(http://stanford.edu/)

CS 329T: Trustworthy Machine Learning

Stanford, Spring 2021

Schedule & syllabus

The lecture slides,abs, and assignments will be posted online here as the course progresses. All the pre-recorded lectures would be uploaded Monday every week on Canvas.
Lecture times are **2:30-3:50pm PST**. All deadlines are at **11:59pm PST**.

This schedule is subject to change according to the pace of the class.

Date:
Description:
Part I: Background (Week 1)
Events:

Date: **Mon Mar 29**
Description: **Week 1 Presentation topics:** **Course overview** **Background: Deep learning** **Background: Vision** **Background: Keras**
Slides
Events: Pre-recorded lecture

Date: **Tue Mar 30**
Description: **Orientation, overview Fireside chat, course QA and introduce final project**
Slides (slides/part1-week1-video01-overview.pdf)
Events: Fireside chat Lecture

Date: Thu Apr 1

Description: Troubleshooting Homework 0
Intro to Homework 1

Slides (slides/CS329T_Lab1.pdf)

Events: Lab

Homework 1 Released:
[pdf (homeworks/hw1/CS329T_HW1.pdf)]
[Code (homeworks/hw1/CS329T_HW1_Code.zip)]
[Written Template (homeworks/hw1/CS329T_HW1_Written.zip)]

Description: Homework 1 is designed to make sure you are comfortable with ML fundamentals that will be needed in this course. If you are struggling with parts of this assignment, consider whether you meet the prerequisites.

Learning outcomes: Background checkpoint

Content: XGboost, Python, Sci-kit learn, Tensorflow for vision

Date:

Description:

Part II: Explanations (Weeks 2 and 3)

Events:

Date: Mon Apr 5

Description: Week 2 Presentation topics:
Explanations overview
Local explanations
Input importance and Shapley values

An Evaluation of the Human-Interpretability of Explanation
(https://finale.seas.harvard.edu/files/finale/files/an_evaluation_of_the_human-interpretability_of_explanation.pdf)
Why Should I Trust You?": Explaining the Predictions of Any Classifier
(https://dl.acm.org/doi/pdf/10.1145/2939672.2939778)
Axiomatic Attribution for Deep Networks (https://arxiv.org/pdf/1703.01365.pdf)

Events: Pre-recorded lecture

Date: Tue Apr 6

Description: Shapley values in explanations: SHAP & QII

Slides (slides/part1-week2-video01-explanations.pdf)
Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems
(https://www.andrew.cmu.edu/user/danupam/datta-sen-zick-oakland16.pdf)
A Unified Approach to Interpreting Model Predictions
(https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf)<

Events: Fireside chat Lecture

Date: Thu Apr 8

Description: Intro to Homework 2

Slides (slides/CS329T_Lab2.pdf)

Events: Lab

Date: Fri Apr 9

Description: Homework 1 due

Events:

Date: Sat Apr 10

Description: Homework 2

Events: Homework 2 Released:
[pdf (homeworks/hw2/CS329T_HW2.pdf)]
[Code (homeworks/hw2/CS329T_HW2_Code.zip)]
[Written Template (homeworks/hw2/CS329T_HW2_Written.zip)]

Date: Mon Apr 12

Description: Week 3 Presentation topics:

- Vision attributions (saliency maps, integrated gradients, layerwise relevant propagation, etc.)
- Evaluations for attributions
- Training point influence

Slides
Interpreting Interpretations: Organizing Attribution Methods by Criteria (https://arxiv.org/pdf/2002.07985.pdf)
Representer point selection for DNN (https://papers.nips.cc/paper/2018/file/8a7129b8f3edd95b7d969dfc2c8e9d9d-Paper.pdf)
Understanding Black-box Predictions via Influence Functions (https://arxiv.org/pdf/1703.04730.pdf)

Events: Pre-recorded lecture

Date: Tue Apr 13

Description: More deep learning introspection methods

Slides (slides/explanations-Week2.pdf)

Towards Automatic Concept-based Explanations (https://arxiv.org/pdf/1902.03129.pdf)
Influence-Directed Explanations for CNNs (https://arxiv.org/abs/1802.03788)

Events: Fireside chat Lecture

Date: Thu Apr 15

Description: Homework 2 Q/A

Slides (slides/CS329T_Lab3.pdf)

Events: Lab

Date:

Description:

Part III: Fairness (Weeks 4 and 5)

Events:

Date: Mon Apr 19

Description: Week 4 Presentation topics:
Fairness overview
Mitigation in Data
Individual Fairness

Slides
Big Data's Disparate Impact (https://pdfs.semanticscholar.org/1d17/4f0e3c391368d0f3384a144a6c7487f2a143.pdf?_ga=2.198712170.499045504.1611253703-113508275.1611253703)
Certifying and Eliminating Disparate Impact (https://arxiv.org/pdf/1412.3756v3.pdf)
Fairness through Awareness (http://www.cs.toronto.edu/~zemel/documents/fairAwareItcs2012.pdf)

Events: Pre-recorded lecture

Date: Tue Apr 20

Description:
How fair do we need to be? Disparate impact/connections to legal sector
Problems with measuring fairness in the real world

Slides
Certifying and removing disparate impact (https://arxiv.org/abs/1412.3756)
The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning (https://arxiv.org/pdf/1808.00023.pdf)

