This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2022-9199C

MLDL

Machine Learning and Deep Learning Conference 2022

Decision Science for Machine Learning (DeSciML)

- Rich Field / 5553
- Michael Darling / 8732, J.D. Doak / 8765, James Headen / 6754, Mark Smith / 5493
- J. Eric Bickel / UTA, Jason Boada / UTA, Zack Smith / UTA

- LDRD Project 222397

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

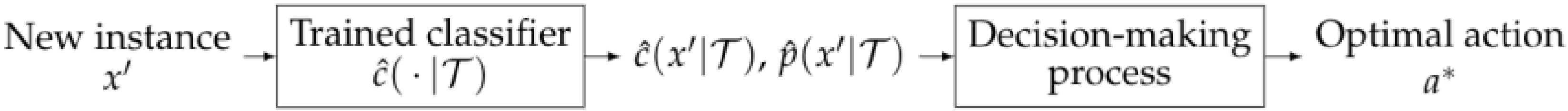
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Abstract

While the use of machine learning (ML) classifiers is widespread, their output is often not part of any follow-on decision-making process. To illustrate, consider the scenario where we have developed and trained an ML classifier to find malicious URL links. In this scenario, network administrators must decide whether to allow a computer user to visit a particular website, or to instead block access because the site is deemed malicious. It would be very beneficial if decisions such as these could be made automatically using a trained ML classifier. Unfortunately, due to a variety of reasons discussed herein, the output from these classifiers can be uncertain, rendering downstream decisions difficult. Herein, we provide a framework for: (1) quantifying and propagating uncertainty in ML classifiers; (2) formally linking ML outputs with the decision-making process; and (3) making optimal decisions for classification under uncertainty with single or multiple objectives.

Overview of the Problem

- Objective: Use outputs from a ML classifier for decision-making
  - Make decisions for individual instances
  - Account for uncertainty in ML model outputs
- ML classification is not the same as decision-making
- Cost-sensitive ML is limited
  - Minimizes costs averaged over a population instead of optimizing individual decisions
  - Cost may not be the only objective
- Decision theory can be used
  - Account for cost / consequences of each decision
  - Can consider multiple objectives (cost, risk, security, etc.)
  - Standard DT must be modified to handle uncertainty in ML model outputs
- Example application: URL classification



New instance x' → Trained classifier $\hat{c}(\cdot|\mathcal{T})$ → $\hat{c}(x'|\mathcal{T}), \hat{p}(x'|\mathcal{T})$ → Decision-making process → Optimal action a^*

Motivating Example: URL Classification

- **Objective**: Decide whether or not to allow a connection with an external website
- Training data: 127,684 labeled examples
  - 50% labeled **Malicious**, 50% labeled **Benign**
  - 87 features: text, length, counts, patterns
- Trained ML classifier
  - CART decision tree with probabilistic class predictions
- Let x' be a previously unseen URL
  - ML model provides an estimate $p(x')$ that URL x' is **Malicious** and $1 - p(x')$ that x' is **Benign**
- Possible actions
  - ***Allow*** access to the website
  - ***Block*** access to the website

$$\underbrace{\text{https://www.facebook.com}}_{\text{HostName}}\underbrace{\text{/help/cookies/}}_{\text{Path}}\underbrace{\text{?ref=sitefooter}}_{\text{Parameters}}$$

- What action should we take on x'?

Decision Theory Applied to ML Classification

A decision problem has 4 ingredients:

1. The collection of possible actions to take

$$\mathcal{A} = \{\text{all candidate actions}\} = \{a_i\} = \{\texttt{Allow}, \texttt{Block}\}$$

2. The collection of possible states of nature (the true labels)

		Nature, $\mathcal{C}$	
		benign	malicious
Action, $\mathcal{A}$	allow	0	20
	block	5	1

$$\mathcal{C} = \{\text{all possible labels}\} = \{C_i\} = \{\texttt{Benign}, \texttt{Malicious}\}$$

