****

**STAT 992: Interpretable and Explainable Machine Learning**

**1 Credit**

**Course Designations and Attributes**

Natural Science

LAS Credit

Intermediate

**Course Description**

Artificial intelligence and machine learning are transforming science and society. In this context, the need to provide faithful, context-specific explanations of model predictions and properties has become essential, leading to the development of a range of techniques for building intrinsically interpretable models and explaining black box models. This course will survey this literature, examining the computational and mathematical foundations of cutting-edge techniques, and providing hands-on experience in realistic modeling settings. Techniques covered include Shapley values, saliency maps, concept activation vectors, distillation, Rashomon ensembles, and dictionary learning. We will also review strategies for evaluating the effectiveness of explainability techniques. Techniques will be illustrated through applications to scientific discovery, algorithmic fairness, model control, and methods development.

**Requisites**

Graduate/professional standing, member of Statistics Visiting International Scholars program

**Meeting Time and Location**

Thursdays 2:30 – 3:45pm, B215 Van Vleck Hall.

**Instructional Modality**

In-person

**How Credit Hours are Met by the Course**

The credit standard for this course is met by an expectation of a total of 45 hours of student engagement

with the courses learning activities, which include regularly scheduled: readings, in-class discussions, and homework assignments described in this syllabus.

**Regular and Substantive Student-Instructor Interaction**

Participation in regularly scheduled lectures each week will include the opportunity for direct interaction between students and the instructor. The instructor will also frequently interact and post announcements in Canvas and email students about academic aspects of the class.

Instructors & Teaching Assistants

**Instructor**

Kris Sankaran ([ksankaran@wisc.edu](mailto:ksankaran@wisc.edu)). Office hours can be arranged by appointment. They can be held either at Medical Sciences Center 7225C or at this [zoom link](https://uwmadison.zoom.us/j/8622164885).

Course Learning Outcomes

By the end of this course, you will be able to:

1. **Explain predictions**: Analyze individual predictions from machine learning and AI models using algorithms and software from the interpretability literature.
2. **Explain models**: Investigate components of trained machine learning and AI models using algorithms and software from the interpretability literature.
3. **Explain learning**: Investigate the learning dynamics of machine learning and AI models using algorithms and software from the interpretability literature.
4. **Modify models**: Apply methods from the interpretability literature to enhance the interpretability of an initial machine learning or AI model.
5. **Evaluate methods**: Apply mathematical and computational analysis to critique applications of interpretability and explainability techniques.

**Grading**

Assignments: 75%

Participation: 25%

Grades will be assigned according to the percentage scale, A = 92-100, AB = 88-91.9, B = 82-87.9, BC = 78-81.9, C = 70-77.9, D = 60-69.9, F = 0-59.9 (92% of points => A); and according to the percentile scale, A = 75, AB = 65, B = 45, BC = 30, C = 10, D = 5, F = 0 (performing better than 75% of the class => A). Your grade will be the higher of these two grades.

Required Textbook, Software & Other Course Materials

* There are no required textbooks. All readings are provided in the table below.

Campus provides students with [technology guidelines and recommendations](https://it.wisc.edu/learn/guides/learning-online-technology-tips-tools/) for instruction. Students should consult these resources prior to the start of the semester.

Homework & Other Assignments

* Readings and Discussions
  + We will discuss the papers included in the reading list below. You are expected to have completed the readings before each course session.
  + Each class will include a discussion component, and your attendance is important. You will be excused from two absences. Any absences beyond this will only be excused for documented medical or family emergencies.
* Assignments
  + This course includes three short reports. The reports give you an opportunity to explore the application of interpretability and explainability techniques within your areas of interest/expertise. Deadlines are indicated on Canvas.
  + It is possible to work together with another student on a series of related reports – e.g., if you think your joint analyses could be submitted to a workshop in the future. Please contact the instructor if you would like to pursue this option.
* All assignments must be submitted on Canvas.
  + For every 24 hours late that a submission is made, it will be penalized 5%, for up to 4 days, after which no submissions will be accepted. The only exception for late acceptance will be in documented medical or family emergencies.

Exams, Quizzes, Papers & Other Major Graded Work

* There are no exams.

