XAA101: Designing Reliable and Robust AI Systems

Course Outline

Instructor: Dr. Anthony Corso

Last Updated: July 18, 2023

******This course outline may change.******

Schedule and Logistics

Overview

All activities will be held live online. Lectures and assignment reviews will be recorded and the recordings made available shortly after the lecture concludes.

Short coding assignments will provide you practice applying the ideas and techniques covered in the course. To help you with the assignments and materials, Dr. Corso will hold regular office hours. In addition, daily problem sessions will be led by a course facilitator, giving you a chance to ask questions and to collaborate on the projects with other learners.

You will also have access to a slack channel where you can exchange ideas and questions with other learners and the course facilitator.

Upon completing the course, you will receive a digital record of completion. In order to earn the record of completion, you must attend or watch the recordings of **4 out of 5 lectures by Sunday, July 30th**. We will record attendance by daily course surveys sent after lecture.

Daily Schedule

All times are Pacific Time. This schedule may change.

Orientation: Thursday, July 20, 4:00-5:00 pm (recording available)
Course run dates: Monday, July 24, through Friday, July 28th.

- Lectures and Assignment Reviews: 8:30-10:30 a.m. (PT)
- Problem Session: 11:00-12:00 a.m. (PT)*
- Office Hours: 4:00-5:00 p.m. (PT)**
- Submit Assignments: 6:00 p.m. (PT)

*There will be no problem session or office hours on Friday, July 28.
**Dr. Corso will be available for limited additional office hours during the week.

Preparing for the Course

To get the most out of the course, you should have experience with programming in Python and conceptual familiarity with Machine Learning. If you aren't familiar with Machine Learning or

Python or you need to refresh your knowledge, see the end of this document for recommended resources.

Course Outline

These topics may change.

Day 1: Robust Machine Learning

Learning Objectives:

After this lecture, you will be able to…

- Understand the key assumptions modern ML is based on and identify when those assumptions are being violated.
- Describe various types of distribution shifts and their effects on model performance
- Be able to apply several approaches for improving robustness of ML models

Lecture Outline:

- Introduction
    - Use of AI in safety critical applications
    - Failure case studies
- The assumptions of statistical ML
    - The IID (Independent and Identical Distribution) assumption and when it breaks down
    - Overfitting and the use of train/val/test splits
- (Lack of) Robustness beyond the IID
    - Examples of failures of distribution shift robustness
    - Types of distribution shifts
    - Interpretations of the lack of generalization
- Techniques for improving robustness
    - Data augmentation
    - Loss functions to promote subgroup parity and model invariance
    - Fine-tuning from large pre-trained models

Day 2: Model Explanation and Verification

Learning Objectives:

After this lecture, you will be able to…

- Understand and articulate the limits of single metrics of performance
- Understand the various types of explainable ML and implement at least one black-box method
- Understand the core concepts in neural network verification

Lecture Outline:

- Why go beyond test set accuracy?
    - Explainability can help build trust in models and provide insights
    - When designing for extreme reliability, test sets aren't big enough or diverse enough
- Explainability in Machine Learning
    - Goals of explainability
    - Categorization of explainability methods
    - Algorithms for post-hoc explainability
    - Shortfalls of explainability methods
- Verification
    - A motivating example: Adversarial Robustness
    - Neural network verification (methods and tools)
    - Open challenges for formal approaches

Day 3: Uncertainty Quantification

Learning Objectives:

After this lecture, you will be able to…

- Articulate the difference between aleatoric and epistemic uncertainty
- Quantify the quality of uncertainty estimates using common metrics
- Apply common algorithms for estimating aleatoric and epistemic uncertainty

Lecture Outline:

- Aleatoric vs. Epistemic Uncertainty
    - Uncertainty that is present in the data
    - Uncertainty due to unknown unknowns
- Aleatoric Uncertainty Quantification
    - Modeling distributions to estimate uncertainty

    - Quantifying the quality of uncertainty estimates through calibration
    - Algorithms for uncertainty quantification
        - Temperature Scaling
        - Conformal Prediction
- Epistemic Uncertainty Quantification
    - Algorithms for modeling epistemic uncertainty
        - Bayesian Neural Networks
        - Scalable Approximations: Dropout and ensembling
    - Out of model scope vs. out of distribution

