**COS 598: Interpretability and Explainability in Machine Learning**

**Spring 2021**

Dr. Chaofan Chen, Assistant Professor

Email: [chaofan.chen@maine.edu](mailto:chaofan.chen@maine.edu)

Office Hours:

Lectures: Tuesdays and Thursdays, 12:30 pm - 01:45 pm

Location:

*Course Description*

As machine learning models are playing increasingly important roles in many real-life scenarios, interpretability has become a key issue for whether we can trust the predictions made by these models, especially when we are making some high-stakes decisions. Lack of transparency has long been a concern for predictive models in criminal justice and in healthcare. There have been growing calls for building interpretable, human understandable machine learning models, and “opening the black box” has become a debated issue in the media.

This graduate-level course aims to familiarize students with recent advances in interpretable and explainable machine learning. In this course, we will discuss seminal papers of the field, the notion of interpretability and explainability, traditional interpretable models, posthoc model interpretability analysis, and interpretability in deep learning. This course will also introduce various applications that will benefit from model interpretability and explainability, such as healthcare and finance.

*Learning Objectives and Outcomes*

By the end of this course, you will:

* have a broad knowledge of the field of explainable and interpretable machine learning;
* be familiar with the most recent advances in explainable and interpretable machine learning;
* develop or apply explainable and/or interpretable machine learning techniques for a problem of your interest.

*Prerequisites*

COS 598 Machine Learning, or ECE 498/598 Deep Learning, or concurrent enrollment in either of the above courses, or instructor consent.

*Grading*

Grades are based on:

* Homework (15%)
* Course project (75%)
* Participation (10%)

The ranges for the grades are:

* A: ≥ 90, A-: [85, 90)
* B+: [80, 85), B: [75, 80), B-: [70, 75)
* C+: [65, 70), C: [60, 65), C-: [55, 60)
* D: [50, 55)
* F: < 50

*Homework*

There are 3 homework assignments for this course. Homework assignments involve written questions and coding exercises. Late homework assignments will not be accepted.

The homework assignments account for 15% of your final grade.

*Course Project*

An important part of this course is to develop or apply explainable and/or interpretable machine learning techniques for a problem of your interest. To that end, you will be required to complete a semester-long course project with me for the duration of this course. There will be bi-weekly (or as frequent as necessary) project meetings to ensure that you are making steady progress in your course project. You must discuss with me, and decide on your course project topic within the first five weeks of the course. Once you decide on your project topic, you will submit a *project proposal*, detailing your plan for completing the course project. You will also submit a *midterm report* before Week 11, detailing the progress you will have made, and any additional tasks remaining that need to be completed. You will give a *presentation* of your course project during the last lecture of this course, and you will submit a *final report*, which is due at the time when the scheduled final exam for the course ends. There will be no in-class final exam for this course.

The course project accounts for 75% of your final grade.

*Participation*

Since this is a research seminar, and we will be discussing papers, your *active* participation is important for this course. You are expected to complete the readings assigned for each lecture *prior to* the lecture, and you are expected to come with your questions and/or comments about the assigned readings.

Participation accounts for 10% of your final grade. Note that unexcused absences will lower your participation score, which may in turn negatively impact your final grade.

