Assured Machine Learning

Xiaozhe Gu, Arvind Easwaran

School of Computer Science and Engineering
Energy Research Institute (ERI@N)

Nanyang Technological University, Singapore

June, 2018

Outline

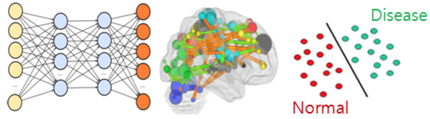
Outline

Machine Learning Applications in Safety-Critical Environments

Machine learning algorithms are increasingly influencing our lives, and moving into safety-critical applications.

- Decision making in life-threatening conditions, e.g., ML based medical decision support systems.

Figure: ML Based Brain Disease Diagnosis[1]



Machine Learning Applications in Safety-Critical Environments

Machine learning algorithms are increasingly influencing our lives, and moving into safety-critical applications.

- Autonomous Robots, e.g., surgical robots, rescue robots, industrial robots, etc.

Figure: Surgical Robots[2]

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Machine Learning Applications in Safety-Critical Environments

Machine learning algorithms are increasingly influencing our lives, and moving into safety-critical applications.

- Self-Driving Vehicles, e.g, autonomous shuttle.

Figure: Autonomous Shuttle

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Machine Learning Applications in Safety-Critical Environments

Machine learning algorithms are increasingly influencing our lives, and moving into safety-critical applications.

- Autonomous Weapons

Figure: SGR-A1 (wikipedia)

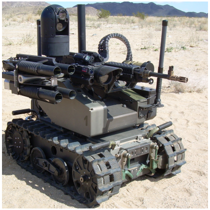
Machine Learning Applications in Safety-Critical Environments

Machine learning algorithms are increasingly influencing our lives, and moving into safety-critical applications.

- Autonomous Weapons

Figure: SGR-A1 (wikipedia)

Figure: keelvar.com



Traditional Programing versus Machine Learning

Figure: Traditional Programming

Traditional programming involves static program instructions strictly specifying what we need the computer to do.

Traditional Programing versus Machine Learning

Figure: Traditional Programming

Traditional programming involves static program instructions strictly specifying what we need the computer to do.

Figure: Machine Learning

ML is a technique that enables computers to learn autonomously and to improve from experience without being explicitly programmed

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Machine Learning Safety

Amodei et al. [5] refer to ML safety as *" mitigating risk in the context of unintended or harmful behaviour that may emerge from machine learning systems when we "*

- *specify the wrong objective function,*
- *are not careful about the learning process,*
- *or commit other machine learning-related implementation errors.*

Machine Learning Safety

Amodei et al. [5] refer to ML safety as *" mitigating risk in the context of unintended or harmful behaviour that may emerge from machine learning systems when we "*

- *specify the wrong objective function ,*
- *are not careful about the learning process ,*
- *or commit other machine learning-related implementation errors.*

Bostrom et al.[6] refer to safety as *" techniques that ensure that machine learning systems behave as intended ".*

Challenges to Safety Assurance

- Non-transparency: The reasoning behind some powerful ML models is not known.

Challenges to Safety Assurance

- Non-transparency: The reasoning behind some powerful ML models is not known.
- Unmodeled phenomena: It is not impossible and not desired to model everything.

Challenges to Safety Assurance

- Non-transparency: The reasoning behind some powerful ML models is not known.
- Unmodeled phenomena: It is not impossible and not desired to model everything.
- Instability: A small change in the training process may produce a different result.

Challenges to Safety Assurance

- Non-transparency: The reasoning behind some powerful ML models is not known.
- Unmodeled phenomena: It is not impossible and not desired to model everything.
- Instability: A small change in the training process may produce a different result.
- Difficulty in verification: Formal verification of ML components is a difficult, and somewhat ill-posed problem.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Challenges to Safety Assurance

- Non-transparency: The reasoning behind some powerful ML models is not known.
- Unmodeled phenomena: It is not impossible and not desired to model everything.
- Instability: A small change in the training process may produce a different result.
- Difficulty in verification: Formal verification of ML components is a difficult, and somewhat ill-posed problem.
- Incorrect specification: The incorrect specification objective function can result in harmful and unintended results.

Outline

Motivation

Improve Interpretablilty and Transparency

Safe Fail

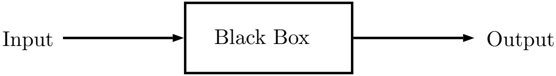
Verification of ML

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Non-transparency

Many machine learning models (e.g., deep neural network) behave mostly as black boxes.

