**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI**

**WORK INTEGRATED LEARNING PROGRAMMES**

**Digital**

Part A: Content Design

| **Course Title** | FAccT Machine Leaning |
| --- | --- |
| **Course No(s)** |  |
| **Credit Units** | 4 |
| **Credit Model** | 1 - 0.5 - 1.5  1 unit for class room hours, 0.5 unit for Tutorial, 1.5 units for  Student preparation.  1 unit = 32 hours |
| **Content Authors** | Dr. Sugata Ghosal |
| **Version** | 1.0 |
| **Date** | January 11th, 2023 |

Course Description

Machine Learning systems have demonstrated remarkable learning capabilities. A growing area in machine learning focuses on improving the Fairness, Accountability, and Transparency (FAccT) of a model in addition to its accuracy-based performance metrics. Although FAccT will be difficult to achieve, emerging technical approaches in this topic show promise in making better FAccT ML systems. In this course, we will study the rigorous computer science techniques necessary for FAccT machine learning and dive into the technical underpinnings of topics including fairness, robustness, interpretability, accountability, and privacy. These topics reflect state-of-the-art research in FAccT, are socially important, and they have strong industrial interest due to government and other policy regulation. This course will focus on the algorithmic and statistical aspects of FAccT ML. We will also discuss several application areas where we can apply these techniques. This course requires students to have mathematical and programming backgrounds in machine learning.

Topics

Biases and fairness, fair representation learning, Interpretability and Transparency, Example and Visualization Based Methods for Interpretability, Interpreting deep neural networks, Fairness Through Input Manipulation, Fair NLP/Vision, Robustness and adversarial attacks/defence, ML auditing, privacy

**Course Objectives**

| **No** |  |
| --- | --- |
| **CO1** | Introduce students to the concepts of bias and fairness and techniques for incorporating these in ML |
| **CO2** | Introduce students to the concepts of interpretability and transparency and techniques for incorporating these in ML |
| **CO3** | Introduce students to the concepts of robustness and techniques for robust ML |
| **CO4** | Introduce students to the concepts of privacy in ML |

**Text Book(s)**

| T1 | Barocas, Solon, Moritz Hardt, and Arvind Narayanan. [Fairness and Machine Learning](https://fairmlbook.org/), 2018. |
| --- | --- |
| T2 | Molnar, Christoph. [Interpretable machine learning](https://christophm.github.io/interpretable-ml-book/), 2019. |

**Reference Book(s) & other resources**

| R1 | Christopher M. Bishop, Pattern Recognition & Machine Learning, Springer, 2006 |
| --- | --- |
| R2 | https://www.borealisai.com/research-blogs/tutotial\* |
| R3 |  |

**Content Structure**

1. **Introduction**
   1. Review - traditional performance criteria for ML
   2. Emerging requirements
      1. Fairness
      2. Accountability
      3. Transparency
      4. Privacy
      5. Robustness
   3. Motivating use cases

1. **Fairness and Bias** 
   1. Sources of Bias
   2. Real world examples
      1. school admissions, criminal justice, hiring, gender/occupation bias
   3. Sensitive Features
   4. Fairness through unawareness
2. **Learning Fair Representations**
   1. Major Fairness criteria
      1. Direct Solution Method
      2. Demographic Parity
      3. Equality of Odds/Opportunity
      4. FICO Case Study
   2. Prejudice Removing Regularizer
      1. Prejudice index (PI)
      2. Optimizing PI
   3. Adult Income Case Study
3. **Fairness thru input manipulation**
   1. Basic Data Manipulation Techniques
      1. Reweighing
      2. Universal Sampling
      3. Preferential Sampling
   2. Individual Fairness
   3. Optimized Pre-processing
   4. Learning to Defer
4. **Fair Casual Reasoning**
   1. Causal Fairness and Inherent Bias
   2. Counterfactual Fairness
      1. Formal Methods
      2. Case Study – Success in Law School
      3. Case Study - Crime Rates and Arrest
   3. Equalized Counterfactual Odds
   4. Multiple Causal Worlds

