LAMBDA Lab Presents
Seminar on
Explainable Artificial Intelligence
with Python

Duration: 4 Weeks (2 Sessions Per week)
Contact Hours: 16 Hours
Tentative Start Date: 25th Oct 2021

Mode of Communication: English
Specially Designed for Students and Researchers who are interested to apply XAI in their own Research Area
Platform: Zoom

Course Instructor
Dr. Parteek Bhatia
Visiting Professor
LAMBDA Lab, Tel Aviv University, Israel
www.parteekbhatia.com
Explainable Artificial Intelligence with Python

Seminar overview: This seminar provides a broad introduction to latest developments in the field of Explainable Artificial Intelligence and its application in various research domains. As our reliance on AI models is increasing day by day, it's also becoming equally important to explain how and why AI is making a particular decision. Recent laws have also caused the urgency about explaining and defending the decisions made by AI systems.

In this seminar, you will learn about tools and techniques using Python to visualize, explain, and build trustworthy AI systems. We will discuss all-important XAI techniques like LIME, SHAP, DiCE, LRP, Contrastive and Counterfactual Explanations Method in this course. You will be introduced to several open-source explainable AI tools for Python that can be used throughout the machine learning project lifecycle.

Who is the seminar for?
This seminar is designed for the learners who are:
- Beginner Python programmers who already have some foundational knowledge with machine learning libraries.
- Researchers who already use Python for building AI models and can benefit from learning the latest explainable AI open-source toolkits and techniques.
- Data analysts and data scientists that want an introduction to explainable AI tools and techniques using Python for machine learning models.

Interested researchers and students
Please register by 25th Oct 2021
For Registration, please fill following form
Or visit www.parteekbhatia.com/XAI
Scan QR Code to Register
## Tentative Detailed Contents and Plan

<table>
<thead>
<tr>
<th>Meeting-1</th>
<th>Introduction to Explaining Artificial Intelligence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Need and Applications of Explaining Artificial Intelligence</td>
</tr>
<tr>
<td></td>
<td>• Categorization of XAI</td>
</tr>
<tr>
<td></td>
<td>• Various Case Studies XAI</td>
</tr>
<tr>
<td></td>
<td>• Overview of various techniques for XAI</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Meeting-2</th>
<th>XAI through Local Interpretable Model-Agnostic Explanations (LIME)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Working Principle of LIME</td>
</tr>
<tr>
<td></td>
<td>• Understanding Mathematical representation of LIME</td>
</tr>
<tr>
<td></td>
<td>• Applications of LIME in various Case studies</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Meeting-3 (One hour of Demonstration and one hour of Self Practice)</th>
<th>Implementing LIME over various Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Case Study-1: Applying LIME over Stroke/No Stroke Health Dataset</td>
</tr>
<tr>
<td></td>
<td>Case Study-2: Applying LIME over Newsgroup Dataset</td>
</tr>
<tr>
<td></td>
<td>Through these case studies following topics will be covered:</td>
</tr>
<tr>
<td></td>
<td>• Getting started with LIME</td>
</tr>
<tr>
<td></td>
<td>• An experimental AutoML module</td>
</tr>
<tr>
<td></td>
<td>• Interpreting the scores</td>
</tr>
<tr>
<td></td>
<td>• Training the model and making predictions</td>
</tr>
<tr>
<td></td>
<td>• Creating the LIME explainer</td>
</tr>
<tr>
<td></td>
<td>• Interpreting LIME explanations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Meeting-4</th>
<th>XAI through SHapley Additive exPlanations (SHAP)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Working Principle of SHAP</td>
</tr>
<tr>
<td></td>
<td>• Key SHAP Principles Symmetry, Null Player and Additivity</td>
</tr>
<tr>
<td></td>
<td>• Understanding Mathematical representation of SHAP</td>
</tr>
<tr>
<td></td>
<td>• Applications of SHAP in various Case studies</td>
</tr>
</tbody>
</table>
Research Directions for XAI By
Prof. Irad E. Ben-Gal Prof. and Head, Laboratory of
AI Machine Learning Business & Data Analytics
(LAMBDA)

Implementing SHAP over various Datasets—Parteek
Bhatia

**Case Study-1:**
Applying SHAP over Stroke/No Stroke Health Dataset

**Case Study-2:**
Applying SHAP for understanding results of Sentiment
Analyser

Through these case studies following topics will be covered:

- Installing SHAP
- Intercepting the dataset
- Vectorizing the datasets
- Creating, training, and visualizing the output of a linear
  model
- Agnostic model explaining with SHAP
- Explaining the original IMDb reviews with SHAP

**Assignment-1**

**Objective:**
To apply LIME and SHAP on various classifiers for
XAI

- Applying LIME over mushroom dataset to explain
  prediction that is edible or poisonous.
- Applying LIME and SHAP over MNIST Digit
  classification and to perform comparative analysis of
  both techniques for XAI.

References

https://github.com/marcotcr/lime
https://github.com/slundberg/shap

**Meeting-6**

**AI Fairness with What-If Tool (WIT) and
Counterfactual Explanations Method for XAI**

- Understanding Fairness in AI
• Demonstration of What-If Tool (WIT) for COMPAS dataset
• Working Principle of Counterfactual Explanations Method
• Understanding Mathematical representation of Counterfactual
• Diverse Counterfactual Explanations (DiCE)
• Concept of Belief, Truth, Justification and Sensitivity
• Understanding various distance functions

**Case Study-1**
Applying Counterfactual Explanations over CelebA dataset for identification of Smile
Through these case studies following topics will be covered:
• Installing DICE
• WIT datapoint explorer and editor
• Visualizing counterfactual distances in WIT
  Exploring various data point distances

**Meeting-7**

**XAI for Neural Networks with Layer wise Relevance Propagation (LRP) and Its Implementing**
• Working Principle of LRP
• Understanding Mathematical representation of LRP

**Case Study**
Applying LRP over Brain MRI dataset
Through this case study following topics will be covered:
• Loading the dataset
• Pre-processing the dataset
• Building VGG16 Model for identification of tumour
• Layerwise relevance propagation for VGG16
• Calculating relevance for images

**Assignment-2**
**Objective:** Compare income classification on UCI census data binary classification model comparison

**DATA SOURCE**
To demonstrate the use of What-If Tool for XAI.

**UCI Census Income Dataset**
Compare two binary classification models that predict whether a person earns more than $50k a year, based on their census information. Examine how different features affect each models' prediction, in relation to each other.

**Text toxicity classifiers**
- binary classification
- model comparison
- keras
- model
- custom distance

**DATA SOURCE**
**Wikipedia Comments Dataset**
Use the What-If Tool to compare two pre-trained models from ConversationAI that determine sentence toxicity, one of which was trained on a more balanced dataset. Examine their performance side-by-side on the Wikipedia Comments dataset. These are keras models which do not use TensorFlow examples as an input format.

---

**Meeting-8**

**Concept of Contrastive XAI, Cognitive XAI and Future Research Directions**
- Understand the working principle of Contrastive Explanations Method (CEM)
- Concept of Cognitive XAI Explanations
- Discussion of Future Directions for XAI
Interested researchers and students

Please register by 15\textsuperscript{th} Oct 2021

For Registration, please fill following form


Or Visit

\url{www.parteekbhatia.com/XAI}

Scan QR Code to Register