*Course Schedule*

|  |  |  |
| --- | --- | --- |
| Week | Lecture: Date | Topics |
| 1 | Lecture 1: January 26 | Course overview  A review of mathematical and statistical foundations of machine learning  Readings:  [Linear algebra](https://www.deeplearningbook.org/contents/linear_algebra.html)  [Probability and information theory](https://www.deeplearningbook.org/contents/prob.html)  [Numerical computation](https://www.deeplearningbook.org/contents/numerical.html) |
| Lecture 2: January 28 | Understanding machine learning  Readings:  [Machine learning basics](https://www.deeplearningbook.org/contents/ml.html)  [Bayesian learning](https://ermongroup.github.io/cs228-notes/learning/bayesian/) |
| 2 | Lecture 3: February 2 | Deep learning  Readings:  [Deep feedforward networks](https://www.deeplearningbook.org/contents/mlp.html)  [Regularization for deep learning](https://www.deeplearningbook.org/contents/regularization.html)  [Optimization for training deep models](https://www.deeplearningbook.org/contents/optimization.html) |
| Lecture 4: February 4 | Explaining black-box models  Readings:  [Ribeiro et. al., 2016](https://arxiv.org/pdf/1602.04938.pdf) |
| 3 | Lecture 5: February 9 | Explaining black-box models  Readings:  [Rudin and Shaposhnik, 2019](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3395422) |
| Lecture 6: February 11 | Convolutional neural networks  Readings:  [Convolutional networks](https://www.deeplearningbook.org/contents/convnets.html)  [Zeiler et al., 2010 (Deconvolutional networks)](https://www.matthewzeiler.com/mattzeiler/deconvolutionalnetworks.pdf)  [Dumoulin and Visin, 2016](https://arxiv.org/pdf/1603.07285.pdf) |
| 4 | Lecture 7: February 16 | Explaining convolutional neural networks: activation maximization  Readings:  [Erhan et al., 2009](https://www.researchgate.net/profile/Aaron_Courville/publication/265022827_Visualizing_Higher-Layer_Features_of_a_Deep_Network/links/53ff82b00cf24c81027da530.pdf)  [Nguyen et al., 2016](https://arxiv.org/pdf/1605.09304.pdf) |
| Lecture 8: February 18 | Explaining convolutional neural networks: deconvolution, occlusion  Readings:  [Zeiler and Fergus, 2014](https://arxiv.org/pdf/1311.2901.pdf)  [Springenberg et al., 2015](https://arxiv.org/pdf/1412.6806.pdf)  [Yosinsk et al., 2015](https://arxiv.org/pdf/1506.06579.pdf)  [Zhou et al., 2015](https://arxiv.org/pdf/1412.6856.pdf) |
| 5 | Lecture 9: February 23 | Explaining convolutional neural networks: gradient-based saliency  Readings:  [Simonyan et al., 2014](https://arxiv.org/pdf/1312.6034.pdf)  [Smilkov et al., 2017](https://arxiv.org/pdf/1706.03825.pdf)  [Sundararajan et al., 2017](https://arxiv.org/pdf/1703.01365.pdf) |
| Lecture 10: February 25 | Explaining convolutional neural networks: layer-relevance propagation  Readings:  [Bach et al., 2015](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140)  [Montavon et al., 2015](https://arxiv.org/pdf/1512.02479.pdf)  [Shrikumar et al., 2019](https://arxiv.org/pdf/1704.02685.pdf) |
| 6 | Lecture 11: March 2 | Recurrent neural networks  Explaining recurrent neural networks  Readings:  [Sequence modeling: recurrent and recursive nets](https://www.deeplearningbook.org/contents/rnn.html)  [Karpathy et al., 2016](https://arxiv.org/pdf/1506.02078.pdf) |
| Lecture 12: March 4 | Explainability versus interpretability: critiques of explainable machine learning  Readings:  [Ghorbani et. al., 2019](https://arxiv.org/pdf/1710.10547.pdf)  [Adebayo et. al., 2018](https://papers.nips.cc/paper/8160-sanity-checks-for-saliency-maps.pdf)  [Rudin, 2019](https://arxiv.org/pdf/1811.10154.pdf) |
| 7 | Lecture 13: March 9 | Understanding interpretability  Readings:  [Doshi-Velez and Kim, 2017](https://arxiv.org/pdf/1702.08608.pdf)  [Lipton, 2017](https://arxiv.org/pdf/1606.03490.pdf)  [Weller, 2019](https://arxiv.org/pdf/1708.01870.pdf) |
| Lecture 14: March 11 | Evaluating interpretability  Readings:  [Lage et. al., 2019](https://ojs.aaai.org/index.php/HCOMP/article/view/5280/5132)  [Poursabzi-Sangdeh, 2018](https://arxiv.org/pdf/1802.07810.pdf)  [Bau et al., 2017](https://openaccess.thecvf.com/content_cvpr_2017/papers/Bau_Network_Dissection_Quantifying_CVPR_2017_paper.pdf) |
| 8 | Lecture 15: March 16 | Attention mechanism in deep learning: activation-based (soft) attention  Readings:  [Zhou et al., 2016](https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Zhou_Learning_Deep_Features_CVPR_2016_paper.pdf)  [Zheng et al., 2017](https://openaccess.thecvf.com/content_ICCV_2017/papers/Zheng_Learning_Multi-Attention_Convolutional_ICCV_2017_paper.pdf) |
| Lecture 16: March 18 | Attention mechanism in deep learning: hard attention  Readings:  [Fu et al., 2017](https://openaccess.thecvf.com/content_cvpr_2017/papers/Fu_Look_Closer_to_CVPR_2017_paper.pdf)  [Jaderberg et al., 2015](https://papers.nips.cc/paper/2015/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf) |
| 9 | Lecture 17: March 23 | Attention mechanism in deep learning: hard attention, natural language processing  Readings:  [Lei et al., 2016](https://people.csail.mit.edu/taolei/papers/emnlp16_rationale.pdf) |
| Lecture 18: March 25 | Case-based reasoning and prototype-based models  Readings:  [Kim et. al., 2014](https://beenkim.github.io/papers/KimRudinShahNIPS2014.pdf) |
| 10 | Lecture 19: March 30 | Case-based reasoning in deep learning  Readings:  [Li et al., 2018](https://arxiv.org/pdf/1710.04806.pdf)  [Chen et al., 2019](https://proceedings.neurips.cc/paper/2019/file/adf7ee2dcf142b0e11888e72b43fcb75-Paper.pdf) |
| Lecture 20: April 1 | Case-based reasoning in deep learning  Readings:  [Nauta et al., 2020](https://arxiv.org/pdf/2011.02863.pdf)  [Ming et al., 2019](https://arxiv.org/pdf/1907.09728.pdf) |
| 11 | Lecture 21: April 6 | Rule-based interpretable models: Bayesian approach  Readings:  [Metropolis Hastings Algorithm](https://ermongroup.github.io/cs323-notes/probabilistic/mh/)  [Letham and Rudin, 2015](https://arxiv.org/pdf/1511.01644.pdf) |
| Lecture 22: April 8 | Rule-based interpretable models: Bayesian approach  Readings:  [Gibbs Sampling](https://ermongroup.github.io/cs323-notes/probabilistic/gibbs/)  [Wang and Rudin, 2015](https://arxiv.org/pdf/1411.5899.pdf) |
| 12 | Lecture 23: April 13 | Rule-based interpretable models: optimization approach  Readings:  A\* search algorithm, branch-and-bound  [Angelino et al., 2018](https://www.jmlr.org/papers/volume18/17-716/17-716.pdf) |
| Lecture 24: April 15 | Rule-based interpretable models: optimization approach  Readings:  [Chen and Rudin, 2018](http://proceedings.mlr.press/v84/chen18a/chen18a.pdf) |
| 13 | Lecture 25: April 20 | Rule-based interpretable models: submodular optimization  Readings:  [Submodular Optimization](https://las.inf.ethz.ch/files/krause12survey.pdf)  [Lakkaraju et. al., 2016](https://www-cs-faculty.stanford.edu/people/jure/pubs/interpretable-kdd16.pdf) |
| Lecture 26: April 22 | Risk scores and additive models  Readings:  [Cutting Plane Methods](https://www.cse.iitb.ac.in/~cs709/2015a/notes/readingAssignment/CuttingPlaneMethod.pdf) [Section 7.1]  [Ustun and Rudin, 2017](https://canvas.harvard.edu/courses/68154/files/?preview=8630586) |
| 14 | Lecture 27: April 27 | Risk scores and additive models  Readings:  [Generalized Additive Models](https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch13.pdf)  [Caruana et. al., 2015](http://people.dbmi.columbia.edu/noemie/papers/15kdd.pdf)  [Chen et al., 2018](https://arxiv.org/pdf/1811.12615.pdf) |
| Lecture 28: April 29 | Project presentations  Discussion: future of interpretable machine learning |