Events: Fireside chat Lecture

Date: Thu Apr 22

Description: Intro to Homework 3
TBD

Slides

Events: Lab

Date: Mon Apr 26

Description: Week 5 Presentation topics:
Mitigation with Adversarial Learning
Bias in NLP: Embeddings
Bias in NLP: Beyond embeddings

Slides
Mitigation with Adversarial Learning (https://arxiv.org/abs/1801.07593)
Man is to Computer Programmer as Woman is to Homemaker? (http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf)
Gender Bias in Neural Natural Language Processing (https://arxiv.org/abs/1807.11714)

Events: Pre-recorded lecture

Date: Tue Apr 27

Description: Ethical implications, bias in non-language settings

Slides
Human-like Bias in Language Models (http://opus.bath.ac.uk/55288/4/CaliskanEtAl_authors_full.pdf)
Understanding bias in facial recognition technologies (https://arxiv.org/ftp/arxiv/papers/2010/2010.07023.pdf)

Events: Fireside chat Lecture

Date: Thu Apr 29

Description: Homework 3 Q/A
TBD

Slides

Events: Lab

Date:

Description:

Part IV: Privacy (Weeks 6 and 7)

Events:

Date: Mon May 3

Description: Week 6 Presentation topics:

- Privacy overview
- Membership inference
- Model inversion

Slides

Use Privacy in Data-Driven Systems: Theory and Experiments with Machine Learnt Programs (http://arxiv.org/pdf/1705.07807.pdf)

Membership Inference Attacks Against Machine Learning Models (https://www.comp.nus.edu.sg/~reza/files/Shokri-SP2017.pdf)

Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures (https://dl.acm.org/doi/pdf/10.1145/2810103.2813677)

Events: Pre-recorded lecture

Date: Tue May 4

Description: White-box vs Black-box: Bayes Optimal Strategies for Membership Inference

Slides

White-box vs Black-box: Bayes Optimal Strategies for Membership Inference (http://proceedings.mlr.press/v97/sablayrolles19a/sablayrolles19a.pdf)

Events: Fireside chat Lecture

Date: Thu May 6

Description: Intro to Homework 4
TBD

Slides

Events: Lab

Date: Mon May 10

Description: Week 7 Presentation topics:

- Location privacy
- Federated learning
- Privacy and Explanations

Slides

Quantifying Location Privacy (https://core.ac.uk/download/pdf/9713419.pdf)

Comprehensive Privacy Analysis of Deep Learning: Stand-alone and Federated Learning under Passive and Active White-box Inference Attacks (https://arxiv.org/pdf/1812.00910)

On the Privacy Risks of Model Explanations (https://arxiv.org/pdf/1907.00164.pdf)

Events: Pre-recorded lecture

Date: Tue May 11

Description:

Differential Privacy: A Survey of Results
No Free Lunch in Data Privacy

Slides

Differential Privacy: A Survey of Results (https://link.springer.com/chapter/10.1007/978-3-540-79228-4_1)

No Free Lunch in Data Privacy (http://www.cse.psu.edu/~duk17/papers/nflprivacy.pdf)

Events: Fireside chat Lecture

Date: Thu May 13

Description: Homework 4 Q/A

TBD

Slides

Events: Lab

Date:

Description:

Part V: Robustness (Weeks 8 and 9)

Events:

Date: Mon May 17

Description: Week 8 Presentation topics:

Robustness overview

Adversarial attacks

Real-world adversarial attacks

Slides

The Limitations of DL in Adversarial Settings (https://arxiv.org/pdf/1511.07528.pdf)

Towards Evaluating the Robustness of Neural Networks (https://arxiv.org/pdf/1608.04644)

DReal and Stealthy Attacks on State-of-the-Art Face Recognition (https://www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf)

Events: Pre-recorded lecture

Date: Tue May 18

Description:

Adversarial Examples Are Not Bugs, They Are Features

How does adversarial robustness play a role in model explainability (to be discussed further in next week's Presentation topics)?

Slides

Adversarial Examples Are Not Bugs, They Are Features (https://arxiv.org/pdf/1905.02175.pdf)

Events: Fireside chat Lecture

Date: Thu May 20

Description: Intro to Homework 5
Implement basic attacks for small models

Slides

Events: Lab

Date: Mon May 24

Description: Week 9 Presentation topics:
Adversarial defenses
Attacks on attributions
Defenses against attacks on attributions

Slides
Towards Deep Learning Models Resistant to Adversarial Attacks (https://arxiv.org/pdf/1706.06083.pdf)
Explanations can be manipulated and geometry is to blame (https://arxiv.org/pdf/1710.10547.pdf)
Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing their Input Gradients (https://arxiv.org/abs/1711.09404)

Events: Pre-recorded lecture

Date: Tue May 25

Description:
How to certify robustness?
Fast Geometric Projections for Local Robustness Certification
Non-deep net adversarial attacks on explanations?
Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods

Slides
Fast Geometric Projections for Local Robustness Certification (https://arxiv.org/abs/2002.04742)
Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods (https://arxiv.org/abs/1911.02508)

Events: Fireside chat Lecture

Date: Thu May 27

Description: Homework 5 Q/A
TBD

Slides

Events: Lab

Date:

Description:

Part VI: Synthesis and Takeaways (Week 10)

Events:

Date: Tue Jun 1
Description: Final assignment presentations
Events: Fireside chat Lecture

Date: Thu Jun 3
Description: Final assignment presentations
Slides
Events: Lab