3. A loss function

$$L \colon \mathcal{A} \times \mathcal{C} \to [0, \infty)$$

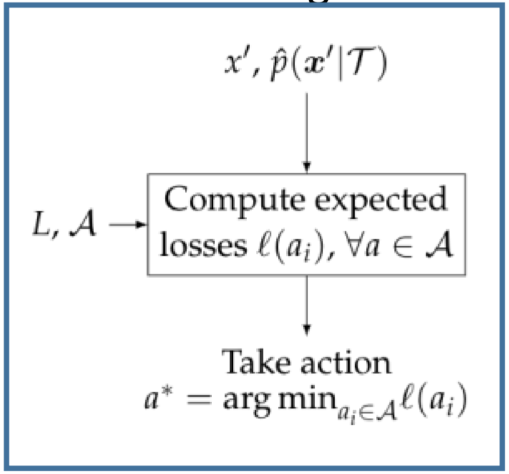
- Quantifies the consequences of taking an action given the true label

4. The optimal action

$$\ell(a_i) = \mathbb{E}[L(a_i, \mathcal{C})] = \sum_{j=1}^{K} L(a_i, C_j)\hat{p}_j(x')$$

- Choose the action with the minimum expected loss

Decision-Making Process



$x', \hat{p}(x'|\mathcal{T})$

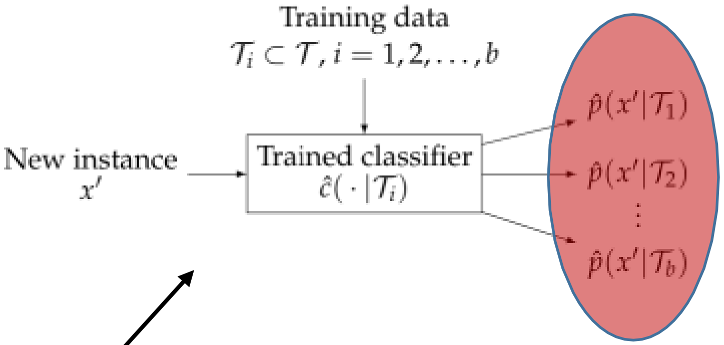
$L, \mathcal{A}$

Compute expected losses $\ell(a_i), \forall a \in \mathcal{A}$

Take action $a^* = \arg\min_{a_i \in \mathcal{A}} \ell(a_i)$

Uncertainty in ML Models

- Data
  - Pre-processing
  - Errors on features and labels
  - Limited training data
  - Extrapolation
- Model
  - Model form
  - Feature selection
- This study
  - Bootstrap sampling of the training set $\mathcal{T}$
  - Retrain the ML model for each sample
  - Predict on new instance x' with the ensemble of trained models
  - The output probability score is a random variable



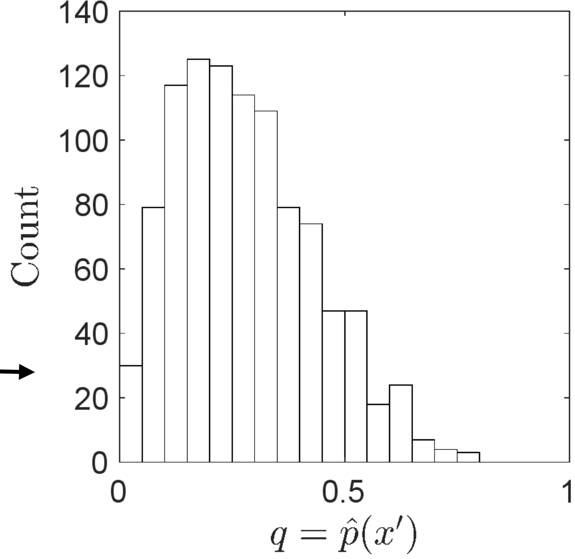
Training data
$\mathcal{T}_i \subset \mathcal{T}, i = 1, 2, \ldots, b$

New instance
x'

Trained classifier
$\hat{c}(\cdot\,|\mathcal{T}_i)$

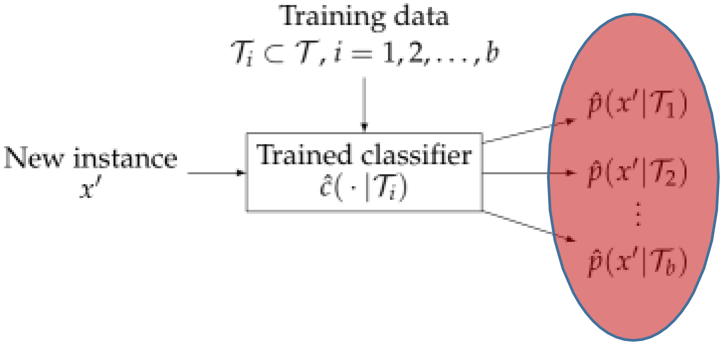
$\hat{p}(x'|\mathcal{T}_1)$
$\hat{p}(x'|\mathcal{T}_2)$
$\vdots$
$\hat{p}(x'|\mathcal{T}_b)$