*Reference Materials*

There is no required textbook for this course, since we will not be following a single book or source for the course material, and we will be reading papers most of the time. If you need to learn some background materials, the *Deep Learning* book by Goodfellow et al. is a great resource, and is accessible at:

<https://www.deeplearningbook.org/>

*Attendance*

Attendance is required for this course. Unexcused absences will lower your participation score, which may in turn negatively impact your final grade.

*Special COVID-19* *Requirements and* *Accommodations*

Please show respect to your fellow classmates and instructor, by wearing a face covering and maintaining social distance during in-person lectures. *Violations of the mask wearing policy during in-person lectures will be handled by the Office of Student Life and will be subject to student conduct procedures. In the most serious cases, students can be sent home for such violations.*

Eating or drinking during in-person lectures is considered a violation of the mask wearing policy, and is therefore *strictly prohibited*.

If you are feeling unwell, please do not come to in-person lectures. There is a concurrent remote section of the course, and you can always attend the lectures remotely on Zoom.

If you absolutely cannot complete the homework assignments and/or the course project on time due to serious health conditions, please contact me through email for possible extensions or “Incomplete.”

*Academic Honesty*

Academic honesty is very important. It is dishonest to plagiarize works written by another person and pass those works off as your own without proper citations. Students committing academic dishonesty may be given failing grades for an assignment or for an entire course, at the discretion of the instructor. In addition to any academic action taken by an instructor, these violations are also subject to disciplinary actions under the University of Maine System Student Code of Conduct. The maximum possible penalty under the Student Code of Conduct is dismissal from the University.

*Civil Discourse*

Interaction with the instructor is expected be carried out in a civil and respectful manner. This does not mean that you need to use titles or be overwhelmingly polite to the instructor, but it does mean that the instructor reserves the right to pass summary judgment on issues if a student acts profoundly disrespectful or uncivil.

*Disabilities*

Students with disabilities who believe that they may need accommodations in this course are encouraged to contact the Student Accessibility Services at (207) 581-2319 as soon as possible, to ensure that such accommodations can be made.

*Sexual Discrimination/Violence Reporting*

The University of Maine is committed to making the campus a safe place for students. This commitment means that, if you tell any of your instructors about sexual discrimination/violence involving members of the campus, your instructor is required to report this information to the Office of Equal Opportunity or the Office of Sexual Assault and Violence Prevention.