Written Exam - Example

Explainable and Trustworthy AI

Eliana Pastor

Written exam

- Structure

- 1-3 multi-choice questions – approx. 1-1.5 points each
- 2-3 open questions – short/medium answers – approx. 3-4 points each
- 1-2 open questions – medium/long answers – approx. 6-8 each

Multi-choice questions - Q1 – 1 point

What does local interpretability in Explainable Artificial Intelligence (XAI) refer to?

A) Understanding the overall behavior of the model

B) Understanding individual predictions or decisions made by the model

C) Understanding the data preprocessing steps in the model

D) Understanding the training process of the model

Multi-choice questions - Q1 – 1 point

What does local interpretability in Explainable Artificial Intelligence (XAI) refer to?

A) Understanding the overall behavior of the model

B) Understanding individual predictions or decisions made by the model

C) Understanding the data preprocessing steps in the model

D) Understanding the training process of the model

Open questions – short/medium answer – 4 points

Discuss how counterfactual explanations can be used to provide insights into model predictions. Provide an example scenario where counterfactual explanations would be useful.

Open questions – short/medium answer – 4 points

Definition

- Explanation by example
- Local explanation – explain individual prediction
- A counterfactual explanation of a prediction describes the smallest change to the feature values that changes the prediction to a predefined output.

Insights

- Smallest change to the feature values that changes the prediction to a predefined output.
- Understand the decision boundary
- Actionability – how to change the input to have a different outcome.

Scenarios, e.g.,

- Loan Application rejection
  - Understand what an applicant should change to qualify for a loan (e.g., ask for a lower amount, increase the income)
- Rating – Employee performance Evaluation
  - An employee can understand how to improve her performance (e.g., additional training)

Open questions – long answer – 7 points

Consider a trained black box healthcare AI system operating with tabular data that predicts the likelihood of disease. How would you ensure that this system provides understandable explanations to doctors? Discuss some methods and the types of explanations that would be appropriate.

Open questions – long answer – 7 points

Multiple good answers

+ **motivate the choice of why appropriate**

Use the black box model and explain it

Use post-hoc global approaches

- E.g., provide attribution-based: perturbation feature importance,
- E.g., provide visualization-based explanations via partial dependent plots
- e.g., define a global surrogate

Use post-hoc local approaches

- E.g., provide feature attribution explanations
  - Provide an importance score to each input feature for each specific prediction
  - E.g., LIME, SHAP
- **E.g., provide counterfactual explanations**
  - Explanation by example
  - Describe smallest change to change the prediction outcomes
- **Example-Based Explanations**
  - Prototype: show representative examples with similar predictions and contrasting examples
  - Case-Based Reasoning: Provide similar past cases from the training data and their outcomes

Also, hybrid of the solutions

Open questions – short/medium answer – 3 points

Outline the concept of surrogate models in XAI. How can surrogate models be used to interpret machine

Open questions – short/medium answer – 3 points

- A surrogate model involves training an interpretable model, such as decision trees or linear models, on data points labeled by the black box model to explain. The surrogate model should closely mimic the behavior of the black box model, globally or globally.
- Post-hoc explainability
- Model agnostic

- To explain
  - Globally the model– global surrogate
    - Use the whole set of data points
  - Local/individual predictions - local surrogate model
    - The data points are sampled or generated (e.g., via perturbation or genetic algorithms) in the locality of the instance to explain
- Often use only for tabular data as interpretable models often designed and work better for this type of data or require an interpretable representation

Open questions – short/medium answer – 4 points

You are provided with a black-box credit scoring model. Design an approach using LIME to explain why a particular individual was denied credit. Detail the steps you would take and the type of explanations you would generate.

Open questions – short/medium answer – 4 points

- **Def**

- Post-hoc explainability
- Local/Individual prediction
- Model agnostic
- Local surrogate model

- **High level steps of LIME**

Given an instance to explain

- Generate the locality of the instance to explain x
- Label the generated local points with the model to explain
- Weight by proximity to the instance to explain x
- Train a local interpretable model (linear)

- **Type of explanation**

- Feature attribution: importance score for each feature or interpretable feature representation
- If tabular data as we expect for scoring model: importance scores for each feature. We can use a bar plot representation

- See slide XAI_05 Local surrogate

Multi-choice questions - Q1 – 1 point

Which of the following best describes accountability in the context of Trustworthy AI?

A) The requirement that AI systems operate without any human oversight
B) The responsibility of the entities involved to ensure and answer for the outcomes produced by AI systems
C) The ability of AI systems to self-correct when errors occur
D) The focus on minimizing the cost of deploying AI systems

Multi-choice questions - Q1 – 1 point

Which of the following best describes accountability in the context of Trustworthy AI?

A) The requirement that AI systems operate without any human oversight
B) **The responsibility of the entities involved to ensure and answer for the outcomes produced by AI systems**
C) The ability of AI systems to self-correct when errors occur
D) The focus on minimizing the cost of deploying AI systems