 2020. doi: 10.1109/FIE44824.2020.9274204.

Hogan, R., Ford, M. R., Battaglia, M., Sturdivant, J., and Bresnahan, P. Predicting water quality estimates using satellite images in coastal and estuarine environments. In *The Journal of Computing Sciences in Colleges*. Consortium for Computing Sciences in Colleges, 2022.

Kazmi, H. Teaching machine learning through end-to-end decision making. In *ECMLPKDD 2021 Workshop TeachML*, 2021. URL `https://openreview.net/forum?id=4rOjJXIcU0`.

Kurpiewski, E., Samokhvalov, I., Layne, B., Rasaq, L., Dogan, G., and Heijnen, M. Classification of activities and falls within a multimodal dataset. In *2021 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)*, pp. 1–5, 2021.

Müller, R., Fasemore, A. M., Elhossary, M., and Foerstner, K. U. A lesson for teaching fundamental machine learning concepts and skills to molecular biologists. In *ECMLPKDD 2021 Workshop TeachML*, 2021. URL `https://openreview.net/forum?id=knwKgaspObQ`.