Evaluation of Explanations

Explainable and Trustworthy AI

Eliana Pastor

Evaluating explainability: from anecdotal evidence to systematization and quantification

Anecdotal evidence

Evaluating explanation by showing convincing examples

- seems valid, plausible to us as humans, seems clear

Do not allow a systematic, quantifiable, comparable analysis on the quality of explanations

A **systematization of evaluating explanability** of AI systems is need for ensuring their quality and trustworthiness in various applications.

- Definition of properties of explanation quality
- Definition of the evaluation methods and measure to quantify the properties

Systematization framework – A note

For this slide module, we will mostly rely the characterization of the properties of explanation quality and their evaluation proposed by the survey of Nauta et al. (2023)

Note that different systematizations have been proposed and the field of evaluating explainability is an on-going research landscape

Nauta, Meike, et al. "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai." *ACM Computing Surveys* 55.13s (2023): 1-42.

Properties explanation quality

Evaluation of Explanations

Properties of explanation quality – I

Content/model

- Faithfulness
  - Correctness
  - Completeness
- Consistency
- Continuity
- Contrastivity
- Covariate complexity

Properties of explanation quality – II

Presentation

- Compactness
- Composition
- Confidence

Properties of explanation quality – III

User

- Context
- Plausibility
- Controllability

A first important distinction: Faithfulness and plausibility

- **Plausibility**: alignment of explanation with human reasoning, what we expect as humans
- **Faithfulness**: alignment of explanation with model behavior, inner working

- We cannot assume that the explanations provided by an explanation method are by default faithful!
  - We should test it. Explainers may fail
- We have no guarantee that a plausible explanation is indeed reflecting the inner reasoning of the model, and viceversa
  - A non-plausible explanation could indicate either an error in the reasoning of the model, or an error in the explainer producing the explanation

Properties of explanation quality – Content/model

- Faithfulness
  - Correctness
  - Completeness
- Consistency
- Continuity
- Contrastivity
- Covariate complexity

Faithfulness

Faithfulness of the explanation with respect to the model f to be explained

- Alignment with its inner working
- ‘Is the explanation reflecting the behavior of the model?’

Also divided into:

- **Correctness or comprehensiveness**: whether the explanation captures all elements relevant for the outcome of f
- **Completeness or sufficiency**: the extent to which the explanation covers the output of model f, i.e., whether the set of elements highlighted by the explanation is sufficient to explain the output of f

Consistency

- Identical inputs should have identical explanations
- Assess how much the explanation method is deterministic

- Implementation invariance for explanation methods that observe input and output (and not the internal working)
  - two models that give the same outputs for all inputs should have the same explanations

Continuity

- Similar inputs should have similar explanation
- Describe how continuous/smooth the explanation function is
  - For small variations of the input, we not only expect similar/nearly identifical model response, but also similar/equal explantion

Constrastivity

- Describe how the discriminativeness the explanation is with respect to other targets or events
- An explanation should not only explain the why, but also the why not, i.e., why some other event did not occur

Also separability property:

- Non-identical instances from different populations must have dissimilar explanations.

Covariate complexity

- Complexity of the covariates, i.e. features, used in the explanation I
- The ‘covariates’ should be comprehensible, e.g., using interpretable data representation

Properties of explanation quality – Presentation

- Compactness
- Composition
- Confidence

Compactness

- Size of the explanation
- Motivated by the limitation of human cognitive capacity

- Explanations should be sparse, short and not redundant
  - The more the more the explanation in compact, the better
  - More understandable for us as humans

Composition

- Describe the presentation format, organization and structure of the explanation
- Focus on *how* the prediction/model is explained
- Prioritize clear form of explanation
  - E.g., prefer higher-level information

Note that different forms of explanation can be preferred based on the target users

Confidence

- Describe if an explanation as a measure of uncertainty
  - Confidence of the explanations

Properties of explanation quality – User

- Plausibility
- Context
- Controllability

Plausibility/Coherence

- Assess the alignment of explanation with human reasoning, what we expect as human
  - with relevant background knowledge, beliefs and general consensus

- Also known as reasonableness and agreement with human rationales

Context

- Describes how relevant the explanation is to the user and their needs
- Explanations should be designed and be useful to the user, also based on level of expertise
- Designed based on the stakeholders involved: data scientist, data controllers, domain experts, policy makers

Controllability

- Assess how much a user can control, correct or interact with an explanation

Methods for evaluating explanation quality

Evaluation of Explanations

Properties of explanation quality – I

Content/model

- Faithfulness
  - Correctness
  - Completeness
- Consistency
- Continuity
- Contrastivity
- Covariate complexity

Faithfulness - Removal-based evaluation methods

- **Removal-based evaluation methods**
  - Study the effect of removing/perturbing what the explanation highlights and measure the effect on the output of f
  - Used for feature attribution methods
  - Examples: single deletion or addition, incremental deletion or addition
  - Problem: as for removal-based explanations, out of distribution samples

○ **Single Deletion**

- Evaluates the change in output when removing/perturbing one feature
  - Omitting the feature with the highest importance score for the explanation should lead the highest change in the output of f
  - Omitting the one with least importance should have no impact
  - Omitting a feature that as no effect on the output should have importance 0

Faithfulness - Removal-based evaluation methods

- **Removal-based evaluation methods**

○ **Incremental Deletion**

- Iteratively remove features
  - Descending order (from the most important to the least) or ascending order
  - Often removed subsets, e.g., top-k most influential and least
  - The impact is then summarized, e.g., Area over the Perturbation Curve, average difference in prediction scores by f

○ **Incremental Addition**

- Iteratively adding, starting from ‘empty’ input

Faithfulness - Removal-based evaluation methods Example

The Incremental Deletion can be used to evaluate the **Comprehensiveness** (or correctness) of the explanation

- Measure the drop in model probability if the important attributes are removed → are them all? If we remove indeed the important ones, we expect an high drop. The higher, the better
  - We filter out attributes with a negative contribution (i.e., they pull the prediction away from the chosen label)
  - We progressively consider th *k* most important attributes, e.g., with k ranging from 10% to 100% (step of 10%)
  - We average the result.