Output from 1,000 ML models for one instance



The Distribution of Model Outputs

- The distribution of ML model outputs is important
    - The 3 examples below have identical mean

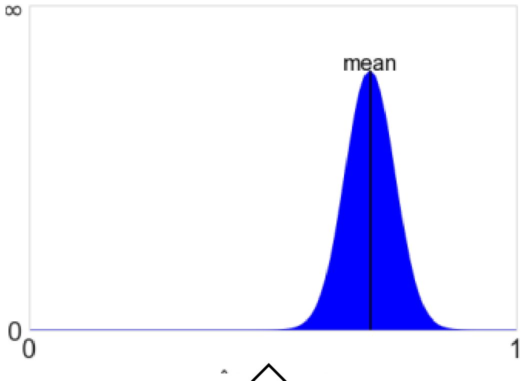


Training data
$\mathcal{T}_i \subset \mathcal{T}, i = 1, 2, \ldots, b$

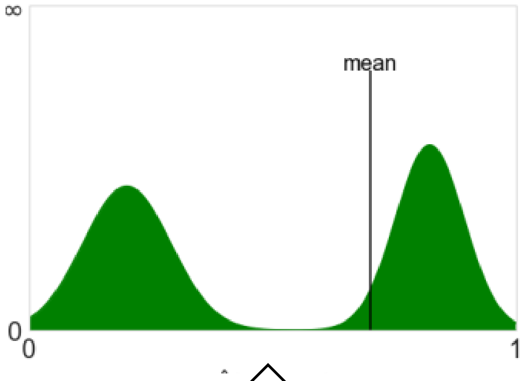
New instance
x'

Trained classifier
$\hat{c}(\cdot|\mathcal{T}_i)$

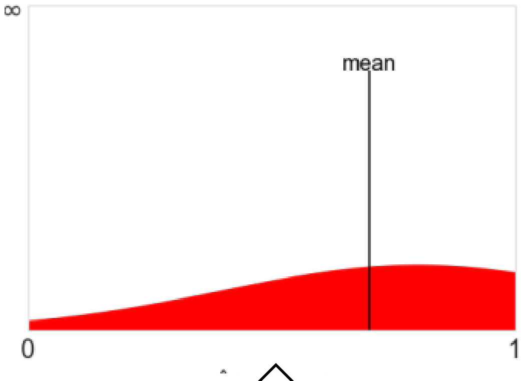
$\hat{p}(x'|\mathcal{T}_1)$
$\hat{p}(x'|\mathcal{T}_2)$
$\vdots$
$\hat{p}(x'|\mathcal{T}_b)$