Figure: Black Box



Understanding the reasons behind model/prediction is, however, quite important in assessing trust without which if the users will not deploy/use it.

Interpretable Models

Very common model types of interpretable models are:

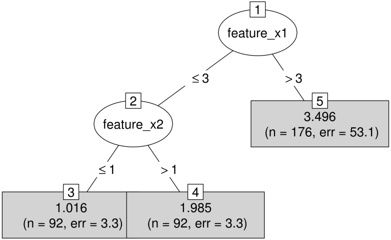
- Linear regression model
  $y = w^T.\mathbf{x} + \epsilon = w_0 + \sum_{i=1}^{m} w_i x_i + \epsilon.$

Interpretable Models

Very common model types of interpretable models are:

- Linear regression model
  $y = w^T.\mathbf{x} + \epsilon = w_0 + \sum_{i=1}^{m} w_i x_i + \epsilon.$
- Decision trees.

Figure: Regression Tree



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

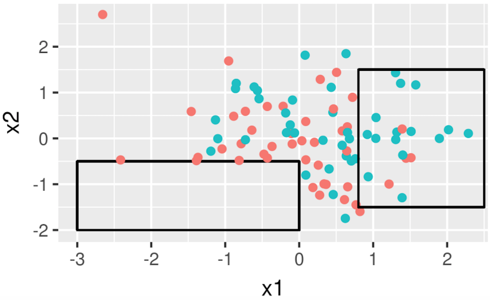
Interpretable Models

Very common model types of interpretable models are:

- Decision Rules:

  `If` $-3 \leq x1 \leq 0$ `and` $-2 \leq x2 \leq -0.5$`:` `then` $y = +1$

Figure: Decision Rule

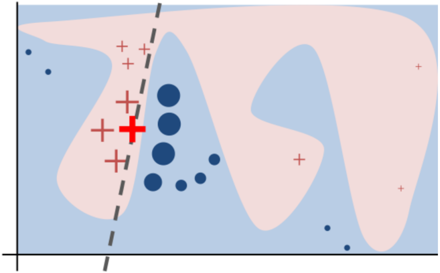


NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Prediction Explanation

Explain the prediction of a classifier by approximating it locally with an interpretable model [3].

Figure: The black-box model's complex decision represented by the blue/pink background is approximated locally by a linear model.

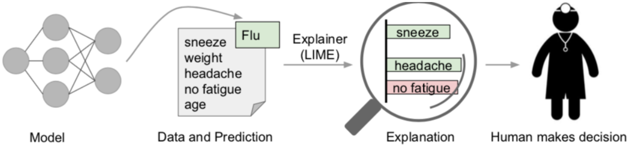


NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Prediction Explanation

With the explanation of a prediction, a doctor can make an informed decision about whether to trust the model's prediction.

Figure: Explaining individual predictions



Global Surrogate Models

A global surrogate model is an interpretable model that is trained to approximate a black box model

How well the surrogate replicates the black box model?

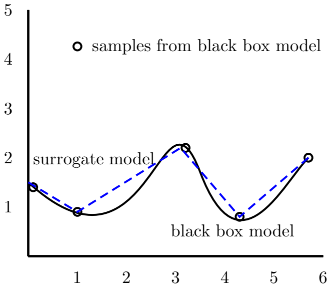
$$\text{Metric:}1 - \frac{\sum_{i=1}^{n}(\hat{y}_i^* - \hat{y}_i)^2}{\sum_{i=1}^{n}(\hat{y}_i^* - \bar{\hat{y}}_i)}$$

where $\hat{y}_i^*$ and $\hat{y}_i$ is the prediction of the surrogate model and respectively of the black box model.

Global Surrogate Models

A global surrogate model is an interpretable model that is trained to approximate a black box model

Figure: Linear Surrogate Model



Outline

Motivation

Improve Interpretablilty and Transparency

Safe Fail

Verification of ML

Learning to Reject

A technique used in machine learning when predictions cannot be given confidently is the reject option [4].

$$f(x_i) = \text{rejection if } g(f, x_i) \leq \sigma$$

where $g(f, x_i)$ measures the confidence level of function f's prediction for x_i, and σ is a threshold.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Learning to Reject

A technique used in machine learning when predictions cannot be given confidently is the reject option [4].

$$f(x_i) = \text{rejection if } g(f, x_i) \leq \sigma$$

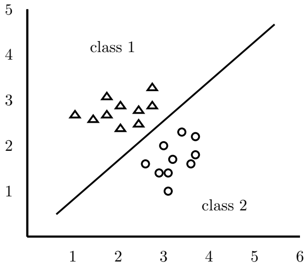
where $g(f, x_i)$ measures the confidence level of function f's prediction for x_i, and σ is a threshold.