1. **Fairness in NLP, Computer Vision**
   1. Biases in NLP Models
   2. Data Augmentation
   3. Debiasing Word Embedding
   4. Counterfactual Fairness
   5. Counterfactual Face Attribution
   6. Gender Equalized Image Captioning
   7. Adversarial Removal of Gender Features
2. **Interpretability and Transparency** 
   1. ML Interpretability
   2. Intrinsically Interpretable Models
      1. Simple interpretable models
      2. Intrinsically interpretable techniques for deep learning
   3. Interpretability Concepts
      1. Intrinsic and post hoc methods
      2. model-specific and model-agnostic methods
      3. Local and global interpretable methods
   4. Interpretability and performance trade-offs Instance-based Learning
3. **Feature interaction for interpretability**
   1. Feature Interaction
   2. Layerwise Relevance Propagation
   3. DeepLift
   4. Shapley Additive Explanations (SHAP)
      1. Coalitional Game and Shapley Values
      2. Kernel SHAP
      3. Deep SHAP
      4. Tree SHAP
   5. Equitable Value of Data
4. **Example and Visualization Based Methods for Interpretability**
   1. Example Based Methods
   2. Counterfactual Explanations
      1. Targeted counterfactual examples
      2. Untargeted counterfactual examples
   3. Contrastive Examples
   4. Concept Based Methods
5. **Interpreting Deep Networks**
   1. Visualization Based Methods
   2. Activation Visualization
      1. Saliency Maps
      2. GoogLeNet Activation Atlas
      3. Interpretability via Activation Visualization
   3. Gradient Based Feature Attribution ○
      1. Integrated Gradient
      2. Baselines for Integrated Gradient
6. **Robustness and Adversarial Attacks / Defense**
   1. Adversarial Attack
   2. White-box Evasion Attack
      1. FGSM
      2. C&W
      3. Physical Attack
   3. Transferability of Attack
   4. Black-box Evasion Attack
      1. Jacobian-based Data Augmentation
   5. Adversarial Defense
   6. Defense Strategies
      1. Adversarial Training
      2. Input Transformations
      3. Stochastic Gradients
      4. Obfuscated Gradients and BPDA
   7. Robust Optimization
   8. Certified Defense
7. **ML Auditing and Privacy**
   1. ML Auditing
      1. Distill-and-Compare
   2. Privacy in ML
      1. Differential Privacy with Deep Learning
      2. Model Inversion Attack and Differential Privacy
      3. Local Differential Privacy
      4. Federated Learning

**Learning Outcomes:**

| No | Learning Outcomes |
| --- | --- |
| LO1 | A strong understanding of the foundations of Machine Learning algorithms |
| LO2 | Able to solve Machine Learning problems using appropriate learning techniques |
| LO3 | Evaluate machine learning solutions to problems |
| LO4 | Identify appropriate tools to implement the solutions to machine learning problems |

**Part B: Learning Plan**

| **Academic Term** |  |
| --- | --- |
| **Course Title** | Machine Learning |
| **Course No** | ZG 565 |
| **Lead Instructor** | Dr. Sugata Ghosal |

| Session No. | Topic Title | Study/HW Resource Reference |
| --- | --- | --- |
| 1 | **Introduction**  Review - traditional performance criteria for ML; Emerging requirements such as Fairness, Accountability, Transparency, Privacy, Robustness, Motivating use cases | T1 – Ch1 |
|
|
| 2 | **Fairness and Bias**  Sources of Bias, Real world examples such as school admissions, criminal justice, hiring, gender/occupation bias, Sensitive Features, Fairness through unawareness |  |
| 3 | **Learning Fair Representations**  Major Fairness criteria such as Demographic Parity, Equality of Odds/Opportunity, FICO Case Study, , Prejudice Removing Regularizer, Adult Income Case Study |  |
| 4 | **Fairness thru input manipulation**  Basic Data Manipulation Techniques, such as Reweighing, Universal Sampling, Preferential Sampling, Individual Fairness, Optimized Pre-processing, Learning to Defer |  |
| 5 | **Fair Casual Reasoning**  Causal Fairness and Inherent Bias, Counterfactual Fairness, Case Study – Success in Law School, Crime Rates and Arrest, Equalized Counterfactual Odds, Multiple Causal Worlds |  |
| 6 | **Fairness in NLP, Computer Vision**  Biases in NLP Models. Data Augmentation, Debiasing Word Embedding, Counterfactual Fairness, Counterfactual Face Attribution, Gender Equalized Image Captioning, Adversarial Removal of Gender Features |  |
| 7 | **Disentangled Fair Representation**  Disentangled Representations, Flexibly Fair Representation, Orthogonal Disentangled Fair Representations, Measurements for Disentangled Fair Representations |  |
| 8 | Review of Session 1 to 7 | Books, Web references and Slides |
| 9 | **Interpretability and Transparency**  ML Interpretability, Intrinsically Interpretable Models, Interpretability Concepts such as Intrinsic and post hoc methods, model-specific and model-agnostic methods, Local and global interpretable methods, Interpretability and performance trade-offs Instance-based Learning |  |
| 10 | **Feature interaction for interpretability**  Feature Interaction such as Layerwise Relevance Propagation, DeepLift, Shapley Additive Explanations (SHAP) and variants |  |
| 11 | **Example and Visualization Based Methods for Interpretability**  Example Based Methods**,** Counterfactual Explanations, Contrastive Examples**,** Concept Based Methods |  |
| 12 | **Interpreting Deep Networks**  Visualization Based Methods, Activation Visualization, saliency maps, Interpretability via Activation Visualization, Gradient Based Feature Attribution ○ |  |
| 13 | **Robustness and Adversarial Attacks / Defense**  Adversarial Attack, White-box Evasion Attack, Transferability of Attack, Black-box Evasion Attack |  |
| 14 | **Robustness and Adversarial Attacks / Defense**  (contd.)  Adversarial Defense, Defense Strategies such as Adversarial Training, Input Transformations, Stochastic Gradients, Obfuscated Gradients  **ML Auditing and Privacy**  ML Auditing, Distill-and-Compare, Privacy in ML, |  |
| 15 | **ML Auditing and Privacy (contd)**  Differential Privacy with Deep Learning, Model Inversion Attack and Differential Privacy, Local Differential Privacy, Federated Learning |  |
| 16 | Review of session 9 to 15 | Books, Web references and Slides |