Teaching & Learning Data Transparency Statement

*The privacy and security of faculty, staff and students’ personal information is a top priority for UW-Madison. The university carefully evaluates and vets all campus-supported digital tools used to support teaching and learning, to help support success through*[learning analytics](https://teachlearn.provost.wisc.edu/learning-analytics/)*, and to enable proctoring capabilities. View the university’s full*[teaching and learning data transparency statement](https://teachlearn.provost.wisc.edu/teaching-and-learning-data-transparency-statement/)*.*

Privacy of Student Records & the Use of Audio Recorded Lectures Statement

*View* [more information about *FERPA*](https://registrar.wisc.edu/ferpa-facstaff/)*.*

Lecture materials and recordings for this course are protected intellectual property at UW-Madison. Students in this course may use the materials and recordings for their personal use related to participation in this class. Students may also take notes solely for their personal use. If a lecture is not already recorded, you are not authorized to record my lectures without my permission unless you are considered by the university to be a qualified student with a disability requiring accommodation. [Regent Policy Document 4-1] Students may not copy or have lecture materials and recordings outside of class, including posting on internet sites or selling to commercial entities. Students are also prohibited from providing or selling their personal notes to anyone else or being paid for taking notes by any person or commercial firm without the instructor’s express written permission. Unauthorized use of these copyrighted lecture materials and recordings constitutes copyright infringement and may be addressed under the university’s policies, UWS Chapters 14 and 17, governing student academic and non-academic misconduct.

Course Evaluations

Students will be provided with an opportunity to evaluate this course and your learning experience. Student participation is an integral component of this course, and your confidential feedback is important to me. I strongly encourage you to participate in the course evaluation.

UW-Madison uses a digital course evaluation survey tool called [AEFIS](https://kb.wisc.edu/luwmad/page.php?id=81069). For this course, you will receive an official email two weeks prior to the end of the semester, notifying you that your course evaluation is available. In the email you will receive a link to log into the course evaluation with your NetID. Evaluations are anonymous. Your participation is an integral component of this course, and your feedback is important to me. I strongly encourage you to participate in the course evaluation.

Students Rules, [Rights & Responsibilities](https://guide.wisc.edu/undergraduate/#rulesrightsandresponsibilitiestext)

Diversity & Inclusion Statement

[Diversity](https://diversity.wisc.edu/) is a source of strength, creativity, and innovation for UW-Madison. We value the contributions of each person and respect the profound ways their identity, culture, background, experience, status, abilities, and opinion enrich the university community. We commit ourselves to the pursuit of excellence in teaching, research, outreach, and diversity as inextricably linked goals.The University of Wisconsin-Madison fulfills its public mission by creating a welcoming and inclusive community for people from every background – people who as students, faculty, and staff serve Wisconsin and the world.

Academic Integrity Statement

By virtue of enrollment, each student agrees to uphold the high academic standards of the University of Wisconsin-Madison; academic misconduct is behavior that negatively impacts the integrity of the institution. Cheating, fabrication, plagiarism, unauthorized collaboration, and helping others commit these previously listed acts are examples of misconduct which may result in disciplinary action. Examples of disciplinary action include, but is not limited to, failure on the assignment/course, written reprimand, disciplinary probation, suspension, or expulsion.

Accommodations for Students with Disabilities Statement

The University of Wisconsin-Madison supports the right of all enrolled students to a full and equal educational opportunity. The Americans with Disabilities Act (ADA), Wisconsin State Statute (36.12), and UW-Madison policy (Faculty Document 1071) require that students with disabilities be reasonably accommodated in instruction and campus life. Reasonable accommodations for students with disabilities is a shared faculty and student responsibility. Students are expected to inform faculty [me] of their need for instructional accommodations by the end of the third week of the semester, or as soon as possible after a disability has been incurred or recognized. Faculty [I], will work either directly with the student [you] or in coordination with the McBurney Center to identify and provide reasonable instructional accommodations. Disability information, including instructional accommodations as part of a student's educational record, is confidential and protected under FERPA. (See: [McBurney Disability Resource Center](https://mcburney.wisc.edu/))

[Academic Calendar & Religious Observances](https://secfac.wisc.edu/academic-calendar/)

*You can use the link above to provide your students with information about the current and future academic calendars, along with the university’s religious observance policy.*