This shape suggests the ML can provide a stable estimate



This shape suggests at least two plausible interpretations

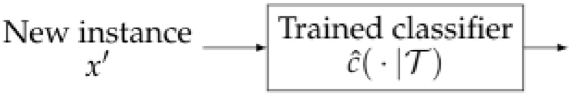


This shape suggests the model is sensitive to particular training samples

Modify the Expected Loss

- The optimal action minimizes the expected loss

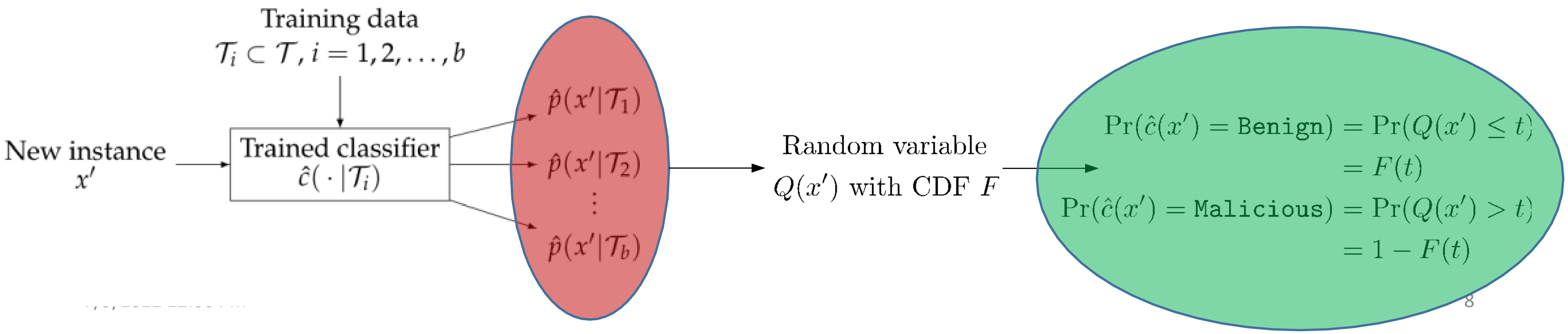
$$\ell(a_i) = \mathbb{E}[L(a_i, \mathcal{C})] = \sum_{j=1}^{\kappa} L(a_i, C_j)\hat{p}_j(x')$$


New instance x' → Trained classifier $\hat{c}(\cdot\,|\mathcal{T})$ →

$$\hat{p}(x') = \Pr(x' \text{ is } \texttt{Malicious})$$

$$\hat{c}(x') = \begin{cases} \texttt{Benign} & \text{if } p(x') \le t \\ \texttt{Malicious} & \text{if } p(x') > t \end{cases}$$

- But the probability is now a random variable How do we handle this?
    - We could use the mean, majority voting, etc., but that doesn't utilize all information


Training data $\mathcal{T}_i \subset \mathcal{T}, i = 1, 2, \ldots, b$

New instance x'

Trained classifier $\hat{c}(\cdot\,|\mathcal{T}_i)$

$\hat{p}(x'|\mathcal{T}_1)$

$\hat{p}(x'|\mathcal{T}_2)$

$\vdots$

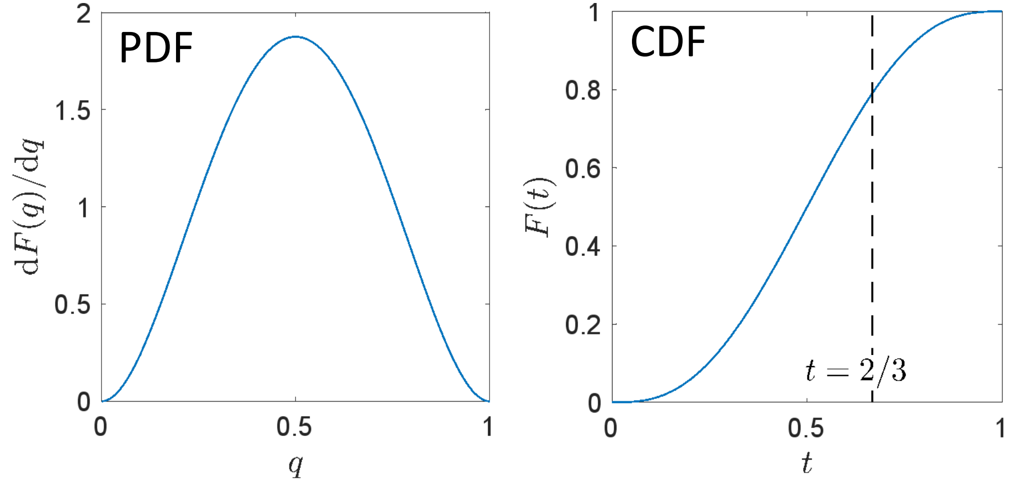
$\hat{p}(x'|\mathcal{T}_b)$

Random variable $Q(x')$ with CDF F

$$\begin{aligned} \Pr(\hat{c}(x') = \texttt{Benign}) &= \Pr(Q(x') \le t) \\ &= F(t) \\ \Pr(\hat{c}(x') = \texttt{Malicious}) &= \Pr(Q(x') > t) \\ &= 1 - F(t) \end{aligned}$$