Day 4: Closed-Loop Analysis

Learning Objectives:

After this lecture, you will be able to…

- Understand the importance of closed-loop analysis
- Understand the elements required to analyze a closed-loop system
- Understand and apply approaches for failure discovery and probability of failure estimation

Lecture Outline:

- The importance of closing the loop
    - Why we need to consider the downstream effects of the model decision
- How to model the closed loop system
    - Defining requirements
    - Building a prototype
    - Modeling the environment
    - Analyzing the property
- The problem of rare events
    - Falsification
    - Most-likely failure search
    - Probability of failure estimation

Guest Lecture:

Guest Speaker: Mykel Kochenderfer, Associate Professor of Aeronautics and Astronautics and, by courtesy, of Computer Science.

Time: 10:00 - 11:00 am (the guest lecture will be recorded)

Topic: TBD

Day 5: Emerging Directions in AI Safety

Learning Objectives:

After this lecture, you will be able to…

- Understand the need for a diverse and robust safety case involving multiple types of validation and verification
- Understand the landscape of upcoming AI regulation and how the technical approaches we've discussed may play a role in future certification processes

Lecture Outline:

- What is the relationship between all of these ingredients
    - Models of safe design
    - The swiss-cheese model of safety
- Upcoming AI Regulation
    - EU AI Act
    - US AI legislation
- Open research directions in AI safety

Slack

The slack workspace <xaa101-scpd.slack.com> is for you to network and ask questions during the quarter. It is available to students who are actively enrolled in a course. You will receive an invitation to join shortly before the course begins.

- Use the "networking" channel to introduce yourself and meet other students.
- Use the "logistics" channel for logistics related questions for staff.
- Use the "questions" for content related questions.

Although staff and the course teaching team will monitor the channels between 9:00 am and 5:00 pm pacific time during the course, we also encourage you to answer each other's questions.

We encourage you to use these channels to form study groups, link up with students who are pursuing similar courses or programs, and in general to share insights into your experiences with your fellow students.

Please use good judgment. Keep interactions professional and focused on coursework or career networking. Avoid using offensive language and respect your colleagues' preferences regarding direct messaging. Please respect and uphold the rights and dignity of others regardless of race,

color, national or ethnic origin, sex, age, disability, religion, sexual orientation, gender identity, or socio-economic status. Our team is always available either here in Slack or via email, so please feel free to reach out to us if you have any questions or concerns, or if any situation arises.

You can review SCPD's terms of service here, including rules for online conduct.

Review Materials

If you need a refresher to prepare, you should consider the following self-study resources. This list includes resources that other learners have found useful. Their inclusion does not imply any specific endorsement by SCPD.

If time is short, we recommend reviewing the Deep Learning and Probability materials starred** below.

Overview and Essentials

- CS221: Artificial Intelligence: Principles and Techniques (Course Website)
- CS229: Machine Learning (Course Website)
- DeepLearning.AI: Deep Learning**
- EE104: Introduction to Machine Learning (Course Website)

Mathematics

- Coursera: Mathematics for Machine Learning series (Imperial College London)
- Stanford CS229 Linear Algebra Review and Reference
- MIT Single Variable Calculus
- MIT Multivariable Calculus
- MIT Linear Algebra
- MIT Mathematics for Machine Learning
- The Matrix Calculus You Need For Deep Learning
- CS221 pre-recorded materials
- ENGR108: Introduction to Applied Linear Algebra (Textbook, Course Website)

Statistics and Probability

- Stanford CS229 Review of Probability Theory**
- Stanford CS229 Statistics and Probability Refresher
- CS109 Course Reader: Probability for Computer Scientists
- MIT Probabilistic Systems Analysis and Applied Probability (2010 version here)
- STATS216: Statistical Learning

Python

- Introduction to Computer Science and Programming Using Python
- Coursera: Python 3 Programming Specialization
- Coursera: Applied Data Science with Python Specialization
- Stanford CS231n Python/Numpy Tutorial

If you are interested in self-study, you can also find the lectures from multiple Stanford courses available at Stanford Online's YouTube channel.