Data Science Project 2 @ HSLU

Version **FS24_01**

Authors: Umberto Michelucci, Ludovic Amruthalingam, Aygul Zagidullina, Daniela Wolff, Alberto Todeschini; © 2024 HSLU

Introduction

DSPRO2 (Data Science Project 2) course (offered by HSLU in the spring semester) will focus on machine learning frameworks and machine learning operations (MLOps).

The course uses the project based learning approach. **Problem-Based Learning (PBL)** [1,2] is an educational approach that revolves around using real-world problems as a context for students to learn critical thinking, problem-solving skills, and acquire knowledge about a specific subject. Here's a brief overview:

1. **Focus on Real-World Problems**: In PBL, learning starts with a problem that is complex and mirrors real-life situations. These problems don't necessarily have a clear, correct answer, which encourages students to engage in deep and critical thinking.
2. **Student-Centered Learning**: PBL shifts the traditional teacher-led approach to a more student-centered one. Students take charge of their learning process, with the teacher acting as a facilitator or guide. This empowers students to conduct research, apply knowledge, and arrive at their solutions.
3. **Collaborative Effort**: It often involves students working in groups, promoting collaboration and communication skills. Through discussion and teamwork, students

share knowledge, challenge each other's understanding, and develop interpersonal skills.
4. **Integration of Multiple Disciplines**: PBL typically requires knowledge from various disciplines, encouraging students to integrate and apply concepts from different subject areas to solve the problem.
5. **Development of Lifelong Learning Skills**: PBL helps in developing critical thinking, research, and analytical skills. It also nurtures self-directed learning, as students learn to find resources, evaluate information, and reflect on their learning process.
6. **Assessment**: In PBL, assessment is often based on the process as well as the final outcome. This includes assessing the student's research process, participation, contribution, and the practical solutions they develop.
7. **Application and Reflection**: Students are encouraged to apply what they have learned to real-world situations and reflect on their learning journey, understanding both the content and the context in which it is used.

PBL is particularly effective in engaging students actively in the learning process, making education more relevant and interesting, and preparing them for real-life challenges.

Students are required to select a project that possesses specific traits outlined in this document and tackle it uniquely, employing a variety of tools and techniques that will be explored throughout the course.

Course Learning Goals

The main **learning goals** of the course are:

- Learn to setup a data science project with a cloud service
- Learn to setup and use a data science and experiment tracking tool
- Learn the hands-on aspects of MLOps
- Learn to use a kanban board to work in a team and overcome challenges and obstacles
- Learn to setup a clean and structured github repository
- Learn to justify the choices done in the project

The course is structured in a series of lectures at the beginning of the semester, followed by a series of coaching sessions.

Language

The official language course is english.

Course GitHub

https://github.com/LucerneUniversityOfAppliedScience/DSPRO2-FS24

Coaching Sessions

Coaching sessions will be in Rotkreuz. No online session will be possible unless specifically declared in the timeplan. Thursday evenings are reserved for project work. Friday mornings for lectures or coaching sessions.

Project Requirements

To be acceptable for the course, the project must use **all** the following components:

- Cloud service: google cloud, AWS or MS Azure (to ensure that this is feasible, a discussion with the coaches and a detailed description in the proposal is required)
- A kanban board for project management (which tool will be discussed during the course)
- Machine Learning Experiment Tracking tool (the tool wandb will be taught during the course, but you can choose another one if you prefer)

Additionally it is strongly recommended (but not required) that the students use the following component:

- Unstructured data (images, sound, video, etc.) as inputs

The requirements for the project are designed to provide a comprehensive, real-world experience in handling modern data science and machine learning projects. Here's a brief justification for each requirement:

1. **Cloud Service (Google Cloud, AWS, etc.)**:
    a. **Exposure to Industry-Standard Tools**: Using a cloud service like Google Cloud or AWS provides hands-on experience with tools that are widely used in the industry. It prepares students for real-world scenarios where cloud computing is essential due to its scalability, performance, and flexibility.
    b. **Access to Advanced Technologies**: Cloud platforms offer advanced machine learning and data processing capabilities, enabling students to work with state-of-the-art technologies.
2. **Machine Learning Experiment Tracking**:
    a. **Facilitates Model Development Process**: Experiment tracking is crucial for managing and documenting the various experiments conducted during model development. It helps in comparing different models, tuning parameters, and ultimately selecting the best model.

    b. **Promotes Best Practices**: Incorporating experiment tracking emphasizes the importance of organization and documentation in data science workflows, which are key to reproducibility and collaboration in professional environments.
3. **Unstructured Data (images, sound, video, etc.) as Inputs**:
    a. **Dealing with Complex Data Types**: Working with unstructured data like images, sound, and video presents unique challenges and is highly relevant in today's data-driven world. It enhances students' skills in handling and processing complex data types.
    b. **Diverse Application Areas**: Such data types are common in numerous fields, from medical imaging to entertainment, and learning to work with them opens up a wide range of application possibilities for students.
4. **Provide Infrastructure Management (track and report on costs, usage, etc.)**:
    a. **Understanding of Operational Aspects**: Managing infrastructure includes monitoring costs and resource usage, which is vital for any project, especially in a cloud environment where resources are billed based on usage.
    b. **Real-World Project Management Skills**: This requirement teaches students the importance of budgeting and resource optimization, skills that are crucial in a professional setting.
5. **Use of a kanban board for Project Management**:
    a. **Project Organization and Collaboration**: a kanban board is a popular project management tool that helps in organizing tasks, tracking progress, and facilitating collaboration among team members.
    b. **Exposure to Project Management Tools**: Learning to use such tools is essential for students, as project management is a critical component of any real-world technical project.