When the model selects the reject option, a human operator can intervene and provide a manual prediction.

Distance from Decision Boundary

In classification problems, classifier implicitly assumes that distance from the decision boundary is inversely related to confidence [7].

Figure: Decision Boundary

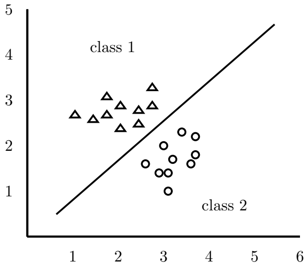


NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Distance from Decision Boundary

In classification problems, classifier implicitly assumes that distance from the decision boundary is inversely related to confidence [7].

Figure: Decision Boundary



This is reasonable because the decision boundary is located where there is a large overlap in likelihood functions.

Ensemble Classification

In ensemble classification, the overall decision $\hat{y}$ is based on the average classification of the base classifiers $\hat{y}_i(.)$,

$$\phi(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m}\hat{y}_i(\mathbf{x})$$

Ensemble Classification

In ensemble classification, the overall decision $\hat{y}$ is based on the average classification of the base classifiers $\hat{y}_i(.)$,

$$\phi(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m}\hat{y}_i(\mathbf{x})$$

The reject option of ensemble binary classification is based on the votes of the base classifiers

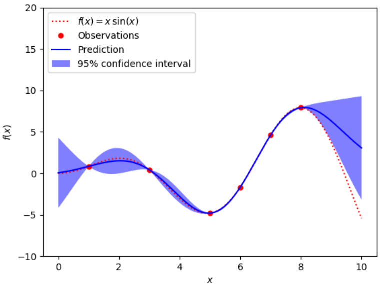
$$\hat{y}(x) = \begin{cases} -1 \ \text{ if } \phi(\mathbf{x}) \leq t_1 \\ \text{reject, if } \phi(x) \in (t_1, t_2) \\ 1 \ \text{ if } \phi(x) \geq t_2 \end{cases}$$

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Gaussian Process

A Gaussian process defines a distribution over functions $p(f)$ where f is a function mapping some input space $\mathcal{X} \rightarrow \mathbb{R}$.

Figure: One-dimensional Regression

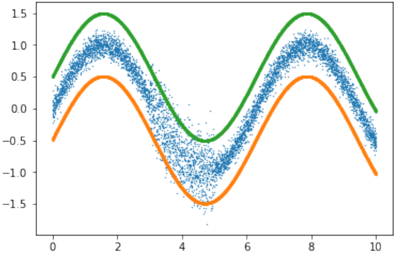


Prediction for input **x** with larger standard deviation implicates lower confidence. Thus, we can reject to predict for input input **x** with larger standard deviation.

Conformal Prediction: Machine Learning with Confidence Intervals

The conform prediction [8] gives us valid bounds $[f(\mathbf{x})-\epsilon_1, f(\mathbf{x})+\epsilon_2]$ for any model such that the prediction region contains the true output with probability α.

Figure: The Bound with 95% Confidence

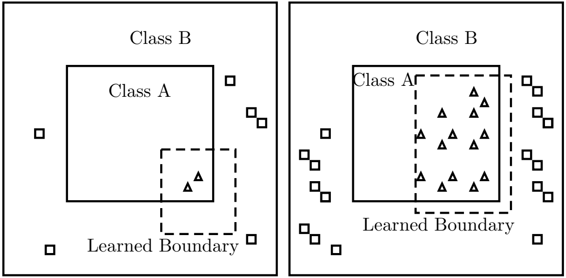


NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Reject Unmodeled Scenario

It is not impossible for ML to model everything. Thus, if the model recognize that instance is from the input space that is not well trained, it could reject to predict.

Figure: Wrong Decision Boundary due to Lack of Data

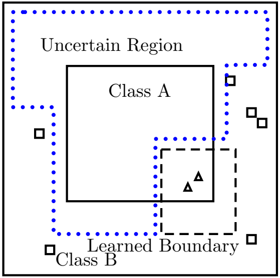


NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Reject Unmodeled Scenario

It is not impossible for ML to model everything. Thus, if the model recognize that instance is from the input space that is not well trained, it could reject to predict.

Figure: Region with Low-confidence



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Outline

Motivation

Improve Interpretablilty and Transparency

Safe Fail

Verification of ML

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Challenges for Verified ML

The formal verification of ML components is a difficult, and somewhat ill-posed problem due to the complexity of the ML algorithms, large feature spaces [9].

