CS 4501/6501 Interpretable Machine Learning

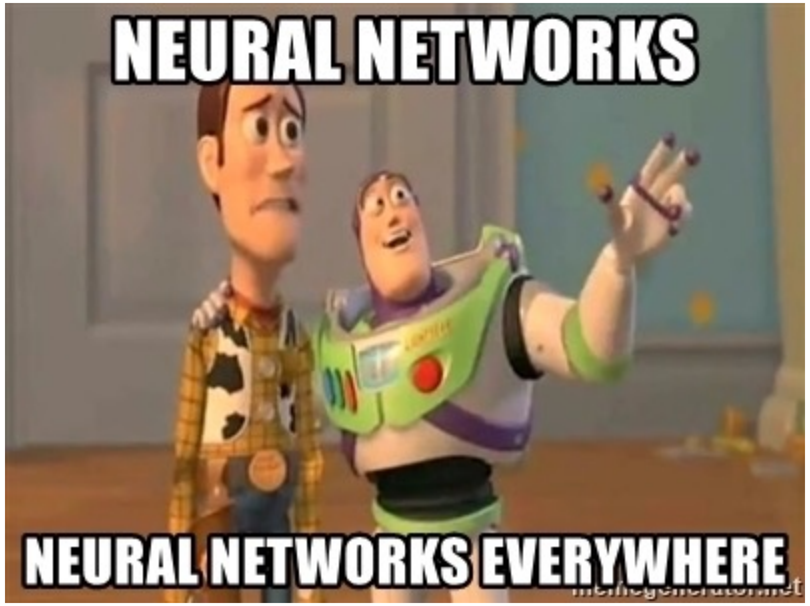
Introduction

Hanjie Chen, Yangfeng Ji
Department of Computer Science
University of Virginia
{hc9mx, yangfeng}@virginia.edu

Interpretable Machine Learning

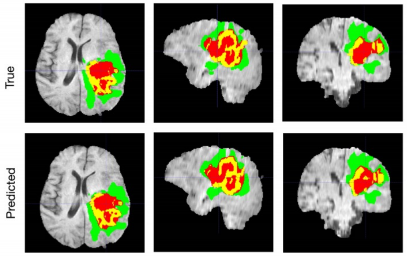
Video source: https://www.youtube.com/watch?v=OZJ1IgSgP9E

Neural Networks



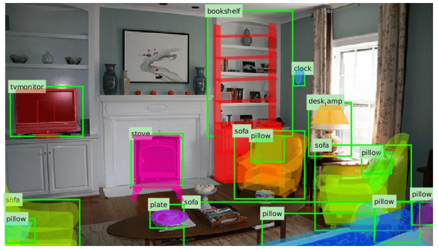
Computer Vision

Health care



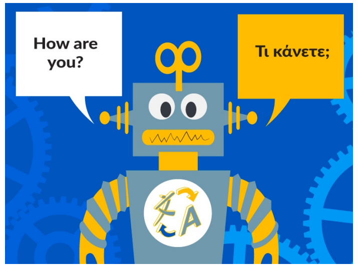
Autopilot

Object recognition



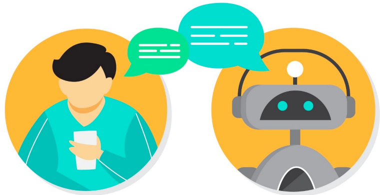
Natural Language Processing

Machine translation



Sentiment analysis

Dialog system



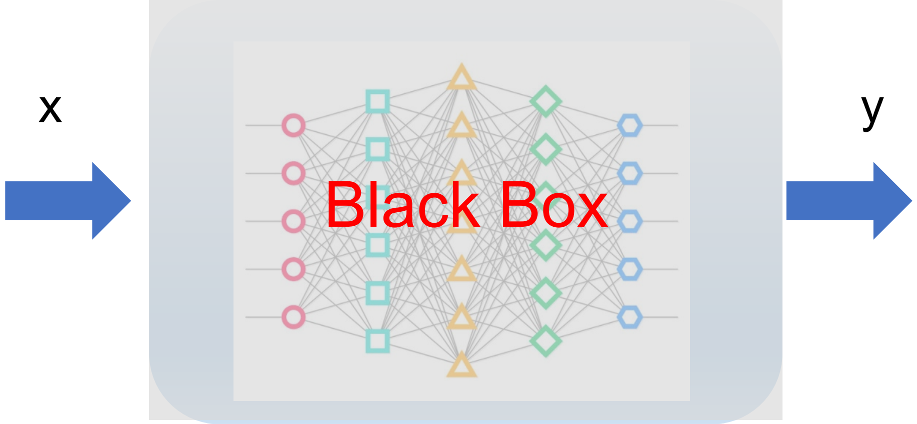
Neural Networks

What is the model learning?