Faithfulness - Removal-based evaluation methods Example

The Incremental Deletion can be used to evaluate the **Sufficiency** (or completeness) of the explanation

- Measure the drop in model probability if the **not** important attributes are removed, keeping only the important ones → are them sufficient? If we preserve indeed the important ones, we expect no drop or small. The close to 0, the better
  - We filter out attributes with a negative contribution
  - We progressively consider th *k* **least** important attributes
  - We average the result.

Faithfulness – Sanity checks

- **Model Parameter Randomization Check – Sanity check**

Measure the sensitivity of the explanation to the model f

- We compare an explanation of model f with the explanation when we randomize the parameters or re-initialize weights
    - --> We expect a change in the explanatio!
- If there is no change after randomization, the explanation is not sensitive to f and hence it does not reflect the reasoning/inner working of f

Faithfulness – White Box Check

- **White box check**

Use interpretable approaches to derive ground truth explanations

- Use an explanation method to explain the prediction of a white box classifier
- Compare the explanation with the ’ground-truth’ explanation from the white box model
- Evaluate how closely the explanation reflects the true one

Faithfulness – Synthetic data check

- **Synthetic Data Check**

Use synthetic data to control the model behavior of *f* and assume the ground trith explanation

- Train a model on controlled synthetic data → we expect the model to learn such patterns
  - E.g., ‘if attribute = 1, the class = 1’
- Compare the explanation of the model with the ground-truth one, based on the controlled data
- Evaluate how closely the explanation reflects the true one

Note that we are assuming that the model *f* has learned the intended reasoning!

Faithfulness – Synthetic data check

- **Synthetic Data Check**

Use synthetic data to control the model behavior of *f* and assume the ground trith explanation

- Train a model on controlled synthetic data → we expect the model to learn such patterns
  - E.g., ‘if attribute = 1, the class = 1’
- Compare the explanation of the model with the ground-truth one, based on the controlled data
- Evaluate how closely the explanation reflects the true one

Note that we are assuming that the model *f* has learned the intended reasoning!

Faithfulness – Fidelity

- **Fidelity**

Agreement between the output of f and the explanation when applied to the input, how well the explanations mimic the output of f if use to make predictions

- Use the explanation to make prediction, e.g., by applying it if we use a surrogate model or use the attribute weights to generate a linear model
- Verify if the outcome of f and of the explanation matches
  - E.g., measure as the fraction of samples for which f and an explanation make the same decision

Different than comprehensivenss/sufficiency → compare the outputs, not the reasoning process

Consistency

- Identical inputs should have identical explanations

- **Implementation Invariance**

Two models that give the same outputs for all inputs should have the same explanations

- E.g., similarity between feature importance scores across random initializations of f

Continuity

- Similar inputs should have similar explanation

- **Stability/Sensitivity/Robustness for Sligh Variations**

Measures the similarity between explanations for an instance *x* and its slightly different version

  - e.g., consider a neighbor sample or a perturbation by adding noise
  - Compute the similarity, e.g., via rank order correlation, cosine similarity

Constrastivity – Target Sensitivity

- Describe how the discriminativeness the explanation is with respect to other targets or events

- **Target Sensitivity**

Evaluate the extent to which features highlighted by an explanation for a certain class should differ between classes

- The explanation should explain a certain class, hence should differ from the explanation for other classes
- Compute similarity between explanations for x with respect to different classes
- The larger the difference, the better

Covariate complexity

- Complexity of the covariates, i.e., human-understandable concepts used in the explanation
- Often used for Concept-based XAI
- **Covariate Homogeneity**
- how consistently a covariate (e.g., prototype/cluster of images) represents a predefined human-interpretable concept
- How disentangled the covariate are – e.g., prototype represents a single concept

Properties of explanation quality – II

Presentation

- Compactness
- Composition
  - Focus on how the prediction/model is explained
  - Often evaluated via anecdotal evidence
  - More adopted via user studies
- Confidence
  - Check if the explanation contains uncertainty information
  - Few methods assess this aspect

Compactness

- **Size** of the explanation

e.g., number of features in the explanation, length of the rule/path

- **Redundancy**
- The lower the overlap among explanation, the higher the interpretability

Composition

- Composition

Describe the presentation format, organization and structure of the explanation

  - Focus on *how* the prediction/model is explained

- Often evaluated via anectodal evidence
- More adopted via user studies

Properties of explanation quality – III

User

- Plausibility
  - Mostly assessed with user studies
- Context
  - Describes how relevant the explanation is to the user and their needs
  - Mostly assessed with user studies
- Controllability
  - Assess how much a user can control, correct or interact with an explanation
  - Mostly assessed with user studies or via anectodal evidence

Plausibility/Coherence

- Assess the alignment of explanation with human reasoning, what we expect as human

- **Alignment with Domain Knowledge.**

Evaluated via

- **User studies**
- Comparison of explanations with **ground truth explanations** from datasets annotated with human rationales
    - Evaluate similarity, e.g., rank correlation for feature importance, intersection-over-union for saliency maps, ROUGE and BLEU for textual explanations

Plausibility/Coherence

- **XAI Methods agreement**

Evaluate the agreement among explainers

We can compare a novel explainer with an established one with certain properties

References

Nauta, Meike, et al. "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai." *ACM Computing Surveys* 55.13s (2023): 1-42.