Homework Assignment 1

Course: AIAA 5047 Responsible AI, Fall 2025

Instructions

This homework assignment is based on Lectures 2 and 3 of the Responsible AI course. If you use any AI tools for assistant, please indicate that in your answer or part of your answer where any AI tools were applied.

Deadline: Sep 28, 23:55 pm.

Submission: Prepare a PDF file using Latex, submit the source file (HW1.tex) to Canvas. We will compile your Latex source file to PDF using pdflatex on Linux/MacOS.

Total Score: 25 points

Problems

1. **Multiple Choice Question (Basic)**
   Which of the following are valid taxonomies of Explainable AI (XAI) according to the slides for Lecture 2?

   (A) Global vs. Local explanations

   (B) Intrinsic vs. Post-hoc explanations

   (C) Supervised vs. Unsupervised learning

   (D) Deterministic vs. Probabilistic explanations

   (E) Rule-based vs. Feature-based explanations

Scoring: 2 points (1 point for each correct choice, -1 for each wrong choice, minimum 0)

2. **Computational Problem (Medium)**

   Given a linear classification model:

   $$h_\theta(x) = \theta_1 x_1 + \theta_2 x_2 + \theta_0$$

   where $\theta = [\theta_1, \theta_2, \theta_0] = [0.5, -0.2, 10]$, and an input sample $x = [x_1, x_2] = [100, 50]$, compute:

   (a) The prediction score $z = \theta^T x$;

   (b) The local importance of features x_1 and x_2 (using the gradient × input method);

   (c) The global importance of features x_1 and x_2.

   Scoring: 2 points per subquestion, total 6 points

3. **Proof Question (Hard)**

   Prove that for Shapley values, the sum of the Shapley values of all features equals the difference between the model's prediction on the sample and the baseline prediction (when all features are missing):

   $$\sum_{i=1}^{n} \phi_i = f(x) - f(\emptyset)$$

   Hint: Refer to the definition of Shapley values and the expected form of marginal contributions.

   Scoring: 4 points (2 for definition, 2 for derivation)

4. **Multiple Choice Question (Easy)**

   Which of the following statements about the Transformer architecture are correct?

   (A) Positional Encoding is used to distinguish the meaning of the same word in different positions.

   (B) Self-attention mechanisms can capture long-range dependencies.

   (C) Residual connections and layer normalization help alleviate the vanishing gradient problem.

(D) Vision Transformer (ViT) outperforms ResNet on ImageNet without large-scale data in the ViT paper's experiments.

(E) CLIP aligns images and texts through contrastive learning.

Scoring: 4 points (1 point per correct choice, -1 per wrong choice, minimum 0)

5. **Comprehensive Question (Medium)**
   Read the paper *Counterfactual Explanations without Opening the Black Box* (Wachter et al., 2017) or *Shapley Values for Feature Importance* (Lundberg & Lee, 2017) and answer the following:

   (a) What is the main difference between counterfactual explanations and Shapley values in interpreting model decisions?

   (b) Design an experiment to show how counterfactual explanations can help doctors understand the decisions of a COVID-19 image classification model.

   (c) Counterfactual explanations may generate "unrealistic" samples. How can this issue be mitigated?

   Scoring: (a) 3 points, (b) 3 points, (c) 3 points; total 9 points