How does the model make a prediction?

Can we trust the model?

How to make the model better?

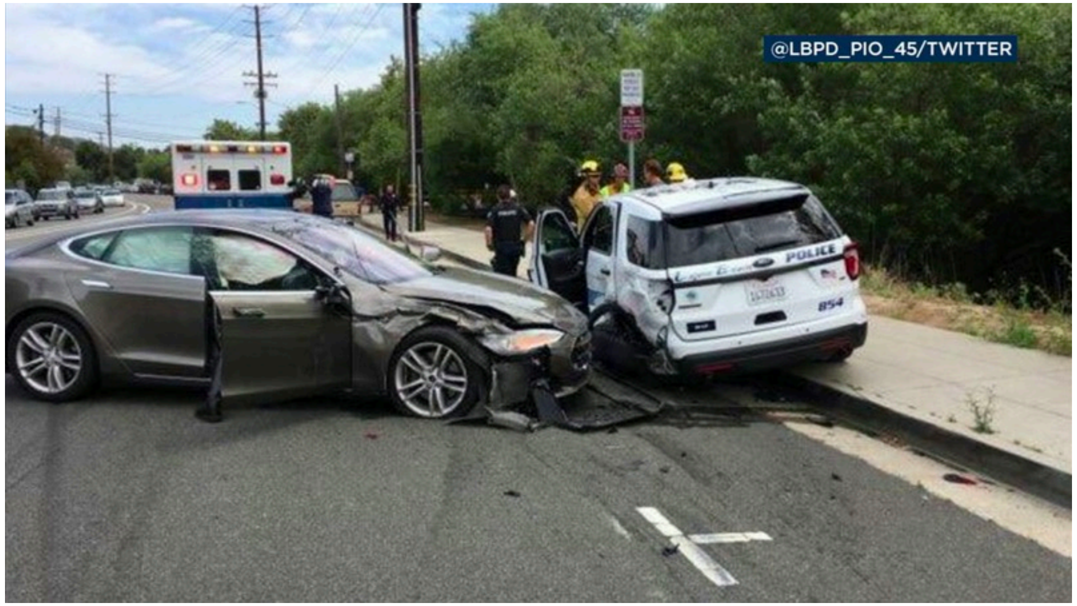


Lack of interpretability…

Black-box models are dangerous…

Unexpected Failures

Tesla hit a parked police car while using Autopilot



Risks of AI in health care

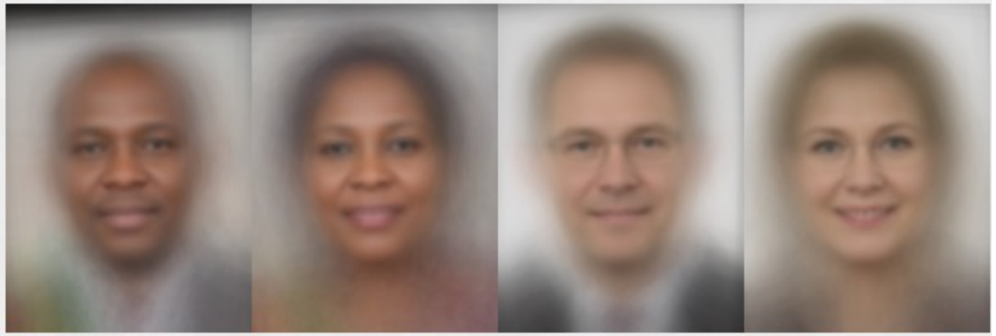
Bias and Unfairness

Machine Learning can amplify bias.