**Detailed Plan for Lab work**

| **Lab No.** | **Lab Objective** | **Lab Sheet Access URL** | **Session Reference** |
| --- | --- | --- | --- |
| 1 | Bias checking and mitigation using AIF360 |  |  |
| 2 | Assessment and Mitigation of unfairness using FairLearn |  |  |
| 3 | Assessing the Fairness of a Classifier using Prejudice Removal Regularizer |  |  |
| 4 | Implementing LIME and SHAP algorithms to interpret different machine learning classifiers |  |  |
| 5 | Visualizing and interpreting a CNN based Image classifier |  |  |
| 6 | Implementing Adversarial attacks on CNN based Image classifier |  |  |

**Evaluation Scheme**:

Legend: EC = Evaluation Component; AN = After Noon Session; FN = Fore Noon Session

| No | Name | Type | Duration | Weight | Day, Date, Session, Time |
| --- | --- | --- | --- | --- | --- |
| EC-1 | Quiz – Best 2 out of 3 | Online | ~1 hour | 10% |  |
|  | Assignment-I | Take Home | ~2-3 weeks | 10% |  |
|  | Assignment-II | Take Home | ~2-3 weeks | 10% |  |
| EC-2 | Mid-Semester Test | Open Book |  | 30% |  |
| EC-3 | Comprehensive Exam | Open Book |  | 40% |  |

**8310758441**

**Note:**

Syllabus for Mid-Semester Test (Open Book): Topics in Session Nos. 1 to 8

Syllabus for Comprehensive Exam (Open Book): All topics (Session Nos. 1 to 16)

**Important links and information:**

Elearn portal: <https://elearn.bits-pilani.ac.in> or Canvas

Students are expected to visit the Elearn portal on a regular basis and stay up to date with the latest announcements and deadlines.

Contact sessions: Students should attend the online lectures as per the schedule provided on the Elearn portal.

Evaluation Guidelines:

1. EC-1 consists of either two Assignments or three Quizzes. Students will attempt them through the course pages on the Elearn portal. Announcements will be made on the portal, in a timely manner.
2. For Closed Book tests: No books or reference material of any kind will be permitted.
3. For Open Book exams: Use of books and any printed / written reference material (filed or bound) is permitted. However, loose sheets of paper will not be allowed. Use of calculators is permitted in all exams. Laptops/Mobiles of any kind are not allowed. Exchange of any material is not allowed.
4. If a student is unable to appear for the Regular Test/Exam due to genuine exigencies, the student should follow the procedure to apply for the Make-Up Test/Exam which will be made available on the Elearn portal. The Make-Up Test/Exam will be conducted only at selected exam centres on the dates to be announced later.

It shall be the responsibility of the individual student to be regular in maintaining the self study schedule as given in the course handout, attend the online lectures, and take all the prescribed evaluation components such as Assignment/Quiz, Mid-Semester Test and Comprehensive Exam according to the evaluation scheme provided in the handout.