Towards Comparing Learned Classifiers

Soaibuzzaman

The most similar real face of ID 8 (no access)

MLDiff Witness:
MLP: ID 8 (no access)
LogReg: access

The most similar real face with access

Introduction & Example

- Numerous complex, real-world applications rely on Machine Learning (ML) classifiers

- Example classification problem
  - A Support Vector Machine (SVM) & Decision Tree (DT) are both trained on the Iris dataset
  - Accuracy: SVM=96% and DT=96%

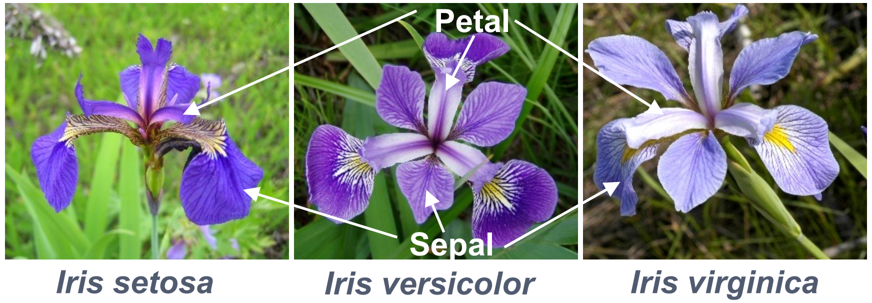

Figure: Iris flower data set (https://en.wikipedia.org/wiki/Iris_flower_data_set)

- Should we use the SVM or the DT?

- How are the two classifiers different?

- Is it relevant which one to use if they agree on known data (the train/test dataset)?

DT classifies some instances as *Virginica* (**medical use**), while SVM classifies them as *Versicolor* (**poisonous**)

Bauhaus-Universität Weimar

Motivation of MLDiff

- State-of-the-art formal analysis of ML models lacks systematic methods to compare multiple classifiers
- Understanding classifier variants during software design and evolution is crucial for improving model quality and trust
- MLDiff aims to uncover and present differences (witnesses), i.e., disagreements, of classifiers (even those not observable in the dataset)

The LogReg classifier grants access to employee ID 8 when the MLP classifier denies it.

Figure: A conflict of a Multi-Layer Perceptron and a Logistic Regression classifier detected by MLDiff
(Source: The Olivetti faces dataset from scikit-learn)

MLDiff Implementation

- For two classifiers cl_1 on features X_1 and cl_2 on features X_2 we **encode SMT assertions** for

$$\forall d \in \mathbb{R}^{|X_1 \cup X_2|}: cl_1 \oplus cl_2\ (d) = cl_1(d|_{X_1}) \times cl_2(d|_{X_2})$$

```
(declare-const x0 Real) ; one constant for each feature
; ...
(declare-const xn Real)
(declare-const cls1 Int) ; predicted class of first classifier
(declare-const cls2 Int) ; predicted class of second classifier
; assertion for classifier 1 relating x1..xn to cls1
; assertion for classifier 2 relating x1..xn to cls2
(assert (not (= cls1 cls2))) ; example query for disagreement
```

Currently supports
- *Decision Trees*
- *Logistic Regression*
- *Multi Layer Perceptron (ReLU, identity)*
- *Support Vector Machine (linear kernels)*

- Use Cases and Queries
  - Differences: $cl_1(d|_{X_1}) \neq cl_2(d|_{X_2})$
  - Extension with **custom/domain constraints**:

classifier disagreement

$$x_1 = 1 \wedge x_3 = 4 \wedge x_6 \leq 0.2 \wedge cl_1(d|_{X_1}) \neq cl_2(d|_{X_2})$$

mammal (categorical x_1) *4 legs (x_3)* *small (weight x_6)*

Bauhaus-Universität Weimar

Challenges and Open Problems

- Supporting a larger set of functions in classifiers (SMT's arithmetic limitations)
- Scaling to complex models and queries (approximation and decomposition)
- Developing a domain-expert-friendly query language
- Generating relevant and interesting in-domain examples
- Exploring examples and explanations (summarizing and explainable AI)