Examples

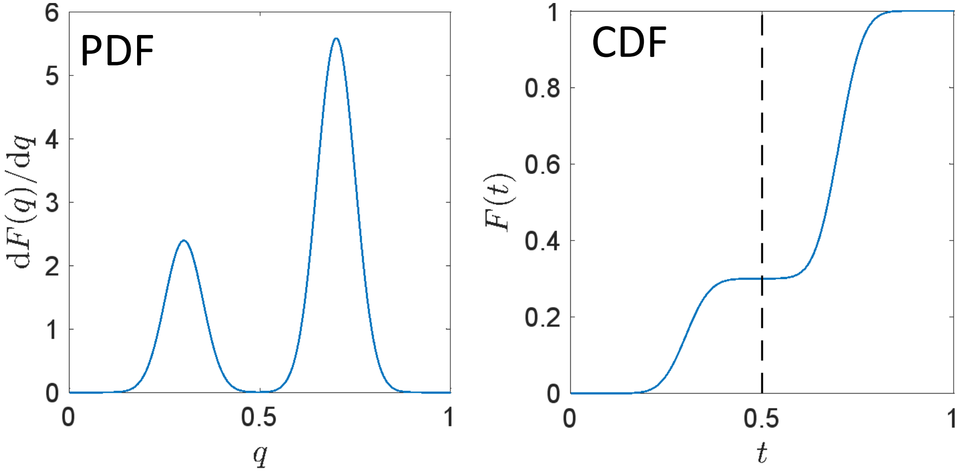
- Assume the ensemble of outputs follows a beta distribution



$$\Pr(\hat{c}(x') = \texttt{Benign}) = 0.8$$
$$\Pr(\hat{c}(x') = \texttt{Malicious}) = 0.2$$
$$\text{Mean} = 0.5$$

The result is the same as the mean value only if the PDF is symmetric and t = 1/2

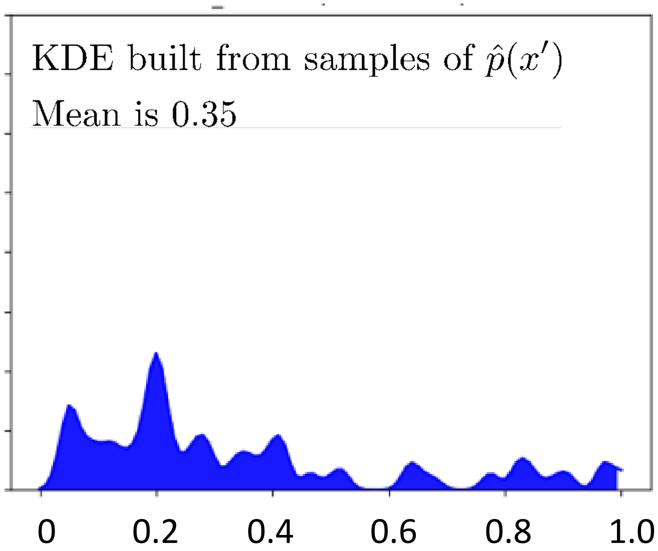
- Or a mixture of 2 Gaussian variables



$$\Pr(\hat{c}(x') = \texttt{Benign}) = 0.3$$
$$\Pr(\hat{c}(x') = \texttt{Malicious}) = 0.7$$
$$\text{Mean} = 0.58$$

Results for URL Classification

- True label = **Malicious**
- PDF estimate from ensemble of ML models is at right
  - Traditional approach would be to train with all the data and run once
  - If we account uncertainty, we would still call this **Benign** and therefore decide to **Allow**
- We apply our approach
  - Asymmetric loss function
  - Decision is to **Block**
- When we consider all test URLs in the dataset



Approach	Accuracy	# Malicious URLs Accessed	Average loss
Traditional approach	90.6%	1,221	$3.10
Cost-sensitive ML	79.4%	185	$0.65
DT (no uncertainty)	84.4%	309	$0.40
DT with uncertainty	98.5%	249	$0.23

Summary

- Objective: Use outputs from a ML classifier for decision-making on individual instances
    - Account for uncertainty in ML model outputs
    - Applied decision theory
- We applied this to URL classification; the proposed approach outperformed
    - Traditional approach
    - Cost-sensitive ML
- Additional work that we don’t have time to share
    - Minimum prediction deviation (MPD)
    - Generalized to more than two classes
    - Multi-objective decisions
    - Image classification