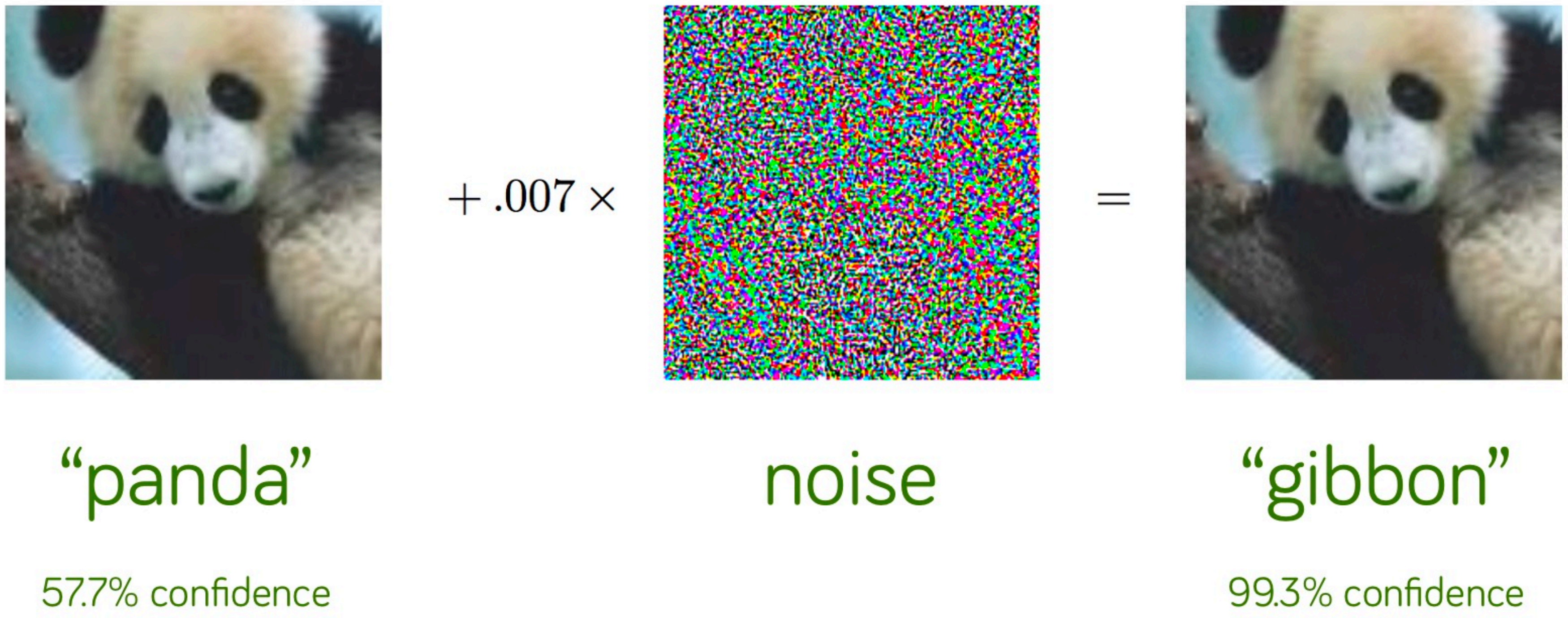
- Data set: 67% of people cooking are women
- Algorithm predicts: 84% of people cooking are women

Higher error rate on darker female

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Vulnerability to Adversarial Attacks



Interpretable Machine Learning

Interpretability → Trustworthiness

Interpretability → Fairness

Interpretability → Robustness

- Revealing model decision-making
- Understanding model prediction behavior
- Debugging and improving model

Course Information

- Website: https://uvanlp.org/iml-2022/

Course Information

- ➢ Website: https://uvanlp.org/iml-2022/

- ➢ Background

  - Machine learning models: remarkable performance, lack of interpretability
  - Interpretable machine learning: building trustworthy and reliable models

Course Information

- Website: https://uvanlp.org/iml-2022/

- Background
  - Machine learning models: remarkable performance, lack of interpretability
  - Interpretable machine learning: building trustworthy and reliable models

- Goal
  - Getting familiar with the emerging problem in machine learning
  - Learning recent advances in interpretable and explainable AI

Course Information

- Instructors

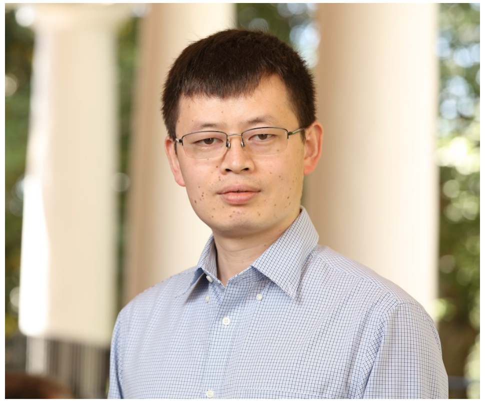
Hanjie Chen

- PhD student (4th year)
- Advisor: Prof. Yangfeng Ji
- Research: Natural Language Processing, Interpretable Machine Learning
- UVA Engineering Graduate Teaching Intern (GTI)
- Website: https://www.cs.virginia.edu/~hc9mx/
- Interests: piano, swimming, yoga, hiking…
- Fun fact: I am living with two cutest cats