**SCHEDULE**

|  |  |  |  |
| --- | --- | --- | --- |
| **Module** | **Week** | **Topic** | **Key Readings** |
| ***Module 1: Explaining Predictions*** | Week 1 | Context and Motivation | - [Towards A Rigorous Science of Interpretable Machine Learning](https://arxiv.org/abs/1702.08608)  - [Definitions, methods, and applications in interpretable machine learning](https://doi.org/10.1145/3236386.3241340)  - [The mythos of model interpretability](https://dl.acm.org/doi/10.1145/3236386.3241340) |
|  | Week 2 | Saliency Maps | - [Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps](https://arxiv.org/abs/1312.6034)  - [Interpretable Explanations of Black Boxes by Meaningful Perturbation](https://arxiv.org/abs/1704.03296)  - [Axiomatic Attribution for Deep Networks](https://arxiv.org/abs/1703.01365)  - [Sanity Checks for Saliency Maps](https://arxiv.org/abs/1810.03292) |
|  | Week 3 | SHAP | - ["Why Should I Trust You?": Explaining the Predictions of Any Classifier](https://dl.acm.org/doi/10.1145/2939672.2939778)  - [Algorithms to estimate Shapley value feature attributions](https://www.nature.com/articles/s42256-023-00657-x)  - [An unexpected unity among methods for interpreting model predictions](https://arxiv.org/abs/1611.07478)  - [L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data](https://arxiv.org/abs/1808.02610) |
|  | Week 4 | Example-based | - [This looks like that: deep learning for interpretable image recognition](https://arxiv.org/abs/1806.10574)  - [Counterfactual Explanations Without Opening the Black Box](https://arxiv.org/abs/1711.00399)  - [Generating Visual Explanations](https://arxiv.org/abs/1603.08507)  - Examples are not enough, learn to criticize! criticism for interpretability |
| ***Module 2: Explaining Models & Learning - For Users*** | Week 5 | ALE/PDP | - [Visualizing the effects of predictor variables in black box supervised learning models](https://doi.org/10.1111/rssb.12377)  - [Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation](https://www.tandfonline.com/doi/full/10.1080/10618600.2014.907095)  - [Visualizing Fit and Lack of Fit in Complex Regression Models](https://www.jstatsoft.org/article/view/v087i09) |
|  | Week 6 | Variable Importance | - [MDI+: A Flexible Random Forest-Based Feature Importance Framework](https://arxiv.org/abs/1908.10610)  - [Conditional variable importance for random forests](https://doi.org/10.1186/1471-2105-9-307)  - [Efficient nonparametric statistical inference on population feature importance](https://arxiv.org/abs/1908.10832)  - [Lazy Estimation of Variable Importance for Large Neural Networks](https://arxiv.org/abs/2007.12213) |
|  | Week 7 | Probes/Concepts | - [Network Dissection: Quantifying Interpretability of Deep Visual Representations](https://arxiv.org/abs/1704.05796)  - [Human-centered concept explanations for neural networks](https://ebookcentral.proquest.com/lib/wisc/reader.action?docID=29070505&ppg=351)  - [Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)](https://arxiv.org/abs/1711.11279)  - [Understanding intermediate layers using linear classifier probes](https://arxiv.org/abs/1610.01644) |
|  | Week 8 | Benchmarking and Evaluation | - [A Benchmark for Interpretability Methods in Deep Neural Networks](https://arxiv.org/abs/1806.10758)  - [Interpreting interpretability: understanding data scientists' use of interpretability tools](https://arxiv.org/abs/2102.12452)  - [Does the whole exceed its parts? the effect of ai explanations on complementary team performance](https://arxiv.org/abs/2006.14779) |
| ***Module 3: Modifying Models*** | Week 9 | Distillation | - [Distill-and-compare: Auditing black-box models using transparent model distillation](https://arxiv.org/abs/1710.06169)  - [**InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders**](https://www.biorxiv.org/content/10.1101/2024.11.14.623630v2)  - [Adaptive wavelet distillation from neural networks through interpretations](https://arxiv.org/abs/2107.10046) |
|  | Week 10 | Regularization | - [Neural Interaction Transparency (NIT): Disentangling Learned Interactions](https://arxiv.org/abs/1812.09837)  - [oi-VAE: Output Interpretable VAEs for Nonlinear Group Factor Analysis](https://arxiv.org/abs/1902.02734)  - [Concept Bottleneck Models](https://arxiv.org/abs/2007.04612)  - [Post-hoc Concept Bottleneck Models](https://arxiv.org/abs/2205.15480) |
|  | Week 11 | Renegades | - [A generalizable and accessible approach to machine learning with global satellite imagery](https://www.nature.com/articles/s41467-021-24638-z)  - [Position: Amazing Things Come From Having Many Good Models](https://arxiv.org/abs/2012.05708) |
| ***Module 4: Explaining Models & Learning - For Developers*** | Week 12 | Representational Analysis I (Vision) | - [How transferable are features in deep neural networks?](https://arxiv.org/abs/1411.1792)  - [Do Vision Transformers See Like Convolutional Neural Networks?](https://arxiv.org/abs/2108.08810)  - [Using latent space regression to analyze and leverage compositionality in GANs](https://arxiv.org/abs/2103.10426) |
|  | Week 13 | Representation Analysis II (Language) | - [Visualizing and Understanding Recurrent Networks](https://arxiv.org/abs/1506.02078)  - [Visualizing and measuring the geometry of BERT](https://arxiv.org/abs/1906.02715)  - [Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet](https://arxiv.org/abs/2402.03616) |
|  | Week 14 | Representational Analysis III (Mechanisms) | - [Discovering Latent Knowledge in Language Models Without Supervision](https://arxiv.org/abs/2212.03827)  - [Representation Engineering: A Top-Down Approach to AI Transparency](https://arxiv.org/abs/2310.01405)  - [Transformer visualization via dictionary learning](https://aclanthology.org/2021.deelio-1.1/)  - [How Do Transformers Learn Topic Structure: Towards a Mechanistic Understanding](https://arxiv.org/abs/2210.11865) |