Challenges to achieving formally-verified ML-based systems:

- Environment Modeling:
  It may be impossible even to precisely define all the variables (features) of the environment that must be modeled, let alone to model all possible behaviors of the environment

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Challenges for Verified ML

The formal verification of ML components is a difficult, and somewhat ill-posed problem due to the complexity of the ML algorithms, large feature spaces [9].

Challenges to achieving formally-verified ML-based systems:

- Formal Specification:
  It is difficult to find an effective method to specify desired and undesired properties of systems that use ML-based components.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Challenges for Verified ML

The formal verification of ML components is a difficult, and somewhat ill-posed problem due to the complexity of the ML algorithms, large feature spaces [9].

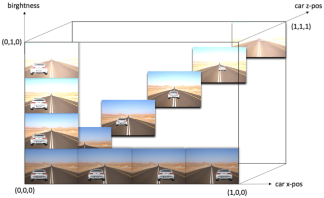
Challenges to achieving formally-verified ML-based systems:

- System Modeling:
  Unlike traditional applications of formal verification, the system S is precisely known, e.g., a C program, ML based system evolves as it encounters new data and new situations.

Approximate Model in Abstract Space

Abstract Feature Space: Instead of the high-dimensional input space, explore realistic and meaningful simplification modifications [10].

Figure: The abstract space A with the three dimensions

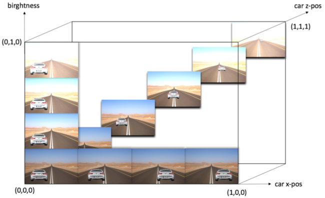


NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Approximate Model in Abstract Space

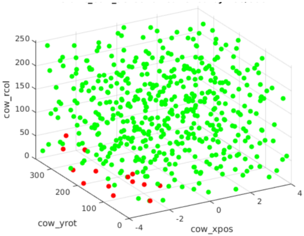
Abstract Feature Space: Instead of the high-dimensional input space, explore realistic and meaningful simplification modifications [10].

Figure: The abstract space A with the three dimensions



A simpler approximate function $\hat{f}$ of origin model f on the abstract domain is analyzed.

Figure: Misclassifying Elements



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Reduce the Problem

- Reduce the input set $\mathcal{X}$ of n dimension into a finite number of regularized subregions. For example, partition an input set into hyper-rectangles [11].

$$\mathcal{P}_i \leftarrow \mathcal{I}_{1,m_1} \times \mathcal{I}_{2,m_2} \times \ldots \times \mathcal{I}_{n,m_n}$$

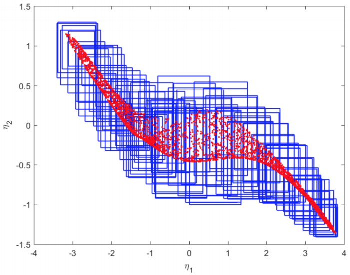
where $I_{i,m_i} = [\eta_{m_i-1}, \eta_{m_i}]$

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Reduce the Problem

- For each hyper-rectangle input set, over-approximated the output set by a hyper-rectangle $[\underline{\phi}_j, \bar{\phi}_j]$

Figure: Blue rectangles: the over-approximated output set, red spots: 5000 random outputs located in the estimated output set [11]



- http://mlcenter.postech.ac.kr/healthcare
- https://www.wired.com/2015/03/google-robot-surgery/
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.
- Bartlett P L, Wegkamp M H. Classification with a reject option using a hinge loss[J]. Journal of Machine Learning Research, 2008, 9(Aug): 1823-1840.
- Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mane. 2016. Concrete Problems in AI Safety.

- Bostrom, N., Dafoe, A., and Flynn, C. 2016. Policy Desiderata in the Development of Machine Superintelligence
- K. R. Varshney, R. J. Prenger, T. L. Marlatt, B. Y. Chen, and W. G. Hanley, Practical ensemble classification error bounds for different operating points, IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 11, pp. 2590-2601, Nov. 2013.
- Vovk V, Gammerman A, Shafer G. Conformal prediction[M]. Springer US, 2005
- S. A. Seshia, D. Sadigh, and S. S. Sastry. Towards verified artificial intelligence. CoRR, abs/1606.08514, 2016.
- Tommaso Dreossi, Alexandre Donze, Sanjit A. Seshia, Compositional Falsification of Cyber-Physical Systems with Machine Learning Components, Preprint/

Weiming Xiang and Taylor T. Johnson, Reachability Analysis and Safety Verification for Neural Network Control Systems,Preprint.