Course Information

- Instructors

Yangfeng Ji



- Assistant professor
- Research: Natural Language Processing, Text Understanding and Generation
- Website: https://yangfengji.net/
- Lead the Information and Language Processing (ILP) Lab https://uvanlp.org/

Course Information

➢ TA

Wanyu Du

- PhD student (2nd year)
- Research: Natural Language Processing, Text Generation, Conversation Modeling
- Website: https://wyu-du.github.io/

Course Information

- Format
  - Hybrid: lectures will be given in person at Rice Hall 340, Zoom online (join via Collab)
  - Lectures will be recorded and uploaded to Collab

Course Information

➢ Format

- Hybrid: lectures will be given in person at Rice Hall 340, Zoom online (join via Collab)
- Lectures will be recorded and uploaded to Collab
- From Week 4 (Feb. 8, 10): one lecture + one discussion per week

Course Information

- Format
  - Hybrid: lectures will be given in person at Rice Hall 340, Zoom online (join via Collab)
  - Lectures will be recorded and uploaded to Collab
  - From Week 4 (Feb. 8, 10): one lecture + one discussion per week
  - Campusewire (online QA, connection, discussion)
  - Office hours:

Name	Time	Location
Hanjie Chen	Thursday 2:00-3:00 PM	Zoom
Yangfeng Ji	TBD	TBD
Wanyu Du	TBD	TBD

Course Information

- Prerequisites
  - Proficiency in Python
  - Basic Calculus and Linear Algebra
  - Basic Probability and Statistics
  - Foundations of Machine Learning

Course Information

- Prerequisites
  - Proficiency in Python
  - Basic Calculus and Linear Algebra
  - Basic Probability and Statistics
  - Foundations of Machine Learning

  Note: This course would not cover basic machine learning
  (please take CS 4774/6316 Machine Learning instead)

Course Information

- Assignments

Two evaluation schemes

- Application-oriented (for undergraduates)
  - 3 programming assignments (3*15%=45%)
  - 1 paper presentation (15%)
  - 10 paper summaries (10%)
  - Final project (20%)
  - In-class discussion + attendance (7% + 3%=10%)
- Research-oriented (for graduates)
  - 2 programming assignments (2*15%=30%) (choose any 2 from 3 assignments)
  - 2 paper presentations (2*15%=30%)
  - 10 paper summaries (10%)
  - Final project (20%)
  - In-class discussion + attendance (7% + 3%=10%)

Course Information

➢ Assignments

Two evaluation schemes

- Application-oriented (for undergraduates)
  - 3 programming assignments (3*15%=45%)
  - 1 paper presentation (15%)
  - 10 paper summaries (10%)
  - Final project (20%)
  - In-class discussion + attendance (7% + 3%=10%)
- Research-oriented (for graduates)
  - 2 programming assignments (2*15%=30%) (choose any 2 from 3 assignments)
  - 2 paper presentations (2*15%=30%)
  - 10 paper summaries (10%)
  - Final project (20%)
  - In-class discussion + attendance (7% + 3%=10%)

No Exam

Course Information

- Assignments

Programming assignment

Implementation of algorithms discussed in class, coding with Python

Course Information

- Assignments

Programming assignment

Implementation of algorithms discussed in class, coding with Python

Paper presentation

- Start from Week 4 (on Thursday)
- 2/3 papers per class, 35/25 mins (25/20 mins presentation + 10/5 mins QA) per paper
- 2 students per paper

Course Information

- Assignments

Programming project

Implementation of algorithms discussed in class, coding with Python

Paper presentation

- Start from Week 4 (on Thursday)
- 2/3 papers per class, 35/25 mins (25/20 mins presentation + 10/5 mins QA) per paper
- 2 students per paper
- Select one paper from the reading list (under the week you are going to present)

https://docs.google.com/spreadsheets/d/1IVlYW_4rN2sMtR4lxstmyzDtKHX9sczSoysKl2JNb88/edit#gid=1712395903