These requirements ensure that students gain practical experience with current technologies and practices, preparing them for careers in data science and machine learning. They also ensure that projects are conducted in a structured, professional manner, reflecting real-world industry standards.

Final Exam

Each student must be physically present at the exam for its **entire** duration. Exceptions will be accepted only with a medical certificate or after approval from the coaching team.

Deliverables

1) **Commitment agreement** signed at the beginning where it is specified the responsibility of each team member. (sent on time filled: +5 points, not sent -5 points) (**Deadline Sunday 10.3.2024 Midnight**).

2) **Proposal** (sent on time filled: +5 points, not sent -5 points) (deadline Sunday 10.3.2024 Midnight)

3) **Scientific Report** (at the end of the course) (word, latex) with the following sections
    - Literature review (2-3 pages with minimum of 7 sources)
    - Data Processing
    - Methods
    - Model Validation
    - Machine Learning Operations (Deployment) (Maximum of 50 points)

4) Always **up-to-date kanban board** (for every coaching session) that will be presented at each coaching session and discussed. (fulfilled +5 points, not fulfilled -5 points)

5) **GitHub repository (**at the end of the course) (GitHub available: +5 points, not available -5 points). We will not grade directly the content, but we consider the possibility of taking points away if the status of the GitHub repository is abysmal. You are warned.

6) Final Presentation (at the end of the course): This will be also judged by an industry jury (maximum of 30 points) (30% from the jury, 70% from the coaches).

For each of the project requirements **NOT** present in the final project 5 points will be subtracted (so you should check this in advance).
The grade will be given on a total of 100 points distributed as explained above.

Scientific Report

Structuring a report with the outlined sections for a course, particularly one involving machine learning or data science, is essential for several reasons:

1. **Literature Review (2-3 pages with a minimum of 7 sources)**:
    a. **Foundation of Knowledge**: A literature review establishes the theoretical foundation and current state of research in the field. By requiring a minimum number of sources, students are encouraged to engage deeply with existing literature, ensuring a comprehensive understanding of the subject.
    b. **Critical Thinking and Contextualization**: Analyzing and synthesizing various sources enhances critical thinking skills. It allows students to understand different perspectives and place their work within the broader context of the field.
2. **Data Processing**:
    a. **Core Competency in Data Science**: Data processing is a fundamental step in any data science project. Demonstrating this process shows the student's ability to handle and prepare data for analysis, which is a critical skill in the field.

    b. **Transparency and Reproducibility**: Detailing the data processing steps ensures transparency and aids in the reproducibility of the results, which are key aspects of scientific research.
3. **Methods**:
    a. **Understanding and Application**: This section allows students to demonstrate their understanding of various methodologies and their ability to apply appropriate techniques to their specific project.
    b. **Rationale and Justification**: Discussing the methods used provides insight into the student's decision-making process and the rationale behind choosing specific approaches.
4. **Model Validation**:
    a. **Ensuring Model Reliability**: Model validation is crucial for assessing the accuracy and reliability of the model. This section shows how the student evaluates the performance and generalizability of their model.
    b. **Critical Evaluation**: It encourages students to critically evaluate their model's performance, understand its limitations, and discuss potential improvements.
5. **Machine Learning Operations (Deployment)**:
    a. **Practical Application**: This section emphasizes the practical aspect of machine learning. It's not just about building models but also about deploying them effectively in real-world scenarios.
    b. **Bridging Theory and Practice**: It allows students to demonstrate their ability to translate theoretical knowledge into practical applications, showcasing their readiness for industry challenges.

Overall, this structure ensures that the report is comprehensive, covering all critical aspects of a machine learning project from theoretical underpinnings to practical implementation. It not only assesses the student's technical skills but also their ability to conduct thorough research, think critically, and communicate their findings effectively.

Industry Jury

Location and composition

To be announced.

Reading Material

The following material is useful for the course.

[1] Aldabbus, Shaban. (2018). Project-Based learning: Implementation & challenges. International Journal of Education, Learning and Development 6(3), 71-79. https://www.researchgate.net/publication/328368222_PROJECT-BASED_LEARNING_IMPLEMENTATION_CHALLENGES

[2] Condliffe, B., Quint, J., Visher, M. G., Bangser, M. R., Drohojowska, S., Saco, L., & Nelson, E. (2017). Project-Based learning A literature review working paper. https://s3-us-west-1.amazonaws.com/ler/MDRC+PBL+Literature+Review.pdf

[3] Michelucci, U. (2022). Applied Deep Learning with TensorFlow 2: Learn to Implement Advanced Deep Learning Techniques with Python. Apress. https://doi.org/10.1007/978-1-4842-8020-1

[4] TFX Pipelines in TensorFlow https://www.tensorflow.org/tfx/guide/understanding_tfx_pipelines

[5] What is Kanban? Here's what your Agile team needs to know https://asana.com/resources/what-is-kanban (last accessed 22.2.2024)

[6] Scott Chacon, Ben Straub, Pro Git - https://git-scm.com/book/en/v2 (last accessed 22.2.2024)