Explainable AI, 5 credits

Förklarbar artificiell intelligens 5 hp

Second cycle
Main field: Computer Science and Engineering, Second cycle, has only first-cycle course/s as entry requirements (A1N)
Syllabus is adopted by the Research and Education Board (2022-09-09) and is valid for students admitted for the spring semester 2023.

Placement in the Academic System

The course is given as a single subject course.

Prerequisites and Conditions of Admission

Degree of Bachelor of Science in Engineering, Computer Science and Engineering including an independent project 15 credits or Degree of Bachelor of Science with a major in Computer Science and Engineering including an independent project 15 credits or the equivalent of 180 Swedish credit points or 180 ECTS credits at an accredited university. Programming 7.5 credits and Mathematics 7.5 credits including linear algebra. Applicants must have written and verbal command of the English language equivalent to English course 6 in Swedish Upper-Secondary School.

Course Objectives

The goal of the course is that the student develop knowledge and skills in variety of topics in explainable AI (XAI) including: the need for and importance of explaining different AI methods, the taxonomy of XAI, and classical and well-known XAI methods. The student will develop the knowledge in both theoretical and practical terms.

Following successful completion of the course the student should be able to:

Knowledge and understanding

- account for familiarity with different categorization of XAI methods
- account for comprehension of different well-known XAI methods
- diskuss familiarity with different metrics of evaluating XAI methods

Skills and ability

- independently provide explainability by implementing XAI methods for a given AI method
- ability to select relevant XAI method for a given AI method and context
- ability to trade-off between different aspects of XAI such as model performance and explainability

Judgement and approach

- evaluate XAI methods by different properties including precision & fidelity, robustness, uncertainty, and representativeness
- evaluate the quality of AI explanation for human considering properties such as comprehensibility, selectiveness, and contrastivity

Primary Contents

The course covers the following topics:

- Introduction to the multidisciplinary topics of explainable AI, what is XAI, why is it important, plus related terminologies
- Broad taxonomy of XAI methods including Intrinsic vs post hoc, model-specific vs model-agnostic, and local vs global
- Trade-off between accuracy and explainability, human-friendly explanations,
- Intrinsically explainable models including Linear Regression, Logistic Regression, Generalized Linear Model (GLM), Generalized Additive Model (GAM), and Decision Tree.
- XAI methods including, Partial Dependence Plot (PDP), Conformal Prediction, Individual Conditional Expectation (ICE), Feature Importance, Saliency Maps, Local Interpretable Model-Agnostic Explanations (LIME), SHAP, Integrated Gradient (IG)
- Evaluation of explainability

Teaching Formats

Both lectures and lab sessions will be given online. Labs are designed in Python in a way to make the concepts given during the lectures internalized. Videos of the lectures will also be put online via the university's learning platform for self-paced learning.

Teaching is in English and fully online.

Examination

The overall grades of Fail or Pass will be awarded for the course.
Exams will consist of labs and a written exam. The practical tasks from the labs are carried out in Python and presented in the form of Jupyter Notebooks.

Name of the test		**Grading**
Written Examination	2,5 credits	U/G
Practical Assignments	2,5 credits	U/G

If there are special reasons, the examiner may make exceptions from the specified examination format and allow a student to be examined in another way. Special reasons can e.g. be a decision on learning support.

For elite sports students according to Riktlinjer för kombinationen studier och elitidrott vid Högskolan i Halmstad, DNR: L 2018/177, the examiner has the right to decide on an adapted examination component or let the student complete the examination in an alternative way.

Course Evaluation

Course evaluation is part of the course. This evaluation should offer guidance in the future development and planning of the course. Course evaluations should be documented and made available to the students.

Course Literature and Other Study Resources

Molnar, Christoph. *Interpretable Machine Learning*. Leanpub 2019
Online version publicly available at:
`https://christophm.github.io/interpretable-ml-book/scope-of-interpretability.html`

Denis Rothman. *Hands-On Explainable AI (XAI) with Python*. Packt 2020

Research articles on the topic of XAI (to be distributed throughout the course).