Concepts in Quality Assessment for Machine Learning - From Test Data to Arguments

Fuyuki Ishikawa
National Institute of Informatics
Background: (Engineering) Machine Learning

- Active effort to apply machine learning (ML)
  - Intensive support on libraries and platforms

- Now engineering support is essential
  - Defining and finalizing products with customers
  - Arguing quality of products
  - ...

- Conceptual models!

  Play the essential role of capturing the essence of the product and its quality

Oct 24 2018  
 f-ishikawa@ER'18
Motivation: Essential Difference in ML

- With ML, we obtain the behavior of a component (e.g., a neural net) \textit{inductively} from training data.
  
  \textit{Black-box, imperfect, non-testable (no oracle), unexplainable, has adversarial examples, …}

\textit{Existing principles do not work}

\textit{When arguing the product and its quality}

\textit{How difficult??}

Questionnaire survey by SIG-MLSE, Japan, 2018
Motivation: Necessary Concepts?

Argue the product and its quality!

*(engineer-engineer or engineer-customer)*

The system shall recognize objects in front of the car!

As usual, a function will be surely implemented and assurance will be given (by “sufficient” tests)

Great!
Motivation: Necessary Concepts?

Argue the product and its quality!

(Engineer-engineer or engineer-customer)

The system shall recognize objects in front of the car!

As usual, a function will be surely implemented and assurance will be given (by “sufficient” tests)

It’s imperfect and we cannot make assurance on the behavior for untested data

Great!

Oh, then we need to carefully assess the quality based on what are actually tested
Proposal: MLQ Framework

- Framework for assessing the quality of ML components and ML-based systems

- Focuses on concepts to capture test data, or empirical evidence in more general

- Reflects the state-of-the-art research on testing ML

- Uses an argument model (e.g., assurance case in Goal-Structuring Notation) to describe the whole picture

- To be linked to test-data management tools
Example: Test Data Attribution

- To specify & check constraints
- To describe current status
- To discuss validity
- To compare with operational data

- carCameralmg1 : TestData
- atr1 : Attribute
- atr2 : Attribute
- weatherReport : AttributionMethod
- middle : Confidence
- situation : AttributeType
- sunrise : AttributeValue
- weatherCondition : AttributeType
- cloudy : AttributeValue

We tested with 100,000 data!

What data … ? Did you test misty days?

Actual representation may be spreadsheet, GUI, or whatever
Different targets of assessment in the ML dev. process
Different ways of assessment that covers the present techniques (accuracy, metamorphic relations, robustness, etc.)
Deductive/logical and inductive/empirical assessment (also covering runtime to tackle the uncertainty)
Argument Example: Whole Picture

**Requirements-based coverage (for accuracy)**

**Whitebox coverage (e.g., neuron coverage)**

**Robustness**

**Whitebox assessment (investigation of weakness)**

**Evidences (test data and other assessment results) are managed with concepts annotated**
Summary: MLQ Framework

- Framework for **assessing the quality** of ML components and ML-based systems
  
  From test data to arguments
  
- Ongoing/Future Work
  - Elaboration with Case Studies
  - Tools, especially connection with test data management
  - Uncertainty-aware argument modeling [ASSURE’18]
    (awareness of risks and continuous engineering)
Thank You!