- Choose different topics for two presentations

Course Information

➢ Assignments

Programming project

Implementation of algorithms discussed in class, coding with Python

Paper presentation

- Start from Week 4 (on Thursday)
- 2/3 papers per class, 35/25 mins (25/20 mins presentation + 10/5 mins QA) per paper
- 2 students per paper
- Select one paper from the reading list (under the week you are going to present)

https://docs.google.com/spreadsheets/d/1IVlYW_4rN2sMtR4lxstmyzDtKHX9sczSoysKl2JNb88/edit#gid=1712395903

- Choose different topics for two presentations
- Sign up before Feb. 3rd

https://docs.google.com/spreadsheets/d/1IVlYW_4rN2sMtR4lxstmyzDtKHX9sczSoysKl2JNb88/edit#gid=0

Course Information

- Assignments

Paper presentation (Rubric)

- Introduction/Background (3’)
- Research problem/Motivation(3’)
- Methodology (3’)
- Experimental results (3’)
- Conclusion/Takeaway (3’)

Course Information

- Assignments

Paper summary

- Start from Week 4 (due on Tuesday)
- Submit one summary at most per week, 10 paper summaries in total

Course Information

- Assignments

Paper summary

- Start from Week 4 (due on Tuesday)
- Submit one summary at most per week, 10 paper summaries in total
- Select one paper from the reading list (under the week you are going to submit)

https://docs.google.com/spreadsheets/d/1IVlYW_4rN2sMtR4lxstmyzDtKHX9sczSoysKl2JNb88/edit#gid=1712395903

Course Information

➢ Assignments

Paper summary

- Start from Week 4 (due on Tuesday)
- Submit one summary at most per week, 10 paper summaries in total
- Select one paper from the reading list (under the week you are going to submit)

https://docs.google.com/spreadsheets/d/1IVlYW_4rN2sMtR4lxstmyzDtKHX9sczSoysKl2JNb88/edit#gid=1712395903

- Use the template to write a short summary (0.5-1 page)

Questions

1. Paper title
2. What is the research problem addressed in this paper?
3. What is the proposed method? How it can address the problem?
4. What are the main observations/conclusions from the experiments?

Course Information

- Assignments

Final project

- Proposal (6%, due on Mar. 24)
- Final presentation (7%, due on May. 3)
- Final project report (7%, due on May. 6)
- 2 – 3 students per group
- Sign up before Mar. 24

https://docs.google.com/spreadsheets/d/1IVlYW_4rN2sMtR4lxstmyzDtKHX9sczSoysKl2JNb88/edit#gid=410460640

Course Information

- ➢ Assignments

Final project

- Related to model interpretation/interpretability
- Implement interpretation methods to solve a real-world problem
- Explore the interpretability of a specific machine learning model
- Reproduce the results in a paper regarding interpretable ML published on top-tier AI conferences (AAAI, NeurIPS, ICLR, ICML, ACL, EMNLP, CVPR, ICCV…)
- …

Course Information

- Assignments

Final project

- Related to model interpretation/interpretability
- Implement interpretation methods to solve a real-world problem
- Explore the interpretability of a specific machine learning model
- Reproduce the results in a paper regarding interpretable ML published on top-tier AI conferences (AAAI, NeurIPS, ICLR, ICML, ACL, EMNLP, CVPR, ICCV…)
- …

Reproducing results in a paper is not easy. There are many factors that may be out of your control (e.g., hyperparameters, environment…)

Course Information

- Assignments

Final project (Rubric-proposal)

- Introduction (2’): background/motivation, research problem
- Models and datasets (1’)
- Proposed method (1’)
- Experiments (2’): plan, evaluation criteria

Course Information

- Assignments

Final project (Rubric-presentation/report)

- Introduction (2’): background/motivation, research problem
- Models and datasets (1’)
- Proposed method (2’): a description of the proposed method, a justification about why you think the proposed method could work
- Experimental results (2’): observations, conclusions

Course Information

- Assignments

In-class discussion

- Ask questions in QA session, leave questions in Zoom channel, post comments on Campuswire

Course Information

- Assignments

In-class discussion

- Ask questions in QA session, leave questions in Zoom channel, post comments on Campuswire
- Remember to move all your questions/comments to Campuswire forum within 30 mins after the class

Course Information

- Assignments

In-class discussion

- Ask questions in QA session, leave questions in Zoom channel, post comments on Campuswire
- Remember to move all your questions/comments to Campuswire forum within 30 mins after the class
- Commenting on one paper (no matter how many comments) would be counted once

