EXPLAINABLE ARTIFICIAL INTELLIGENCE

CS • SEMESTER 2021

INSTRUCTOR: DR. ANDRÉ DOS SANTOS
dossantos@ualberta.ca • andreeds.github.io

LECTURES: **YouTube**

Q&A: TBD • **Zoom**

Final: TBD

OFFICE HOURS: **Discord** or **by email appointment**

TEXT: Deep Learning - https://www.deeplearningbook.org/

PREREQUISITES: TBD

Overview

Introducing eXplainable Artificial Intelligence (XAI). Why is it important to pursue explainability in AI models? How XAI can affect performance, privacy, and safety. Terminology on XAI: understandability, comprehensibility, interpretability, explainability, and transparency. How XAI can be measured and its limitations. Inherent explainable XAI versus Ad-hoc explainable XAI. Main XAI models: Linear regression, decision trees, k-nearest neighbors, rule based learners, general additive models. Bayesian, LIME, SHAP, ELI5. XAI applications. Responsible AI. New topics in XAI. 40 hours course.

Proposing Grading scheme

3 assignments	30%
Project	40%
2 quizzes	10%
1 final exam	20%
Instructor Discretion	+5/-5%
Total	**100%**

Lecture Outline

I. **INTRODUCTION** - 10 hours
   - eXplainable AI
   - Terminology
   - XAI goals
   - Levels of transparency
   - Taxonomy

II. **TRANSPARENT ML** - 10 hours
   - Linear/logistic regression
   - Decision Trees
   - K-Nearest Neighbors
   - Rule Based Learners
   - General Additive Models
   - Bayesian

III. **POST-HOC EXPLAINABILITY** - 14 hours
   - Model-agnostic techniques
     - LIME, G-REX, SHAP, ELI5
   - Shallow ML models
     - tree ensembles, SVM
   - Deep ML models
     - DeepRED, DeepLIFT, GradCAM, DGN, Gradient Boosting Trees, RETAIN, DVBFs, SVAE, New Models

IV. **XAI FUTURE** - 3 hours
   - Responsable AI
   - Safe AI
   - New topics - *If time permits*

V. **XAI APPLICATIONS** - 3 hours
   - NLP
   - Images
   - Time Series
   - Others domains - *If time permits*

Note

Classes

1. Classes will be recorded and posted ahead of time. Exercises, Q&A will be done during regular class time and it is when attendance will be taken. Classes will be done using Zoom meetings.
2. Attendance is expected in Q&A lectures. Little time is available to assist those who have missed lectures and watching relevant classes.

Assignments

3. Due dates for assignments will be given with the assignments. All assignments should be submitted to **University Portal**. Late submissions are not accepted.
4. All assignments should follow the directions and designs given by the instructor.
5. The assignments are weighted equally.
6. It is the responsibility of the student to make sure that their submission has been successfully uploaded before the assignment due date. This can be checked by viewing the uploaded files. Email and Hardcopy submissions are not accepted.

Communication

7. Please try to use Discord for most of the communications. Unless it is a topic that needs to be registered through official means, it will always be asked to establish communication with the Instructor by Discord.

Midterm and Final

8. You will write the examinations by typing your answers. You will not require a camera or microphone to write the exam.

9. The exams are an open book where you are allowed to access any type of reference material but you are not allowed to involve any other person in any way.
10. The examination will cover all parts of the course.

Quizzes

11. You will write the quizzes by typing short answers. You will not require a camera or microphone to write the exam. The quizzes are an open book where you are allowed to access any type of reference material
12. On top of the quizzes marks, there will be a competition where correct answers and timing will be translated into competition points. The points will establish a ranking for bonus marks on the class grade.

Project

13. TBD

Academic integrity

14. Please read the sections of the 2020-2021 University of Alberta General Calendar dealing with Academic Regulations.
15. CHEATING WILL NOT BE TOLERATED. Any close resemblances in the submitted assignments or exams will be assumed to be the result of cheating. Copying of assignments is plagiarism. Allowing your segments to be copied will be treated the same as copying. THE CONSEQUENCE OF PLAGIARISM OR ANY OTHER FORM OF CHEATING WILL RANGE FROM A ZERO GRADE, TO FAILURE IN THE CLASS, TO EXPULSION FROM THE UNIVERSITY. Please note that the dean of the faculty will be informed of any such incident, as per university regulations. Refer to the section on Academic Misconduct and Penalties in the General University Calendar.