Course Information

- Assignments

In-class discussion

Number	Points
≥ 13	7
[11, 13)	6
[9, 11)	5
[7, 9)	4
[5, 7)	3
[3, 5)	2
[1, 3)	1
0	0

Course Information

➢ Assignments

Attendance

- If you attend the class in person, please sign the table after class
- If you join in Zoom, we will count attendance at a random time during the class

Missing classes	Points
≤ 3	3
4	2
5	1
> 5	0

Policy

- Late penalty

  Homework submission will be accepted up to 48 hours late, with 20% deduction per 24 hours on the points as a penalty

Late time (hours)	Penalty
(0, 24]	20%
(24, 48]	40%

Policy

➢ Late penalty

Homework submission will be accepted up to 48 hours late, with 20% deduction per 24 hours on the points as a penalty

Late time (hours)	Penalty
(0, 24]	20%
(24, 48]	40%

For example:

- Deadline: Feb. 8th, 11:59 PM
- Submission timestamp: Feb. 10th, 9:00 AM ($\leq$ 48 hours)
- Original points of a homework: 10
- Actual points: $10 \times (1 - 40\%) = 6$

Policy

➢ Late penalty

Note:

- It is usually better if students just turn in what they have in time
- It’s the students’ responsibility to double check their submission (We DO NOT accept any replacement if the deadline has passed over 48 hours, or we would treat it as a late submission if it is still acceptable)
- If a student submits one homework via multiple files/times, we will use the latest timestamp for deciding and calculating the late penalty

Policy

- Collaboration
  - Students should work on programming projects and paper summaries independently
  - Discussions are encouraged, but copying or plagiarizing homework is NOT allowed
  - In your submission, please list the names of your classmates who have discussions with you on that assignment
  - Students are encouraged to work as a team on paper presentations and final projects
  - Each team only needs to submit one report/presentation
  - All team members will have the same points for each submission

Policy

- Note
  - All assignments will be submitted at Collab
  - Campuswire: in-class discussion, forming teammates, course announcements, online QA, group discussion (with a chatroom)

Policy

- Grades

Point range	Letter grade
[98, 100]	A+
[94, 98)	A
[90, 94)	A-
[88, 90)	B+
[83, 88)	B
[80, 83)	B-
[74, 80)	C+
[67, 74)	C
[60, 67)	C-
[0, 60)	F

Course Schedule

Date		Topic	Assignments/Deadlines
Week 1: Jan. 20		Course overview	-
Week 2	Jan. 25	Introduction to interpretability	-
	Jan. 27	Interpretable generalized additive models (GAMs)	-
Week 3	Feb. 1	Introduction to neural networks	-
	Feb. 3	Introduction to neural networks	Sign up presentation form
Week 4	Feb. 8	Post-hoc explanations for black-box models: perturbation-based methods	Paper summary
	Feb. 10	Paper presentation	Programming project 1 out

Course Schedule

Date		Topic	Assignments/Deadlines
Week 5	Feb. 15	Post-hoc explanations for black-box models: gradient/attention-based methods	Paper summary
	Feb. 17	Paper presentation	-
Week 6	Feb. 22	Post-hoc explanations for black-box models: beyond feature-level	Paper summary
	Feb. 24	Paper presentation	Programming project 1 due Programming project 2 out
Week 7	Mar. 1	Improving neural network intrinsic interpretability	Paper summary
	Mar. 3	Paper presentation	-
Week 8		**Spring Recess**	

Course Schedule

Date		Topic	Assignments/Deadlines
Week 9	Mar. 15	Building interpretable neural network models	Paper summary
	Mar. 17	Paper presentation	Programming project 2 due Programming project 3 out
Week 10	Mar. 22	Rationalized neural networks	Paper summary
	Mar. 24	Paper presentation	Final project proposal, sign up the final project form
Week 11	Mar. 29	Interpretation and human understanding	Paper summary
	Mar. 31	Paper presentation	Programming project 3 due
Week 12	Apr. 5	Robust interpretations	Paper summary
	Apr. 7	Paper presentation	-

Course Schedule

Date		Topic	Assignments/Deadlines
Week 13	Apr. 12	Connections with model performance, robustness, fairness	Paper summary
	Apr. 14	Paper presentation	-
Week 14	Apr. 19	Paper presentation	Paper summary
	Apr. 21	Paper presentation	Paper summary
Week 15	Apr. 26	Paper presentation	Paper summary
	Apr. 28	Paper presentation	Paper summary
Week 16: May. 3		Final presentation	-

Question?